Analysis of the Spread and Evolution of COVID-19

Justin Deterding
Dept. of Physics and Astronomy
University of New Mexico
jdeterding@unm.edu

Catherine Wright
Dept. of Computer Science
University of New Mexico
wrightc@unm.edu

***Abstract*—The COVID-19 virus from the β-coronavirus cluster emerged in December 2019 and has since infected over 1.6 million people worldwide. Our work starts by analyzing the evolutionary adaptation and innovation of several samples of the virus. We simulate a neutral network that traverses on phenotypically-invariant mutations to explore it's relation to the SARS outbreak of 2002 and determine that the virus is quickly modifying itself to become more resilient. We then model the typical infection spread of a simple virus using a cellular automaton to imitate disease behavior in a crowded population. Understanding the spread and mutability of COVID-19 is critical as countries balance the need for individual isolation with the importance of a steady economy. While the results presented are worrisome, we intend to expand our research by creating a more rigorous model to explore infection rates given the postulated incubation periods and rate of transmission of the disease.**

I. Introduction

Since it's creation, the universe has expanded following what Murray Gell-Mann calls the "relentless operation of chance" [1]. Any human born today owes it's existence to the long stream of probabilistic events, both microscopic and macroscopic, that led to its creation. The majority of these events do not cause a noticeable change to the world around them, but there are cases of such probabilistic events having drastic consequences on the future. Gell-Mann ties such events to what Francis Crick called *frozen accidents* after co-discovering the human genetic code and suggesting that our genetic structure was most likely an accident that became frozen in time. We can view the mutation and transmission of the zoonotic SARS-CoV-2 virus from animals to humans as a frozen accident, as the infection of a single (or a small handful) of humans in Wuhan, China has had a large impact on daily lives worldwide.

Such accidents are themselves a measure of complexity; the measure of complexity as thermodynamic depth, which classifies an organism's complexity to be proportional to the amount of thermodynamic and informational resources necessary for its construction. This is one of the many postulated ways of measuring the complexity of systems [2], with some of the more simple measurements being the size of the genome, the entropy, or the algorithmic information content. Such methods are intuitive to what we expect complexity to define, however all fall short when applied to a diverse range of real and simulated genomes. Other measurements capture the essence of complexity, however, they are very difficult to actively measure in a system. One such case was proposed by Charles Bennett, calculated by measuring the logical depth in describing the creation of the organism. More precisely, logical depth is the number of steps a Turing machine would take to recreate the original sequence, which is difficult to measure for a large system. Another measure of complexity that holds a richer connotation is the concept of fractional dimension, described as the measure of details at both course-grained and fine-grained scopes of the organism. While an explicit measurement of the complexity of a system is still actively debated, there is no doubt that viruses are complex organisms.

As with all organisms, viruses must abide by the survival of the fittest to stay alive and grow in the population. To do so the organism typically discovers a balance in exploration vs. exploitation of genetic structure to achieve both robustness as well as search for beneficial mutations that would increase its survival. This adaptability can be achieved by the organism following along what is called a *neutral network* [3]–[6]. Neutral networks exploit genetic changes that do not change the phenotype of the organism both to become more resilient as well as increase the likelihood of finding phenotypic variations vital for Darwinian evolution.

As the world pull together to stop the spread of COVID-19 we see individuals and governments begin to respond. We see individuals taking part in walk outs to force large cooperation's to manufacture life saving medical equipment. We see governments imposing shelter in place orders to prevent individuals from spreading the virus. In this situation, similar to [7], there is information flow to and from large institutions (top) and individuals (bottom) that effect our response and therefore the spread of the virus. How information is processed, and travels at various levels of society is as important to understand as how the virus itself spreads through the population.

As individuals and institutions interact with one another around the globe there is always the potential for chaos to erupt. Similar to chaotic dynamics, as we are displaced from our normal routines and stressed to rise to the occasion there is always the potential civil unrest in the form of looting, panic, and a collapse of national infrastructure. The spread of the virus itself, if it were incredibly infectious, could display chaotic population dynamics similar to the logistic map when the growth rate is sufficiently large. However, for the population dynamics of infectious diseases there are far more informative models than the logistic map.

When modeling infectious disease it is important to choose

the right parameters for correct characterization. In epidemiology a critical parameter is R_0, also called the basic reproduction number, which determines the average number of cases generated by a single positive case [8]. Other important features include (but are not limited to) the incubation period and fatality rate. The SIR model and its variations are popular choices for simulating infections over a closed population.

Section II is divided into the methods and results found from exploring two subsets of epidemiological research around the COVID-19 pandemic. First, Section II-A explores the present and predicted evolution of the virus, using a neutral network to track the mutational ability of the Spike glycoprotein (S) in an early case. Afterward, we look deeper into how the disease may be spreading in the population in Section II-B. This is done by investigating the spread of the virus through the population under varying conditions, model assumptions and a genetic algorithm to explore the parameter space. We highlight our key findings in Section III, and conclude by proposing future research (Section IV) that would further develop our current work.

II. Methods and Results

A. Neutral Networks and the Evolution of COVID-19

Most biological systems have the common goal of achieving robustness, efficiency and evolvability, and there are strong hypotheses that this is often accomplished by organisms utilizing the exploration vs. exploitation trade-off. Populations that exploit a common search space form a *neutral network* with nodes representing genomes (RNA sequences) that all fold into the same phenotype (structure). Edges in the network are connections between sequences that differ by a single nucleotide mutation. Exploration occurs when mutations are not neutral, and there is a small chance that the non-neutral mutation is beneficial to the system and persists. Most of the time however, exploration results in deleterious mutations that are quickly removed from the population.

The genome of 29,903 nucleotides that makes up one of the first human COVID-19 cases was sequenced and released for public access in early February 2020 [9]. The spike (S) glycoprotein responsible for the virus binding to human ACE2 cells is of particular interest since it will be necessary to understand for producing a vaccine. However, the virus is a single-stranded RNA virus that is prone to mutation [10]. By modeling the new coronavirus in a neutral network we can begin to understand the possible evolutions that the virus could take, and the potential non-deleterious mutations that could transform the virus into something more infectious, more symptomatic, or even more lethal than the current strain.

To explore the evolution of COVID-19 along a neutral network it is important to understand the vastness of potential mutations that can occur at even one mutation away from a single genome. Starting with the genome collected in [9] that is approximately 30,000 nt long of characters A, C, G, and U, each nucleotide can be mutated to the three remaining nucleotides. Therefore, there are 90,000 genomes that are exactly one mutation away. The whole network contains every variation of the four nucleotides leading to a graph with $4^{30,000} \approx 6.3 \times 10^{18061}$ nodes. The fraction of total mutations that fold into the same phenotype can be approximated by first calculating the probability that a codon will remain neutral. If there are 64 codons that fold into 20 amino acids there are approximately 3.2 codons that are neutral for an amino acid. If the current genome uses one of these codons, the probability that amino acid $AA[k]$ in the genome will remain neutral after mutation is $P(neutral(AA[k])) \approx \frac{2.2}{63} \approx 3.5\%$. If the probability of a genome being neutral to another is equal to the joint probability that every codon in the genome is neutral to the corresponding codons, and mutations are independent events, the probability that a mutated genome of length N is neutral to the original can be expressed by the following equation.

$$\begin{aligned} P(neutral(genome)) &= \bigcap_{k=1}^{N} P(neutral(AA[k])) \\ &= \prod_{k=1}^{N} P(neutral(AA[k])) \\ &= 0.035^N \end{aligned}$$

For a genome with $N = 30,000$ nt this is on the order of $1 \times 10^{-43676}\%$ of nodes. Since this value is incredibly small for any realistic length of genome, it would be unrealistic to randomly search through mutations for neutral strands. Crawling the neutral network provides a practical way of discovering neutral mutations.

To analyze this problem we can make it tractable by looking at the mutations of key amino acids in the spike protein. Characterization of a few mammalian cross-species β-coronaviruses genotypes revealed mutations at critical residues in the spike protein's receptor-binding domain (RBD) [11] between COVID-19 and the original SARS-CoV virus. The five critical residues aligned in each genome are shown in Table I. If we focus our neutral network analysis on these 5 amino acids, the total number of nucleotide combinations is reduced to $4^{15} \approx 1 \times 10^9$. Since a codon is composed of 3 nucleotides and there are 20 possible amino acids that a codon can fold into, there are $20^5 = 3.2 \times 10^6$ possible combinations of amino acids. Another property of the total network is the number of genomes at the k^{th} mutation away from the root. This can be calculated using equation 1 below, such that N is the length of the genome.

$$mutations(k) = \binom{N}{k} 3^k \tag{1}$$

Figure 1 uses this equation to calculate the total number of genomes from 1 to 15 mutations away using the five critical residues to form the root node ($N = 15$).

Without the aid of simulated neutral networks it is possible to estimate the likelihood of the critical residues of COVID-19 mutating back to those found in the original human SARS-CoV. Table II shows the probability of each amino acid mutating in the shortest path possible into the target amino acid. The probability of COVID-19 mutating sequentially back

TABLE I
CRITICAL RESIDUES (CR) OF SARS-COV AND COVID-19

	CR1	CR2	CR3	CR4	CR5
Human-SARS-CoV	Y	L	N	D	T
Civet-SARS-CoV	Y	L	K	D	S
Bat-SARS-CoV	S	F	N	D	N
COVID-19	L	F	Q	S	N

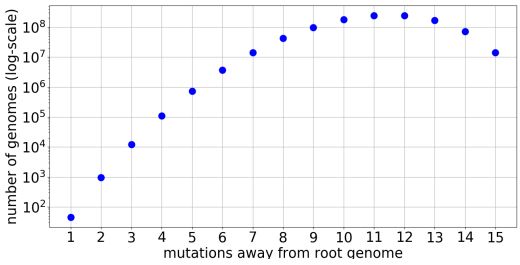

Fig. 1. Number of genomes at each mutation from 5 critical residues of COVID-19

to SARS-CoV could happen in roughly 9 mutations with a likelihood of 0.00001061775%. Based on the hypothesis that COVID-19 mutates twice per month [12], this complete mutation would take 4-5 months.

By implementing a neutral network we can simulate the exploration vs. exploitation evolution that COVID-19 may search while transmitting through the population. The root node genome of a neutral network can be found by translating the amino acid phenotype of the original sequence. This is a one-to-many mapping, with some translations being too distantly mutated to others to be found in a single neutral network cluster. It is because of this disparity that a complete neutral network of a phenotype may involve multiple clusters. The five critical residues of COVID-19 translate into two clusters, and the networks are depicted in Figure 2. All graphs in this section are generated using NetworkX [14]. For all neutral network graphs the root node is shown in black while any node phenotypically neutral to the root is shown in white. There are 288 mutations of the original genome that are silent mutations, i.e. neutral. This shows empirically that the percent of silent mutations out of total mutations is $2.7 \times 10^{-5}\%$. Interestingly enough and perhaps

TABLE II
LIKELIHOOD OF EACH CRITICAL RESIDUE MUTATING BACK TO HUMAN SARS-COV ALONG THE SHORTEST PATH. CALCULATIONS CAN BE FOUND IN OUR APPENDIX [13]

Transition	Probability (%)	# of Mutations
$L \rightarrow Y$	1.3	2.33
$F \rightarrow L$	33	1
$Q \rightarrow N$	2.5	2
$S \rightarrow D$	0.9	2.33
$N \rightarrow T$	11	1
All Events	1.06×10^{-5}	8.66

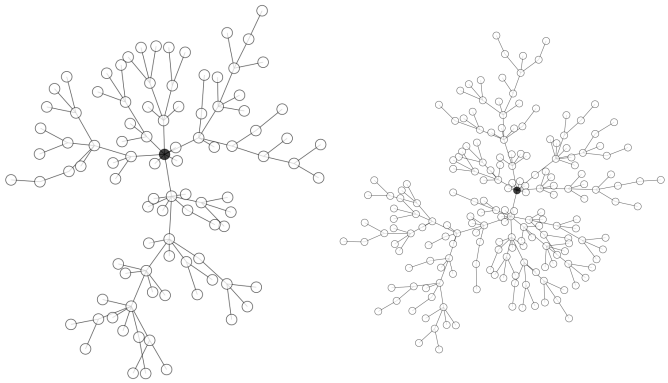

Fig. 2. Complete neutral network of COVID-19. Root node (black) and all silent mutations (white) form two clusters. Neutral network has 288 nodes total.

by coincidence, reverse-engineering our previous calculation estimates $P(neutral(AA[k]))$ to be approximately equal to $\frac{1}{e}$, or 37%. This is much higher than our previously predicted 3.5%, however reasonable since our previous prediction did not take biological probabilities of mutation into account. In genetic structures, mutations are not drawn uniformly as there is a bias for the amino acid to stay neutral. There are also several amino acids that have a much higher statistical chance of remaining neutral.

It is possible yet rare that a neutral network can experience a non-silent mutation that is beneficial to the organism. The majority of the time a mutation will result in a deleterious gene, one that goes against natural selection and does not survive in the population. However, it is reasonable to adjust our implementation to factor in a small $b\%$ of non-silent mutations that persist. Our network made the assumption that no mutations containing the stop protein would be accepted as functional variations. To approximate a realistic value for b we can see that in the four viruses shown in Table I there are 2 or 3 potential amino acids out of the possible 20 that work in each critical residue position. From this we conclude that a value for b between 10% and 15% is practical. For visualization we used a much smaller value for b. Such a simulation is shown in Figure 3, using $b = 1.5\%$ with all non-silent mutations of the same phenotype shown in matching colors. The simulation was stopped after 40 iterations since networks with non-silent mutations grow too large to view graphically.

This neutral network can be used to simulate the likelihood of COVID-19 mutating back into the original SARS-CoV virus. As discussed above, the likelihood of this mutation happening through the least amount of possible mutations is extremely small. However our calculations for the shortest-path mutation do not factor in the biological probabilities of one nucleotide mutating to another. It also does not account for more round-about paths that the virus could take, which was discovered to be possible during simulation. Since such simulations could run for days exploring other paths that might not be the paths leading to original SARS-CoV, we modified our search algorithm to be more parsimonious by limiting the

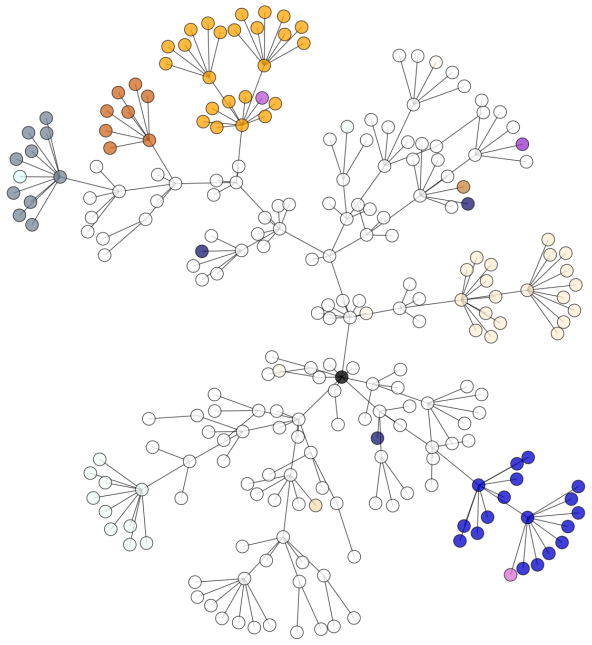

Fig. 3. Neutral network with root node (black) a translation of the five critical residues of COVID-19, with silent mutations shown in white. Any non-silent mutation has a $b = 1.5\%$ chance of being beneficial to the virus and persisting. All non-silent mutations (colored) of the same phenotype are shown in the same colors. Simulation stopped after 40 iterations.

TABLE III
MEAN AND STANDARD DEVIATION OF 40 ITERATIONS CALCULATING SHORTEST PATH DISTANCE FROM COVID-19 TO SARS-COV. ALL RESULTS CAN BE FOUND IN OUR APPENDIX [13].

	$\mu \pm \sigma$ mutations from COVID-19		
b value	Bat-SARS-CoV	Civet-SARS-CoV	Human-SARS-CoV
.10	19 ± 5	23 ± 5	26 ± 5
.15	18 ± 4	21 ± 5	23 ± 5

overall search space. We define a fitness function equal to the number of amino acids correctly mutated to the target genome. The network begins searching normally with an initial fitness of 0 until it functionally mutates into a sequence with greater fitness. At this point it stops searching from any node that has lower fitness than the current maximum. What we call *productive searching* typically results in the network finding a mutation into the target genome, i.e. a mutation with fitness equal to 5. When it finds a match it uses built-in functionality of NetworkX [14] to count the minimum hop distance between the root genome (COVID-19) and the target genome, and terminates early. We performed this search for the three SARS-CoV genomes (bat, civet, and human) for 40 iterations each with $b = .10$ and $b = .15$. The mean and standard deviation of mutations is reported in Table III. To mutate into human SARS-CoV took much longer than the shortest-path mutations calculated in Table II, and if we assume mutations to occur twice per month [12], our simulated results suggest it would take roughly a year to successfully revert.

Genomes in large networks are more robust to mutation than those on smaller networks [3]. Our research has pointed to this correlation between neutral network size and evolvability of the organism, showing that if the organism has more neutral mutations to explore, it has more potential to mutate often thus making it harder to vaccinate against. It also has a higher probability of finding beneficial non-silent mutations. We expanded our research by analyzing genomes of different COVID-19 cases found on GISAID's EpiFlu Database [15]. We extracted the spike proteins from four genomes. The first being a β-coronavirus in a pangolin, thought to be the host that transmitted the virus to humans. The second being one of the earliest recorded cases of COVID-19 found in Wuhan, China. The third and fourth are some of the most recently sequenced, with one in Washington, USA, and the other from England. These two come from the two opposing strains that are hypothesized to be circulating such that the third sample is thought to be more closely related to the original cases in Wuhan.

TABLE IV
SIZE OF THE NEUTRAL NETWORKS FOR 4 SPIKE PROTEINS OF COVID-19

Origin	**Accession ID**	**Collection Date**	**Nodes** ($\log_{10}$)
Pangolin	410721	2019	604.955
Wuhan, China	402120	12/24/2019	608.317
WA, USA	418915	3/24/2020	608.618
England, UK	420250	3/28/2020	608.618

The number of nodes in each neutral network was computed by calculating the log_{10} sum of all potential neutral mutations. For example, the neutral network containing only silent mutations for genetic sequence `AAUGGA` will have every combination of amino acids Asparagine (`N`) and Glycine (`G`). There are 2 codons that encode `N` and 4 that encode `G`, therefore the total size of the network can be calculated by $\log_{10}(2) + \log_{10}(4) = \log_{10}(8)$, and the network has 8 nodes. A manual adjustment was made for Serine (`S`) because it is the only amino acid that results in differing clusters. Results are shown in Table IV with number of nodes shown in a $\log_{10}$ scale. The two most recently collected proteins have the largest networks, suggesting that the virus has mutated in ways that have increased it's robustness and exploration ability.

B. Understanding the spread of COVID-19

In building a Cellular Automaton (CA) to model the spread of COVID-19 we developed a 3-state CA based on the SIR epidemiological model. In the SIR model individuals can have 3 states, susceptible, infected, or recovered (SIR). In constructing our CA we built our grid using a NetworkX graph [14]. While in this work use a simple grid-graph where each node has a Moore neighborhood, in future work we plan to build on this and analyse the spread of the virus on arbitrary graphs.

In building the CA-SIR model we begin by considering the constraints that the model imposes on the rule set governing the cells. From SIR dynamics, if a cell is susceptible (S) it can either stay susceptible, or becomes infected. From an infected state (I) a cell can only recover (R), and once recovered it remains recovered for the remainder of the simulation. From this the only rules to be applied are if the cell is susceptible, otherwise the cell will transition from infected to recovered. Next we consider the total number of permutations a single cell can have. With eight neighbors, each with

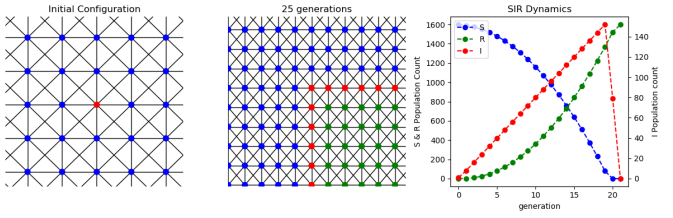

Fig. 4. The left most sub-figure shows a single infected node surrounded by susceptible nodes. The black lines indicate node connections of the graph. The center sub-figure shows the line of infected nodes moving across the graph. Any susceptible node (blue), connected to an infected node (red) will become infected and any currently infected node will become recovered (green). Finally the right most sub-figure shows the population dynamics.

three possible states there are a total of $3^8 = 6561$ possible unique neighborhoods.

At this point we consider the reality of the problem. If you have one infected neighbor, we will argue that the location doesn't matter, only that its infected. This will serve to further reduce the number of possible permutations to 9 possible rules. A cell simply counts the number of infected cells and based on the count determines if it becomes infected or not. With this rule set we will not be able to see any of the interesting dynamics such as in the classic game of life. However, in modeling the spread of disease we are not interested in that class of CA dynamics. Rather we seek to build the simplest model that can provide insight into the real world situation.

For cells with Moore neighbor connections and a rule set outlined above the results are boring but intuitive. With one initially infected cell in the graph the infection simply grows radially outward until the infected cells hit the border. A more interesting dynamic is when the rule changes from deterministic to probabilistic. Figure 5 shows a single infected node (red) in a graph of otherwise susceptible nodes. The probability that a cell transitioned from susceptible to infected (P_I) was 40% for each neighboring infected cell. Then for the number of infected neighboring cells (N_I) the probability that a cell becomes infected is, $P_I(N_I, P_T) = 1 - (1 - P_T)^{N_I}$.

From Figure 5 we see a number of interesting aspects. First, not all of the population became infected as we see a few isolated blue cells. Second we see a clear peak in the number of infected in the population several generation before the virus works its way to boundary, in comparison to the deterministic model. However, because of the statistical nature of the state transition this simple example does not fully illustrate the dynamics of the system. Instead this simulation should be run multiple times to develop a confidence interval for the range of possible dynamics. In Figure 6 we ran

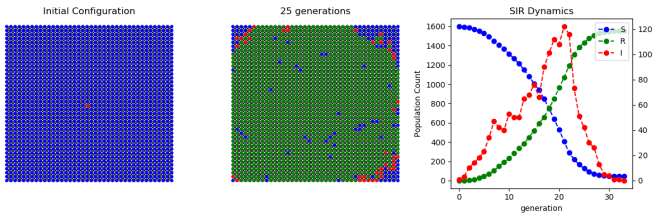

Fig. 5. The left most sub-figure shows the entire graph of 40 by 40 nodes, with a single infected node (red) surrounded by susceptible nodes (blue). The probability that a susceptible node becomes is an independent probability 40% for each neighboring infected node. The center subplot shows the state of the graph after 25 generations. The right most sub-figure shows the number of nodes in any given state.

20 simulations similar to the the simulation done previously with a set probability of transmission, and number of randomly placed

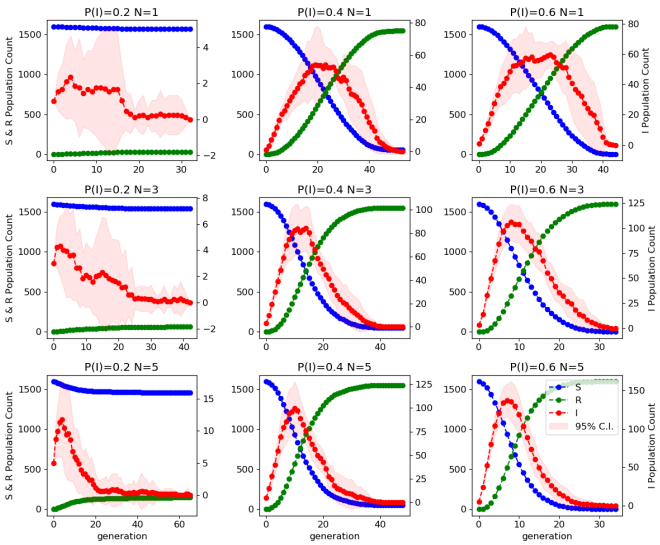

Fig. 6. Each of the sub-figures shows the average of 20 simulations on a grid of 40 by 40 nodes with Moore neighbors. From left to right the probability of transmission ($P(I)$) increases from 20%, 40%, to 60%. From top to bottom, the number of infected (N) individuals randomly placed on the graph increases from 1, 3, 5. The shaded area is the 95 percent confidence interval.

infected individuals. From left to right the probability of transmission is increased from 20%, 40%, to 60%. From the top to bottom the number of randomly placed individuals is increased from 1, 3, to 5. The shaded red area represents the 95% confidence interval for the number of individuals infected for the 20 simulations. The red, blue, and green lines are the average time series populations for the respective states.

After running multiple trials the dynamics of the simulation begin to resemble that of the results from the differential SIR model, particularly for high probability of transitions and initial infected population (bottom right sub-figure). Starting with a single infected individual with a 20% infection rate we see that the the infection fails to effectively spread to the rest of the populations. In fact this is true as the number of initially infected individuals is increased as well. However with increasing number of infected individuals, and higher probability of transmissions we note that more of the populations is infected and that the confidence interval begins to shrink around the average population dynamics. Finally I would like to point out in this figure how the confidence interval is initially narrow and generally begins to widen with each generation. That as time goes on we are less and less certain the amount of the population that will be infected. Next, for evolving two virus (or any number of virus), we began by consider under what conditions are necessary for a single virus to be (continuously) viable in a population. To do this we expand the rule set by adding a probability that a recovered state (R) will return to susceptible (S), which we called the mutation rate $P(S)$. In doing this we decided to modify the initial question to ask what is the maximum sustained population of infected individuals maintainable in a population? We begin by looking at the time series SIR dynamics. Figure 7 shows simulation results similar to the ones done in Figure 6. In this set of simulations recovered nodes transition back to susceptible at some probability $P(S)$. There are several notable differences. First, a sustained population of the virus is possible. Second for this range of parameters we note the system oscillate into an equilibrium after an initial spike in the infected population. In the right most column we see that for the virus with greatest probability of spread also has the highest variation in population over time. From this we developed a genetic algorithm (GA) using

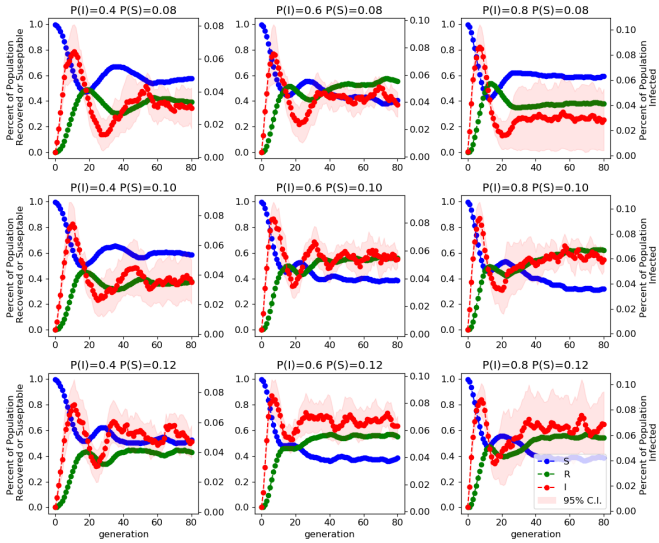

Fig. 7. Each of the sub-figures shows the average of 20 simulations with 5 infected individuals that are randomly placed on a grid of 40 by 40 nodes, with increasing probability of transmission $P(I)$ from right to left, and increasing probability from recovered to susceptible $P(S)$ from top to bottom. Grid is connected by Moore neighbors. The shaded area is the 95 percent confidence interval.

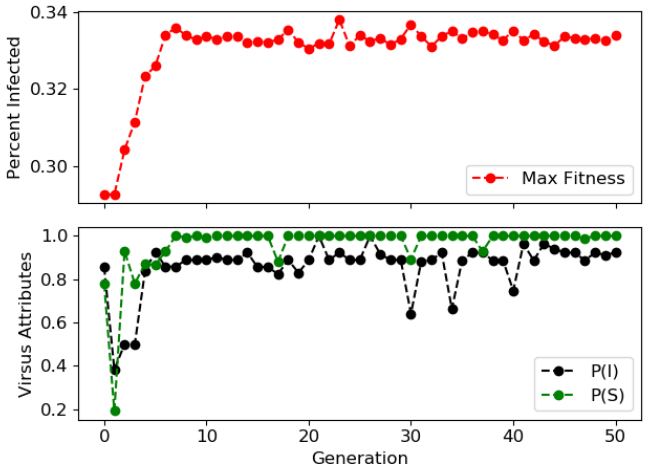

Fig. 8. The top sub-figure is the figure of merit the Genetic Algorithm (GA) is optimizing. This value is average number of infected individuals in a population after 50 generations for 10 independent trials. The bottom sub-figure are the two evolved parameters, the mutation and transmission rate of the individual in the population with the peak fitness. The GA used a population of 10 individual evolved over 50 generations. The individuals (virus) attributes were evolved with 20% probability of mutation. Where a mutation constituted a shift in the transmission or mutation probabilities.

DEAP [16]. The GA had a population size of 10 individuals and was evolved for 50 generations. Each individual had two attributes, a transmission rate, and a mutation rate. The GA had a 50% crossover rate where virus swap probabilities from mutation rate to transmission rate. In addition there was 20% probability of mutation where the probabilities were shifted by a value randomly selected from normal distribution of $\mu = 0$ and $\sigma = 0.125$. The GA evaluates the virus fitness by randomly placing 5 viruses, with the GA evolved mutation and transmission rate, and records the population levels after 50 generations. The average number infected after 10 trials is then evaluated as the fitness score. Furthermore since this fitness score is probabilistic by nature every member of the population gets re-run at every generation of the GA. This prevents any "lucky" runs as members of the population have to continually reprove their fitness. From Figure 8 we find some intuitive and obvious result. First the GA rather quickly converges to an optimal solution (the problem space isn't that large so we shouldn't be particularly surprised). Not surprising that the mutation rate converges to 100%, with everyone constantly susceptible to the virus. Next, the populations reach an equilibrium of equal amounts of the population susceptible, infected and recovered. With some though this also makes intuitive sense, in this simple model individuals continually cycle from susceptible, to infected, to recovered and back. Then maximum that can possibly infected in the population is 33%. Finally, we note that infection rate never reaches 100%, instead it remains around 90%. Looking back at Figure 7 at the lower right sub figure we note that the confidence interval is quite large. By expanding too rapidly the infected population will often collapse, therefore for consistently high infected population we don't want the rate of infection to be set at 100%.

III. Discussion and Conclusions

Our work in Section II-A quantified approximately how small the percentage of possible neutral mutations is, explaining why evolution has led to genomes crawling along neutral paths as opposed to random mutation. We simulated a network centered around one of the early cases of COVID-19, allowing for a small percentage of non-silent mutations to result in beneficial variation (Figure 3). The network quickly grew too large to make decent predictions as to what it would evolve into, but did show the possibility of it reverting to the SARS-CoV virus that caused an epidemic in 2002. Looking at the size of neutral networks centered around both early cases and recent cases of COVID-19 as shown in Table IV suggests mutations that are making the virus more robust, with more exploitation power and more resilience.

Using the SIR epidemiological infection model along with the the principles of cellular automata we were able to show realistic spread of an infection through a population. Unlike the differential SIR model we were able to capture the variation in this model for low numbers of initially infected members of the population, and low transmission rates. With increasing numbers of initially infected individuals and an increasing probability of transmission our model converges to the classic differential SIR model. Furthermore with the introduction of a transition from recovered to susceptible we were able to keep a virus continually viable in our population and evolve the mutation and transmission rate of the virus to infect the maximum number of individuals.

Our results suggest that as COVID-19 becomes more robust, limiting the rate of infection will be critical for manageability as it spreads around the population. In future work we plan to model more precisely the infection rates for different directions of informational flow, by transitioning from a simple grid to an arbitrary graph structure. In addition we would like to expand upon our simple virus by implementing other common virus attributes such as incubation period and recovery time to see how this effects our population dynamics.

IV. Project 3 Proposal

- **Title.** The Spread of Information: Comparing Governmental and Individual Action's Effect on the Spread of Infection
- **Research Question.** Is a top-down or bottom-up flow of information better for slowing the spread of viral infection?
  - **Hypothesis.** A bottom-up approach will produce reduced spread, but is harder to control in a real population.

However, similar to results seen in project 1, it will be a combination of top-down and bottom-up measures that will best slow the spread of infection.

– **Motivation.**
As the COVID-19 virus spreads from nation to nation around the globe there have been varying responses from governments and individuals. How governments and individuals respond to the information they have effects how the virus spreads. Understanding our response and how our response effects the spread is as important as understating the virus itself. There have been top down approaches, such as in China with nation wide lock down, as well as governments closing borders with neighboring countries. We also see a bottom-up effect as individuals practice social distancing (or don't) and limit their travel (or don't).

– **Proposed Research.**
For project three we would like study this top down vs. bottom up approach. At one extreme we can investigate how viruses can spread through a graph where individual nodes make their own decision to cut connections to other nodes similar to how a Cellular Autonoma would operate. We can compare these results to how the virus spreads through a graph when connections are more authoritative on a larger scale. We can then expand this to see what happens to the virus when both methods are balanced over the population.
Pulling from ideas in project 1 we would attempt to use the JIDT tool to measure information flows in our system. Building the model itself we would also not be starting from nothing. For implementing the simulation we would use MESA, a python package for agent based modeling. In addition we would also use NetworkX for the graph structure, and DEAP for any genetic GA implementation.

References

[1] M. Gell-Mann, "What is complexity? remarks on simplicity and complexity by the nobel prize-winning author of the quark and the jaguar," *Complexity*, vol. 1, no. 1, pp. 16–19, 1995. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cplx.6130010105

[2] M. Mitchell, *Complexity: A guided tour*. Oxford University Press, 2009.

[3] A. Wagner, "The role of robustness in phenotypic adaptation and innovation," *Proceedings of the Royal Society B: Biological Sciences*, vol. 279, no. 1732, pp. 1249–1258, 2012.

[4] J. Renzullo, W. Weimer, M. Moses, and S. Forrest, "Neutrality and epistasis in program space," in *Proceedings of the 4th International Workshop on Genetic Improvement Workshop*, 2018, pp. 1–8.

[5] E. Van Nimwegen, J. P. Crutchfield, and M. Huynen, "Neutral evolution of mutational robustness," *Proceedings of the National Academy of Sciences*, vol. 96, no. 17, pp. 9716–9720, 1999.

[6] J. Aguirre, J. M. Buldú, M. Stich, and S. C. Manrubia, "Topological structure of the space of phenotypes: the case of rna neutral networks," *PloS one*, vol. 6, no. 10, 2011.

[7] P. C. Sara Imari Walker, Luis Cisneros, "Evolutionarytransitionsandtop-downcausation," *Artificial Life*, vol. 13, Feb. 2012.

[8] C. Fraser, C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, T. D. Hollingsworth, J. Griffin, R. F. Baggaley, H. E. Jenkins, E. J. Lyons *et al.*, "Pandemic potential of a strain of influenza a (h1n1): early findings," *science*, vol. 324, no. 5934, pp. 1557–1561, 2009.

[9] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei *et al.*, "A new coronavirus associated with human respiratory disease in china," *Nature*, vol. 579, no. 7798, pp. 265–269, 2020.

[10] C. Sun, L. Chen, J. Yang, C. Luo, Y. Zhang, J. Li, J. Yang, J. Zhang, and L. Xie, "Sars-cov-2 and sars-cov spike-rbd structure and receptor binding comparison and potential implications on neutralizing antibody and vaccine development," *bioRxiv*, 2020. [Online]. Available: https://www.biorxiv.org/content/early/2020/02/20/2020.02.16.951723

[11] Y. Wan, J. Shang, R. Graham, R. S. Baric, and F. Li, "Receptor recognition by the novel coronavirus from wuhan: an analysis based on decade-long structural studies of sars coronavirus," *Journal of virology*, vol. 94, no. 7, 2020.

[12] "bedford blog cryptic transmission of novel coronavirus revealed by genomic epidemiology," https://bedford.io/blog/, accessed: 2020-04-6.

[13] "Github repository with source code:," https://github.com/cat-wright/CS523-SARS-CoV-2.git.

[14] P. S. Aric Hagberg, Dan Schult, "Networkx," April 2020. [Online]. Available: https://github.com/networkx/networkx

[15] Y. Shu and J. McCauley, "Gisaid: Global initiative on sharing all influenza data–from vision to reality," *Eurosurveillance*, vol. 22, no. 13, 2017.

[16] F.-A. Fortin, F.-M. De Rainville, M.-A. Gardner, M. Parizeau, and C. Gagné, "DEAP: Evolutionary algorithms made easy," *Journal of Machine Learning Research*, vol. 13, pp. 2171–2175, jul 2012.

V. Contributions

Work done on this project was very evenly distributed between the two group members. Catherine provided the results, analysis and write-up of Section II-A, while Justin provided the results, analysis, and write-up of Section II-B. Both team members took part in writing the abstract, introduction, conclusions, and the project proposal in Section IV. All code used in this project was written by Justin and Catherine or cited appropriately in the paper and commented as such in our programs.

The source code for all results generated in this project can be found at [13].