Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Evolving Extremal Epidemic Networks

Daniel A. Ashlock
Mathematics and Statistics
University of Guelph
Guelph, ON Canada N1G 2R4
dashlock@uoguelph.ca

Fatemeh Jafargholi
Mathematics and Statistics
University of Geulph
Guelph, ON Canada N1G 2R4
fjafargh@uoguelph.ca

Abstract

The susceptible, infected, removed model for epidemics assumes that the population in which the epidemic takes place is well mixed. This strong assumption can be relaxed by permitting the epidemic to spread only along the links of a contact network or *graph*. This study uses evolutionary computation to search for graphs that exhibit one of two extreme behaviors: maximum epidemic duration or maximal number of individuals catching the disease. The focus of the paper is on comparison of two representations for evolvable networks. The first makes local expansions of the network specified by a linear chromosome. The second, a permutation-based representation, joins a large cycle with another cycle specified by the permutation. The linear chromosome representation, based on iterated simplexification, yields inferior results in both fitness measures but creates networks with a structure more like a personal contact network. Location of such behaviorally extreme networks will provide a set of test cases for intervention strategies as well as providing conjectures to focus standard mathematical investigation of the types of networks that yield extreme behavior. This study also proposes a testing protocol for network representations for epidemic modeling.

I. Introduction

The susceptible, infected, removed (SIR) epidemic model is a simple epidemics model that assumes a well-mixed population. It is broadly used in epidemiological research. A pool of individuals is divided into three groups. Those that have not yet contracted the epidemic disease are termed *susceptible*. Individuals that currently have the epidemic disease are termed *infected*. Infected individuals are assumed to be able to infect others with the disease. Those that have had the disease in the past but are no longer able to infect others are termed *removed*. The exact meaning of "removed" depends on the disease, encompassing states as diverse as permanent immunity and death. In the SIR framework, an individual can have a disease at most once; once removed they stay that way.

An SIR epidemic is initialized with all but a few individuals in the susceptible state and those few (in this study one) placed in the infected state. In each time-step of the model each susceptible individual has a chance α of becoming infected for each infected individual in the model. These chances of becoming infected are independent and all infected individuals are assumed to be in contact with all susceptible individuals. After probabilities of infection have been evaluated and newly infected individuals identified, those individuals that were previous infected are moved to the removed state. The epidemic disease is assumed to last for one time step in each individual that contracts it in this study.

One problem with the SIR model is that it does not scale well with the size of the population of individuals on which the epidemic is being simulated. If there are n total individuals and m are currently infected then there is a probability

$$\beta = 1 - (1 - \alpha)^m \tag{1}$$

of any given susceptible individual contracting the disease. As n grows the probability of the epidemic ending on a given time step in which a substantial fraction of the population remains susceptible decays exponentially. This implausible scaling property is a consequence of the assumption that the population is well mixed. A standard method of relaxing this assumption is to place the individuals on a *social contact network* that designates links along which the epidemic disease can spread.

Social contact networks can, with some reasonable set of assumptions, be derived from survey data [8], [3], [5] or may be generated at random. Derivation from survey data is necessarily inaccurate and is also made difficult by the need to respect the privacy of the individuals surveyed. One fact that survey data demonstrates is that in diseases spread by sexual contact the statistics for the number of neighbors at a node in the network has obey a power-law or Poisson distribution. Diseases with airborne spread, on the other hand [4] have contact networks based on shared air space. An airliner or a cubical-farm will create a *clique* of individuals that are all in contact in the disease spread network.

The goal of this study is a preliminary study of evolutionary computation as a tool to search the space of networks for networks that exhibit some form of extreme behavior in a network-limited SIR model. This will bound the range of possible behavior, permitting a valuable comparison with networks derived from contact data. Eventually, the mechanisms by which such networks form must be embedded in the representation of evolvable networks. In [5], for example, it is found that links are more likely to be present between individuals in the same demographic group. In [1] it was found that the links most likely to transmit disease were those crossing demographic boundaries. Sophisticated models of how networks form should not be evaluated in tandem with an initial study of software for evolving networks exhibiting extremal behavior; combining multiple complex systems in an initial study can yield incomprehensible results. This study

thus restricts itself to simple evolvable models of network formation, leaving the incorporation of expert knowledge about the types of networks that occur in natural populations for later.

We now define some useful terminology. In the SIR model the variables S_t, I_t, and R_t denote the number of susceptible, infected, and removed individuals in time step t. The *duration* of an epidemic is the smallest $t > 0$ for which $I_t = 0$.

The terminology of combinatorial graphs is useful for describing networks. Readers desiring a more complete introduction to graph theory can find it in [10]. A *combinatorial graph* or *graph*, G, consists of a set $V(G)$ of vertices and a set $E(G)$ of edges where $E(G)$ is a set of unordered pairs drawn from $V(G)$. Two distinct vertices of the graph are *neighbors* if they are members of the same edge. When drawing a graph, vertices are shown as dots or circles and edges are shown as a Jordan arc joining their members. Examples of graphs appear in Figures 2 and 3. The number of edges containing a vertex is the *degree* of that vertex. If all vertices in a graph have the same degree, the graph is said to be *regular*. If the common degree of a regular graph is k, then the graph is said to be k-regular. Graphs that are 3-regular are called *cubic* graphs. A graph is *connected* if one can go from any vertex to any other vertex by traversing a sequence of vertices and edges. The *diameter* of a graph is the largest number of edges in a shortest path between any two of the vertices. The diameter is, in some sense, the shortest path across the graph.

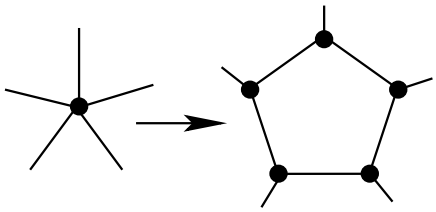

Fig. 1. Replacement of a vertex of degree five with five vertices of degree three; each original neighbor of the degree five vertex becomes a neighbor of one of the vertices of degree three.

All of the topological complexity of networks is available in cubic graphs, but not in networks with lower maximum degree. To see this notice that a vertex of degree $n > 3$ can be replaced with a cycle of n nodes (an example of this is shown in Figure 1) each of which is adjacent to one of the neighbors of the vertex that was replaced. This vertex-to-cycle transformation can turn any graph into a cubic graph while retaining its topological complexity. This does not mean that the resulting graph has the same behavior as a personal contact network; rather it ensures that we have a rich space of connectivities available. Regular graphs are immune to the scaling problems of standard SIR models as each infected individual is able to infect at most a fixed number of other individuals; the probability of infection spreading in no longer exponentially related to the absolute number of infected individuals, rather it depends only on the structure of the network. For this reason, as well as simplicity in an initial study, we choose to use

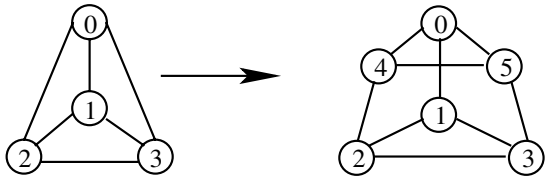

Fig. 2. The basic simplexification move applied to vertex zero of a graph.

representations that generate connected, cubic graphs.

The remainder of this study is structured as follows. Section II specifies the representations for evolvable networks. In Section III the design of the experiments are given. Section IV gives and discusses the results of the experiments. Section V sums up the significance of the results while Section VI outlines next steps.

II. Evolvable Network Representations

There are a large number of ways to represent networks or graphs for evolution. Most of the representations are *incomplete* in the sense that there are many graphs that cannot be specified within the representation. Complete representations, such as a list of edges that could contain any of the possible edges, contain no expert knowledge (e.g. plausible structural bias) about the networks and so tend to pose a very difficult search problem. The representations used in this study are both incomplete and incorporate expert knowledge of one or another sort.

A. Iterated Simplexification

Figure 2 shows an example of *simplexification*. A *simplex* is a collection of vertices that are mutual neighbors. Simplexification of a vertex replaces the vertex with a simplex. The simplex has as many vertices as there were neighbors of the original vertex. Each of these new vertices is a neighbor of one of the vertices that was a neighbor of the original vertex. When a vertex with d neighbors is simplexified then $d-1$ new vertices are created (the vertex that is simplified is retained, vertex 0 in Figure 2). Assuming that the vertices in the original graph are numbered $0, 1, \ldots, n-1$ then the new vertices are numbered $n, n+1, \ldots, n+d-2$. The edges between vertices of the new simplex and the neighbor of the vertex that is simplified are done so that if the members of the simplex are sorted in increasing order then so are their neighbors outside of the simplex. This convention specifies *uniquely* a graph resulting from the simplexification of a vertex.

In order to obtain an evolvable representation we pick an integer bound N larger than the number of vertices in the graph (1000 in this study). The representation consists of a list of numbers in the range $[0..N-1]$. The list specifies successive vertices to be simplexified. As the graph is constructed the list is read from left to right. Each number is reduced modulo the number of vertices currently in the graph to specify the number of a vertex to be simplexified. Each simplexification is performed and the graph structure is updated before the

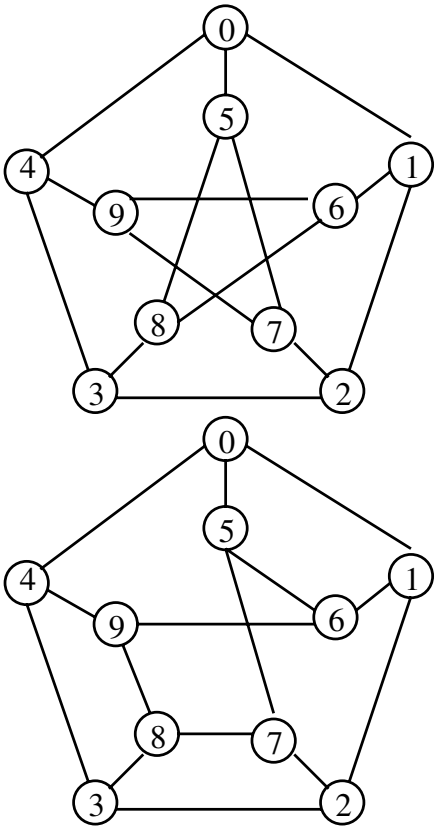

Fig. 3. The generalized Petersen graph with parameters 5,2 and a permutation-generalized Petersen graph for the permutation [5 6 9 8 7] are shown above.

next number is processed. The starting graph is the four-vertex graph shown in Figure 2.

This representation is stored as a string of numbers. Each simplexification adds 2 new vertices to the graph and so, by the end of network development, the number of vertices in a graph is $4+2n$ where n is the length of the string of numbers. The variation operators used in the evolution are two-point crossover and uniform mutation with probability $\gamma = 0.05$. This mutation operator has a probability γ of generating a new number in the range $[0 \ldots N-1]$ at each point in the list.

Two versions of iterated simplexification (ISX) are used. The *full* version uses a list of n numbers to specify n simplexification. The *cyclic* version uses a short list of numbers, cyclically, to make n simplexification. This cyclic method is likely to create graphs with more structural regularity than the full method and also specifies a smaller search space. Since the numbers are still in the range $[0 \ldots N]$ the cyclic method does not simply repeat the same sequence of simplexification over and over. Rather there is an interplay between the size of the graph as it grows and the pattern of simplexification.

B. Petersen Permutation

The *generalized Petersen graph* with parameters n, k is denoted $P_{n,k}$. It has vertex set $0, 1, \ldots, 2n-1$. The vertices

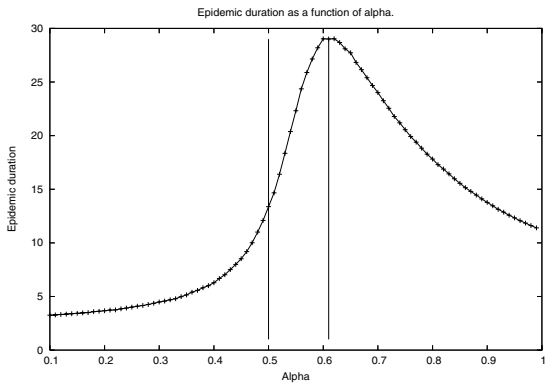

Fig. 4. Epidemic duration as a function of the probability of disease spread α on the generalized Petersen graph $P_{200,1}$. Vertical lines denote $\alpha = 0.5, 0.61$

$0 \ldots n-1$ are connected in a cycle. The vertices $n..2n-1$ are also connected in a cycle but adjacent vertices in this cycle are of the form $i, (i+k);\ (mod\ n)$. Finally, pairs of vertices $i, n+i$ are also connected. An example, $P_{5,2}$ is shown in Figure 3. The *permutation-generalized Petersen graph* (PGPG) replaces the inner cycle, shown as a star in Figure 3, with a permutation of $n \cdots 2n-1$. An example of a PGPG is also shown in Figure 3.

An ordered-gene representation is used for evolving PGPG graphs. The basic data structure is a permutation of the numbers $n \ldots 2n-1$. The PMX crossover operation for permutations [6] is used. This crossover operator choses a position in the permutation. The part of the list before the crossover point are preserved. The part after the list has the same elements as before crossover but they appear in the order that they occur in the other permutation. The mutation operator consists of applying a number of transpositions (exchanges of pairs of elements). This representation for networks is abbreviated PPR.

C. Plausibility of Representations

Graphs produced by the iterated simplexification representations are more plausible as personal contact networks than those produced by the Petersen-permutation representation because they maintain a local structure for the network as it grows. The Petersen-permutation representation can (and usually does) produce graphs with many edges that make long jumps relative to the distances as measured by the outer cycle. As we will see in the results section these representations produce very different searches of the network space as well. The results suggest that plausibility fights with optimality of epidemic duration, an issue that should be considered when using networks in later modeling studies for epidemic intervention.

III. Experimental Design

Two fitness measures for networks are used in this study. The *epidemic duration fitness* (ED-fitness) is the sampled average length of an epidemic in the graph. The *total removal*

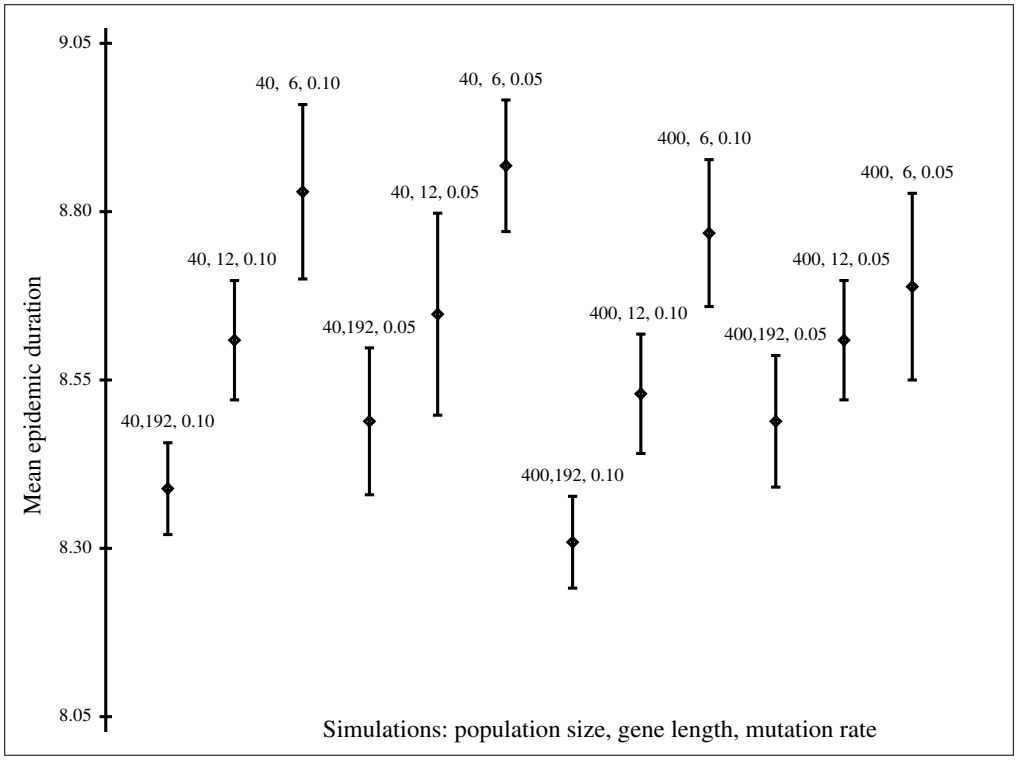

Fig. 5. 95% confidence intervals for the mean epidemic length for the best solutions found in the parameter study for the iterated simplexification representation.

fitness (TR-fitness) is the sampled average number of individuals in the removed category at the end of the epidemic. In all experiments maximizing epidemic duration the probability of an infected individual passing the infection to an adjacent susceptible individual was set to $\alpha = 0.5$. This parameter was chosen to maximize the variability of epidemic behavior because it lies midway between the extreme behaviors of a disease that cannot spread and one that must spread. Figure 4 shows the dependence of epidemic duration, computed by simulation, on the graph $P_{200,1}$. For experiments maximizing the total number of removed individuals, the parameter α was set to 0.61, near the maximum of the curve in Figure 4.

In a cubic graph any α not close to 1.0 dictates that epidemics that do not last long have a nontrivial probability of occurring. Since very short epidemics are not affected much by the structure of the network, epidemics that lasted less than three time steps were discarded in fitness evaluation for all experiments. Since network-limited SIR epidemics are stochastic processes, fitness was estimated by taking the average of 100 acceptable simulations (simulations in which the epidemic duration was at least three time steps).

The iterated simplexification representation is, as far as the authors know, a novel representation presented first in this study. A parameter-setting study was performed for two population sizes (40 and 400), two per-loci mutation rates (0.1 and 0.05), and three variations of the representation (full and cyclic of length 12 and 6) for both fitness functions. The number of simplexification performed in each experiment was 198 yielding a contact network with 400 vertices. The number of vertices in the network is called the *epidemic population size*.

The Petersen-permutation representation uses an ordered gene, a class of representations that have been well studied for problems such as the traveling salesman, bin packing, and scheduling problems[2]. Mutations for this representation consist of exchanging to elements of the list (performing a transposition). A parameter study for the Petersen-permutation representation was performed for both fitness functions using three mutation rates (10, 5, and 1 transpositions) and two population sizes (40 and 400).

Each of the parameter setting experiments consists of 30 runs for each set of parameters studied. A run continues for 100,000 mating events using single tournament selection with tournament size seven. In this model of evolution a group of seven individuals from the population is selected and sorted. The two most fit are copied over the two least fit. The copies are then subjected to the crossover and mutation operator for the run being performed. For the best set of parameters located for each fitness function and both representations larger sets of 100 runs, called *production runs* are performed to search for optimal structures. Examination of the parameter setting runs shows jumps in best fitness in some runs near the end of the run and so the production runs are continued for 250,000 mating events.

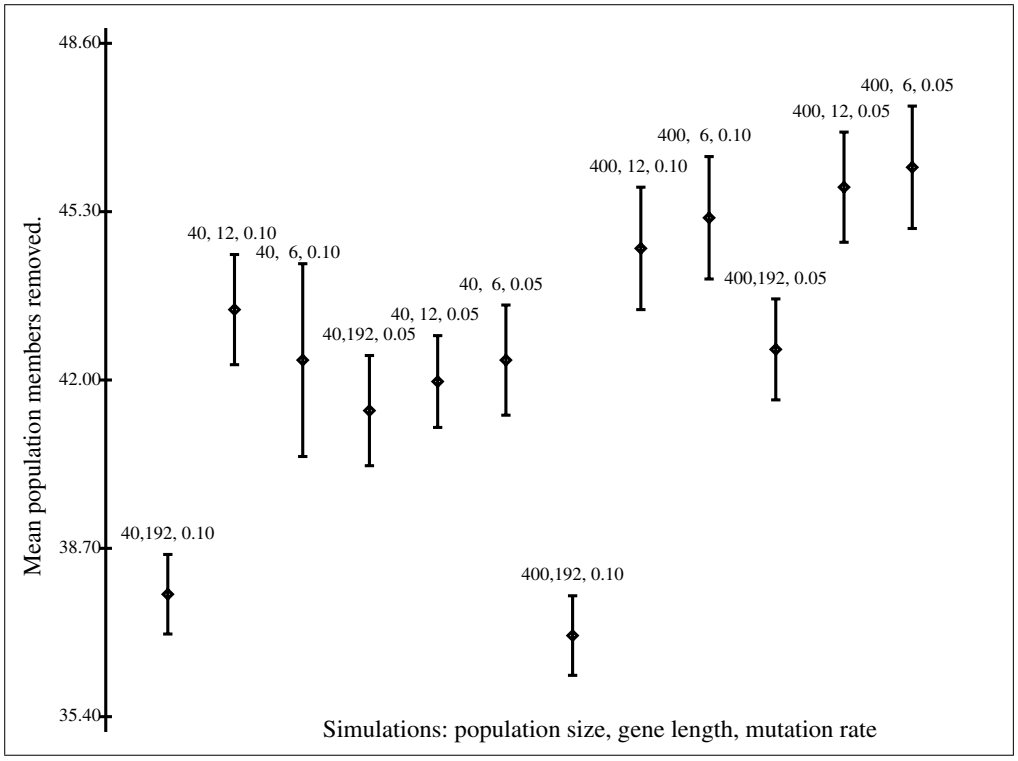

Fig. 6. 95% confidence intervals for the mean number of individuals removed for the best solutions found in the parameter study for the Petersen-permutation representation.

IV. Results and Discussion

The representation that the authors consider more plausible, iterated simplexification, was markedly worse for solving the optimization problem posed by both fitness functions. The initial random populations using the Petersen-permutation representation were more fit than the best optimized structures located with the iterated simplexification representation. This result yields a strong positive on the question of representation having an impact. It also provides evidence that plausibility and optimality may be conflicting goals.

The results of the parameter setting study for iterated simplexification on the ED-fitness function are shown in Table I and Figure 5. The results for the parameter study for the Petersen-permutation representation on the ED-fitness function are shown in Table II. The results of the parameters setting studies for the TR-fitness function and ISX representation are shown in Table III and Figure 6. The results of the parameter setting study on the TR-fitness function for the PPR representation are shown in Table IV.

Parameter setting for the ISX representation on the ED-fitness function shows modest but statistically significant results. Examining Figure 5 we see that a 40-member population with the length-6 cyclic representation and the lower mutation rate is the best and is significantly better than 7 of the other twelve parameter sets tested. There are clear trends the most pronounced of which is that the shorter the representation the better it performs. The lower mutation rate and smaller population sizes are slightly (but not significantly) better.

Population Size	Representation	Mutation Rate	Mean Fitness	Fitness 95% C.I.
40	ISX Full	0.1	8.72	(8.68,8.75)
40	ISX 12	0.1	8.83	(8.78,8.87)
40	ISX 6	0.1	8.94	(8.87,9.00)
40	ISX Full	0.05	8.77	(8.72,8.83)
40	ISX 12	0.05	8.85	(8.77,8.92)
40	ISX 6	0.05	8.96	(8.91,9.01)
400	ISX Full	0.1	8.68	(8.65,8.72)
400	ISX 12	0.1	8.79	(8.74,8.83)
400	ISX 6	0.1	8.91	(8.86,8.97)
400	ISX Full	0.05	8.77	(8.72,8.82)
400	ISX 12	0.05	8.83	(8.78,8.87)
400	ISX 6	0.05	8.87	(8.81,8.95)

TABLE I

Results of the parameter setting study for iterated simplexification while maximizing epidemic length. The numbers following ISX that are not "full" is the length of the cyclic gene used. Mutation rate is per loci. The number of simulations per parameter set is $N = 30$.

The parameter study for the PPR representation on the ED-fitness function shows fewer significant differences than that for the ISX representation; the intermediate mutation rate is

Population Size	Mutation Rate	Mean Fitness	Fitness 95% C.I.
400	10	18.9	(18.8,19.0)
400	5	18.7	(18.6,18.8)
400	1	18.9	(18.7,19.0)
40	10	18.7	(18.7,18.8)
40	5	18.8	(18.7,18.9)
40	1	18.9	(18.7,19.0)

TABLE II

RESULTS OF THE PARAMETER SETTING STUDY FOR ITERATED SIMPLEXIFICATION WHILE MAXIMIZING EPIDEMIC LENGTH USING THE PPR REPRESENTATION. THE MUTATION RATE IS PER LOCI. THE NUMBER OF SIMULATIONS PER PARAMETER SET IS $N = 30$.

Population Size	Represen-tation	Mutation Rate	Mean Fitness	Fitness 95% C.I.
40	ISX Full	0.1	43.2	(42.8,43.6)
40	ISX 12	0.1	46.0	(45.5,46.6)
40	ISX 6	0.1	45.5	(45.0,46.0)
40	ISX Full	0.05	45.0	(44.5,45.6)
40	ISX 12	0.05	45.3	(44.8,45.7)
40	ISX 6	0.05	45.5	(45.0,46.1)
400	ISX Full	0.1	42.8	(42.4,43.2)
400	ISX 12	0.1	46.6	(46.0,47.2)
400	ISX 6	0.1	46.9	(46.3,47.5)
400	ISX Full	0.05	45.6	(45.1,46.1)
400	ISX 12	0.05	47.2	(46.7,47.8)
400	ISX 6	0.05	47.4	(46.8,48.0)

TABLE III

RESULTS OF THE PARAMETER SETTING STUDY FOR ITERATED SIMPLEXIFICATION WHILE MAXIMIZING THE TOTAL NUMBER OF REMOVED INDIVIDUALS. THE NUMBERS FOLLOWING ISX THAT ARE NOT "FULL" ARE THE LENGTH OF THE CYCLIC GENE USED. MUTATION RATE IS PER LOCI. THE NUMBER OF SIMULATIONS PER PARAMETER SET IS $N = 30$.

significantly worse in one comparison, but just barely. The nominally best parameter choice is a population size of 400 and a mutation rate of 10 transpositions and this is the one used for production runs.

The parameter setting study for the ISX representation on the TR-fitness function showed several significant differences. The two cyclic representations using the larger population size and lower mutation rates are the best; the cyclic representation of length 6 was better than the one of length 12 but not significantly. The parameters 400,6,0.05 were chosen for the production runs.

The parameter setting study for the PPR representation on the TR-fitness function showed no significant differences. The parameter set with the best mean was a population size of 400 and a mutation rate of one transposition and so this was used in the production runs.

Population Size	Mutation Rate	Mean Fitness	Fitness 95% C.I.
400	10	287.1	(286.2,287.9)
400	5	286.8	(286.3,287.4)
400	1	287.5	(286.6,288.4)
40	10	287.0	(286.1,288.0)
40	5	286.7	(286.1,287.3)
40	1	287.2	(286.4,288.1)

TABLE IV

RESULTS OF THE PARAMETER SETTING STUDY FOR THE PETERSEN PERMUTATION REPRESENTATION WHILE MAXIMIZING NUMBER OF INDIVIDUALS REMOVED. MUTATION RATE IS THE NUMBER OF TRANSPOSITIONS USED. THE NUMBER OF SIMULATIONS PER PARAMETER SET IS $N = 30$.

Summary data for production runs			
Repre-sentation	Fitness function	Mean Fitness	Fitness 95% C.I.
ISX	ED	9.048	(9.01,9.07)
PPR	ED	19.06	(18.99,19.13)
ISX	TR	50.8	(50.57,51.06)
PPR	TR	288.5	(288.2,288.9)

TABLE V

MEAN FITNESS AND 95% CONFIDENCE INTERVALS ON MEAN FITNESS FOR THE PRODUCTION RUNS FOR BOTH REPRESENTATIONS AND BOTH FITNESS FUNCTIONS. THE NUMBER OF TRIALS IS $N = 100$.

A. Production Runs

The production run for the ISX representation on the ED-fitness found a minimum fitness of 8.79, a maximum of 9.73, and a 95% confidence interval on its mean fitness is $(9.013, 9.077)$. This interval is disjoint from that for the parameter setting run that established the parameters, and so the additional evolution time, 2.5x longer, yielded significantly improved fitness. The distribution of these fitnesses is shown in Figure 7. The significant increase in fitness also appeared in all the other production runs; 95% confidence intervals appear in Table V.

Note that outliers with high fitness appear in the production runs for the ISX representation for the ED-fitness (Figure 7), in the PPR representation on the ED-fitness function (Figure 8) and in those for the PPR representations on the TR-fitness (Figure 10). Such fitness outliers may indicate that he fitness landscape is rough or that the algorithm has, in many of the runs, not stopped improving. This type of outlier was not apparent in the production runs for the ISX representation for the TR-fitness function (see Figure 9) but the distribution has a substantial right tail.

Recall that the *diameter* of a graph is the length of the longest path among those paths in the graph that are shortest paths between some two vertices, a form of "distance across

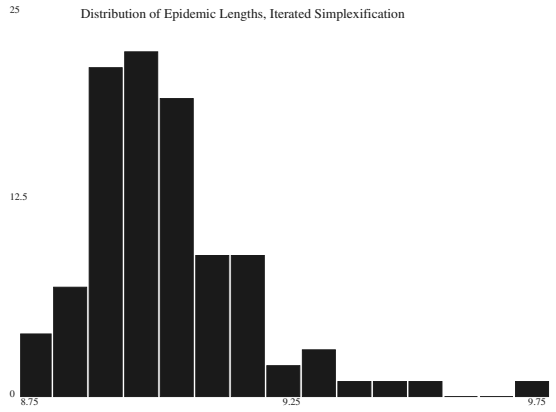

Fig. 7. The above histograms shows how the final finesses were distributed in the production run for maximizing epidemic length using the iterated simplexification representation.

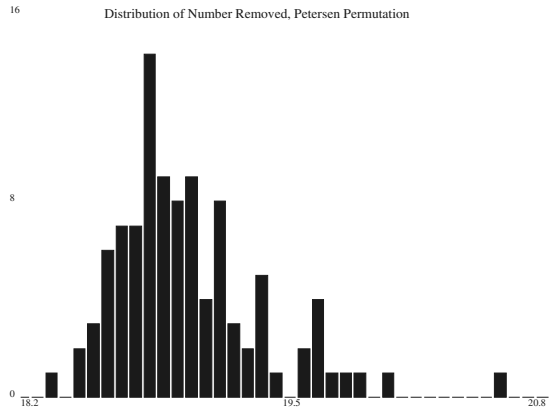

Fig. 8. The above histograms shows how the final finesses were distributed in the production run maximizing epidemic length using the Petersen-permutation representation.

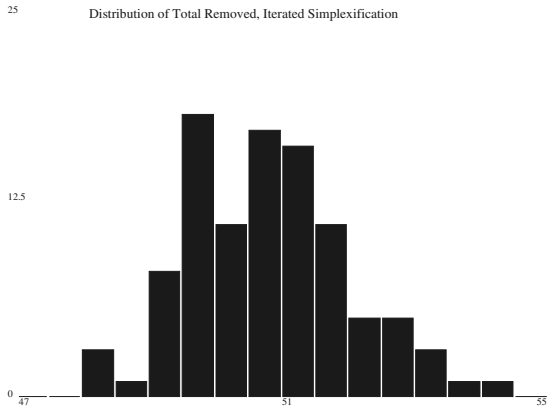

Fig. 9. The above histograms shows how the final finesses were distributed in the production run maximizing total removal using the ISX representation.

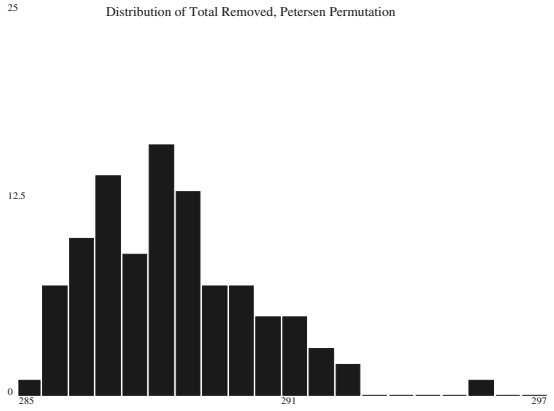

Fig. 10. The above histograms shows how the final finesses were distributed in the production run maximizing total removal using the Petersen-permutation representation.

the graph". If we compute a 95% confidence interval on the diameter of the best graph found in the 100 production runs for both representations and the ED-fitness function then the result for the PPR representation is (9.93,10.1); for the ISX representation it is (24.9,26.5). This is additional evidence that the two representations are exploring completely different parts of network space.

V. Conclusions

The two representations in this study produce markedly different results in optimizing both the ED- and TR-fitness functions. These different results were clearly not because one representation was better at exploring network space; the evidence from the diameters of graph located suggests that they are exploring completely different parts of network space. The portion of network space explored by the PPR representation had far more fit networks in it. To reiterate: for both fitness functions the random structures in the initial populations for the PPR representation were more fit that the best-of-run evolved structures for the ISX representation.

The parameters setting studies were more useful for the ISX representation but were of some value for both representations. A more complete parameter setting study might yield additional benefits both by examining both more values of the parameters that were checked (population size, mutation rate, and length-of-representation for the ISX representation) and by looking at more parameters, e.g. duration of evolution and tournament size for selection. Given the extraordinary difference between the PPR and ISX representations, however, it seems unlikely that any amount of parameter setting will close the gap. This in turn suggests that a search for additional representations is an excellent idea.

The cyclic version of the ISX representation can be generalized for any linear representation in which the individual loci in the gene represent graph construction operators. The ISX representation is a simple version of a *graph grammar* which constructs a graph by applying editing operations to a growing graph. Any graph grammar could use a cyclic string of commands (though it is possible to construct set of operations for which such cycling is vacuous). The cyclic representations have two potential advantages. The first is that they have

exponentially fewer genes and hence an exponentially smaller search space. The second is that a cyclic construction imposes some type of regularity on the resulting graph which, in turn, moves the graph out of the very large region of "random" graphs that dominate network space. This issue requires some additional explanation.

One way of classifying the space of graphs on n vertices is to list the set of $\binom{n}{2} = \frac{1}{2}n(n-1)$ pairs of vertices and say, for each, if an edge is present or not. This yields a simple binary-string representation. The distribution of number of edges in this space is binomial with a quadratic number of possible edges - which means almost all the graphs have about half the possible edges. Personal contact networks, on the other hand, are quite sparse; they lie deeply in the tail of the distribution of number of possible edges. This means that the simple binary-string representation that spans the space represents the desired networks poorly. While it might be possible to make sparseness a secondary fitness goal it seems far more likely that incorporating expert knowledge, in a graph grammar or other structured representation, is a far better way to proceed.

Based on the experience in this study we suggest the following tentative protocol for testing graph representations for locating extremal epidemics.

1) Carefully specify how networks are generated, identifying features representing the incorporation of expert knowledge about natural epidemiological networks. Such expert knowledge is almost certainly required because almost all networks are not as sparse as plausible personal contact networks.
2) A substantial number of random networks should be generated and tested for epidemic fitness measures including but not limited to TR- and ED-fitness. This can exclude representations which manage to squelch all or most of the variability and hence are pointless to evolve.
3) A parameter study or sweep [9] for the representation should be performed. There is substantial room to improve the design of parameter sweeps relative to the ones used in this study. These are full-factorial designs on those parameters the authors intuition suggested would be important. More sophisticated sampling designs[7] would permit the examination of more parameters, e.g. tournament size or length of evolution, or more values of a given parameter such as population size or mutation rate.
4) Production runs should be performed for the best parameters and the resulting best-of-run graphs subject to intensive analysis. This includes comparison with other representations, visualization of the network, and cross-comparison with fitness measures other than those used to produce the graph.

VI. Next Steps

The most important next step is to craft evolvable network representations that incorporate plausible features of natural contact networks. These networks are not typically regular graphs and will be different for different modes of disease spread. The type of networks found in [8] for the spread of sexually transmitted diseases have a connected core featuring a few high-degree nodes and a large number of additional nodes of degree 1, for example. The spread of airborne diseases [4] does not have this type of structure and a good representation here should generate groups of individuals of moderate size that are all in contact, representing shared air spaces.

The maximum number of infected individuals at any time step during the epidemic is another potentially interesting fitness function. The mutual information of this new fitness function and the ED- and TR-fitness functions is an issue that should be investigated; are graphs good at one of these always good at the others? Graphs that are high in one measure but low in another may be a challenging target for a search algorithm but informative about the qualities of a network that are predictive of fitness behavior of networks in general.

Another generalization for this line of research is to increase the sophistication of the epidemic model. The SIR model is quite simple. If can be made more plausible by permitting removed individuals to become susceptible after a delay, by permitting the duration of the infected period to var according to some distribution rather than automatically lasting for one time step. Modeling a latency in the infection process is also a potentially valuable model feature.

Finding good tools to visualize not only the networks but the progress of epidemics across them is another area that would help generate useful intuition about the character of evolved networks. Such visualization may be very helpful in understanding the representations presented in this study and others designed in the future.

References

[1] O. S. Aral, J. P. Hughes, B. Stoner, W. Whittington, H. Hunter Handsfield, R. M. Anderson, and K. K. Holmes. Sexual mixing patterns in the spread of gonococcal and chlamydial infections. *American Journal of Public Health*, 89:825–833, 1999.

[2] Daniel Ashlock. *Evolutionary Computation for Opimization and Modeling*. Springer, New York, 2006.

[3] P De, A E Singh, T Wong, W Yacoub, and A M Jolly. Sexual network analysis of a gonorrhoea outbreak. *Sexually Transmitted Infections*, 80:280–285, 2004.

[4] W. J. Edmunds, C. J. O'Callaghan, and D. J. Nokes. Who mixes with whom? a method to determine the contact patterns of adults that may lead to the spread of airborne infections. *Proceedings of the Royal Society(B)*, 264:949–957, 1997.

[5] K. Ford, W.Sohn, and J. Lepkowki. American adolescents: Sexual mixing patterns, bridge partners, and concurrency. *Sexually Transmitted Infections*, 29:13–19, 2002.

[6] D. E. Goldberg and J. R. Lingle. Alleles, loci, and the traveling salesman problem. In *Proceedings of an International Conference on Genetic Algorithms and their Applications*, pages 154–159. Carnegie Mellon, 1985.

[7] D. R. Huges and F. C. Piper. *Design Theory*. Cambridge University Press, New York, 1985.

[8] A. S. Klovdahl, J. J. Potterat, D. E. Woodhouse, J. B. Muth, S. Q. Muth, and W. W. Darrow. Social networks and infectious disease: The colorado springs study. *Social Science Medicine*, 38(1):79–88, 1993.

[9] Michael E. Samples, Jason M. Daida, Matt Byom, and Matt Pizzimenti. Parameter sweeps for exploring GP parameters. In Franz Rothlauf *et. al.*, editor, *Genetic and Evolutionary Computation Conference (GECCO2005) workshop program*, pages 212–219, Washington, D.C., USA, 25-29 June 2005. ACM Press.

[10] Douglas B. West. *Introduction to Graph Theory*. Prentice Hall, Upper Saddle River, NJ 07458, 1996.