Genetically Modified Network Topologies

Nathan Eagle, Leon Danon and Derek Cummings

MIT Media Lab, 20 Ames St, Cambridge, MA 02319
Dept de Fısica Fonamental, Universitat de Barcelona, Diagonal 647, 08028 Barcelona, Spain
Johns Hopkins University, Baltimore, Maryland 21205
nathan@media.mit.edu
ldanon@ffn.ub.es
derek.cummings@jhu.edu
http://reality.media.mit.edu/gmnets

Abstract. We present a mechanism for constructing networks with a given set of parameters using genetic algorithms. The tunable parameters include number of nodes, number of links, clustering coefficient, entropy and average distance. It is shown that the effects of maximizing entropy while constraining the number of links reproduces an exponential degree distribution, as can be seen in many real networks. We also introduce the concept of the Optimal Network Manifold, a boundary in parameter space that constrains a network's potential characteristics.

1 Introduction

Complex network topologies have received attention from a wide variety of fields in recent years (1–3). For example, the cell is now well described as a network of chemicals connected by chemical reactions; the Internet is a network of routers and computers linked by many physical or wireless links; culture and ideas spread on social networks, whose nodes are human beings and whose edges represent various social relationships; the World Wide Web is an enormous network of Web pages connected by hyperlinks.

Many new concepts and measures have been recently proposed and investigated to characterize such systems. We define and briefly discuss three of the most important concepts:

- **Small Worlds.** The small-world concept describes the fact that in most networks there is a relatively short path between any two nodes, even if the number of nodes is large. The distance between two nodes is defined as the number of edges along the shortest path connecting them. The best known example of small worlds is the "six degrees of separation" found by the social psychologist Stanley Milgram, who showed that there is an average number of six acquaintances between most pairs of people in the United States (4). The small-world property can be observed in most complex networks: the actors in Hollywood are on average within three co-stars from each other, or the

chemicals in a cell are typically separated by three reactions. The small-world concept, however, is not an indication of any organizing principle. Erdos and Renyi demonstrated that the typical distance between any two nodes in a random graph scales as the logarithm of the number of nodes $\langle d \rangle \propto \ln(N)$. Thus, even random graphs are small worlds.

- **Clustering.** A common property of social networks are cliques, circles of friends or acquaintances in which every member knows every other member. This inherent tendency to cluster is quantified by the clustering coefficient (5). Consider a selected node *i* in a network, having k_i edges connected to k_i other nodes. If the first neighbors of the original node were all connected, there would be k_i (k_i - 1)/2 edges between them. The ratio between the number of edges that actually exist between these k_i nodes, E_i, and the maximum number, k_i (k_i - 1)/2, gives the value of the clustering coefficient of node *i*

$$C_i = \frac{2E_i}{k_i(k_i - 1)}. \qquad (1.1)$$

A network's clustering coefficient is the average clustering coefficient of its nodes. In a random graph, since the edges are distributed randomly, the clustering coefficient is $C = p$, where p is the probability of a link existing between any pair of nodes. However, Watts and Strogatz pointed out that in most real networks the clustering coefficient is typically much larger than it is in a random network of equal number of nodes and edges (5).

- **Degree distribution.** Nodes in a network typically do not all have the same number of links, or degree. This variation can be characterized by a distribution function $P(k)$, which gives the probability that a randomly selected node has exactly k links. Since in a random graph the links are placed randomly, the majority of nodes have approximately the same degree, close to the average degree $\langle k \rangle$ of the network. The degree distribution of a random graph is a Poisson distribution with a peak at $P(\langle k \rangle)$. However, recent empirical results show that the degree distributions of most large networks are quite different from a Poisson distribution. In particular, for a large number of networks, including the World-Wide Web (6), Internet (7) and metabolic networks (8), the degree distribution has a power-law tail.

$$P(k) \sim k^{-\gamma} \qquad (1.2)$$

Such networks are called scale-free. While some networks display an exponential tail, often the functional form of $P(k)$ still deviates significantly from the Poisson distribution expected for a random graph.

The discovery of the power-law degree distribution has led to the construction of various scale-free models that, by focusing on the network dynamics, aim to ex-plain the origin of the power-law tails and other non-Poisson degree distributions seen in real systems. The purpose of this work is to explain the observed distributions using the general principle of network evolution rather than a dynamical growth model.

2 Genetically Modified Network Topologies

Genetic algorithms have been used to solve problems in numerical optimization (9), network optimization (10), scheduling (11), circuit design (12) and numerous other disciplines. To our knowledge, genetic algorithms have not been used to construct networks. The present work explores the use of genetic algorithms as a flexible and tunable tool to construct networks with a wide array of characteristics.

The initial population of our genetic algorithm consisted of randomly created adjacency matrices. The random population was created by assigning each index of the adjacency matrix a link with a given probability. Randomly created networks that were not fully connected were discarded and the process is repeated until a full population is created. Our simulations used populations ranging between 20 and 50 individuals.

Mutational Schemes

We explored several mutation schemes:

1) *constant* -the mutation rate is equal for every member of the population
2) *fitness* - the mutation rate is a function of each member's fitness
3) *elite propagation* – the population consists of mutations of only the elite population
4) *hybrid* – a combination of 2 and 3 above, the elites contribute one child each (a mutation of themselves) which they compete with for survival in each round. The non-elite population mutates as in scheme 2.

We compared the performance of these mutation schemes across a suite of several fitness functions. The simulations reported in this paper used they hybrid mutation scheme. Recombination of the adjacency matrices did not yield promising results, however alternative network recombination methods will be explored in further research.

Fitness Functions

There were five components of our fitness function. The average distance, the clustering coefficient, the number of links, the deviation of the number of links from an ideal number and entropy all contributed to the fitness of each network.

Average Distance

The average distance between nodes in a network can be calculated from the network's connectivity matrix A. The shortest distance, *d*, for every pair of nodes (*i, j*) can be calculated by incrementing *d* and noting the first instance $(A^d)_{ij} \neq 0$. The average value over all pairs of nodes, $\langle d \rangle$, is the average distance of the network.

Clustering Coefficient

The clustering coefficient can be derived from the connectivity matrix *A* of the network using the following simple expression:

$$C = \frac{1}{2} tr(A^3) \qquad (2.1)$$

Link Constraints

Two formulations of link constraints were used, the number of links, *M* and $|M - M_{ideal}|$, the difference between the number of links and the ideal number of links.

Normalization

Each component of the fitness function was normalized by subtracting the mean across the population at each generation and dividing by the standard deviation of the component. This was done to eliminate bias of the fitness function toward metrics with higher magnitudes or variances. The contribution of each component to the total fitness was varied for each optimization by multiplying each metric by a coefficient.

Simple Examples

There were several constraint regimes for which the results of our optimization were easily predicted and served as a test for our algorithm. For example, minimizing the average distance should yield optimal networks with $N-1$ links per node, where *N* is the number of nodes, while maximizing the average distance yields sparse graphs with highly separated clusters. Figure 1 presents the optimal networks found under these two sets of constraints. The results presented are the optimal networks after 100 generations.

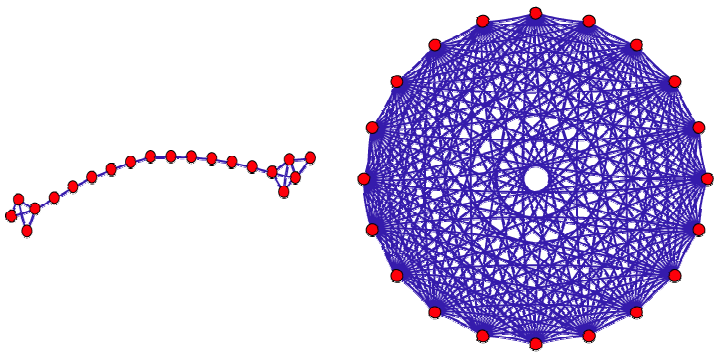

Figure 1: Optimal network under constraint; maximize average distance (left) and minimize average distance (right).

3 Maximizing Entropy

There are several ways of defining the entropy of a network (13). We choose the Shannon interpretation of information entropy:

$$S = \sum_{i=0}^{n} p_i \ln p_i \qquad (3.1)$$

where traditionally n is the total number of states of the system, and p_i is the probability of the system finding itself in state i. For networks, we can consider the p_i to be the probability of finding a node to have degree i. Using this interpretation, the degree distribution $P(k)$ of the network defines its entropy. We can then use the standard Lagrange multiplier method to apply constraints and maximize the entropy. To constrain the number of links we apply the condition

$$\sum_{i=1}^{n} k_i = 2M \qquad (3.2)$$

where $k_{\max}$ is the number of links of the most connected node, k_i is the number of nodes with degree i and M is the total number of links in the particular network. This can be rewritten as

$$\sum_{k=1}^{k_{\max}} P(k) = \frac{2M}{N} = 2\langle k \rangle \qquad (3.3)$$

with N being the number of nodes and $\langle k \rangle$ being the average number of links per node. We then take the derivative with respect to p_k and set it to 0

$$\frac{\delta}{\delta p_k} S = \frac{\delta}{\delta p_k} \sum_{k=1}^{n} p_k \ln p_k + \frac{\delta}{\delta p_k} \left(\sum_{k=1}^{k_{\max}} P(k) - 2\langle k \rangle \right) = 0 \tag{3.4}$$

which gives

$$p_k \propto \exp \frac{k}{\langle k \rangle}. \tag{3.5}$$

Using our genetic algorithm method, we can reproduce the above conditions by building a fitness function that penalizes any deviation from M links and rewards an increase in S. This fitness function has the form:

$$F = S - |M - L| \tag{3.6}$$

where L is the number of links in the modified network. The resulting degree distribution can be seen in Figure 2.

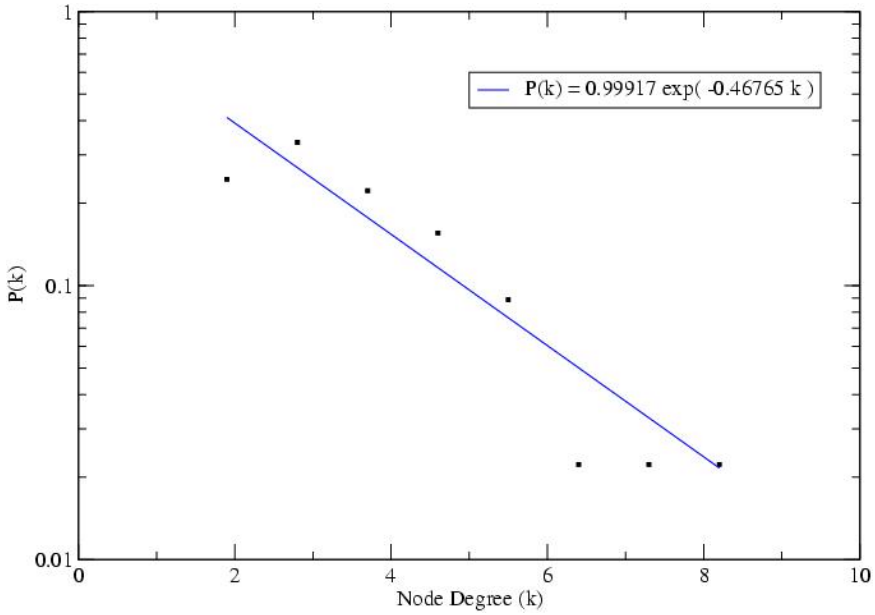

Figure 2: The degree distribution obtained by maximizing entropy while constraining the total number of links in the network

This type of exponential degree distribution can be observed in some social networks (14; 15), and deviates significantly from the Poisson distribution expected for a random graph.

4 Network Manifolds

Despite the significant amount of research on networks, little is known about the interdependencies of a network's characteristics. In Section 2, we have shown simple networks that occur by only optimizing one network parameter; however if we swept across all parameters and stored the very best network for each set of constraints, an interesting structure emerges. A boundary in parameter space is uncovered that provides a hard constraint for all networks topologies. Although many more iterations of the genetic algorithm are needed to fully determine how the boundary manifests itself in this potentially high dimensional space, our initial results give us the main portions of the surface, or Network Manifold, characterized by three network metrics: number of links, clustering coefficient, and average distance.

The implications of Network Manifolds extend beyond the field of theoretical graph analysis. In the design of physical communication networks, both the average distance and the number of nodes are minimized in a tradeoff between cost and performance. As a secondary constraint, clustering is maximized to provide network robustness in specific areas. The optimal configuration strategy has remained elusive and currently networks are designed by what amounts to trial and error. Until now there have been no rules that determine whether a set of network characteristics are even possible. Lastly, after a network has been designed it can be mapped to a unique point on the manifold. The gradient of the manifold at this point corresponds to how quickly one parameter can be improved by slightly varying the other parameters.

Manifold Results

Over the course of evolution within the genetic algorithm, tens of millions of network topologies have been generated and parameterized with average distance, number of links, clustering coefficient and entropy. We have shown that all of these topological configurations lie on a continuous surface in a parameter space consisting of number of links, clustering coefficient and average distance. This implies that we have empirically uncovered the governing equation that constrains the dependencies of these parameters. The size of the network does not appear to play a role in the shape of the manifold. Although much of our research has been on graphs consisting of 20 and 50 nodes, our initial results suggest that the manifold will scale with nodes, and thus our findings are applicable to much larger networks as well.

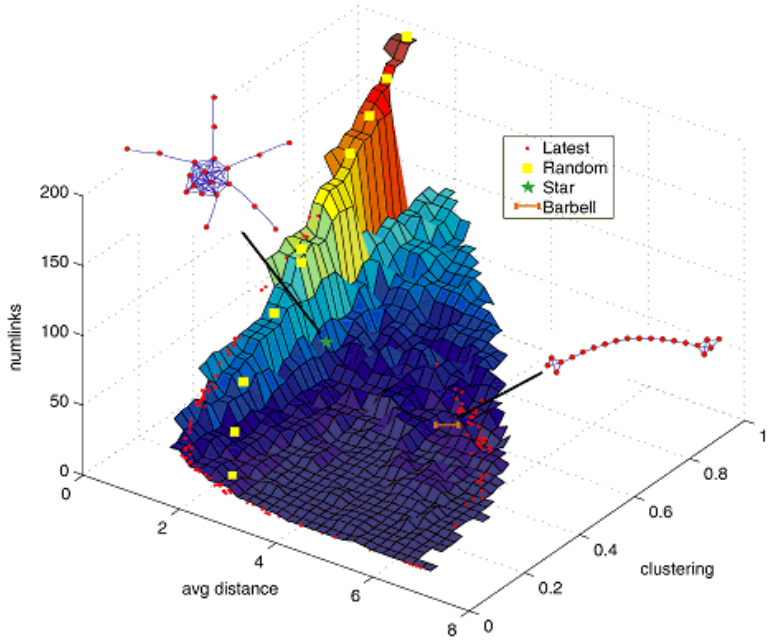

Figure 3: The Network Manifold governing the relationship of three parameters collected by generating over two million network topologies given a variety of constraints.

In Figure 3 we have plotted the manifold and highlighted several distinctive networks. The red dots are the latest networks generated by the genetic algorithm. The yellow squares are random networks with varying link probability. The green star represents a star network while the orange barbell is the barbell graph shown in Figure 1.

Towards the Optimal Network Manifold

The Network Manifold we present should not be considered complete. As the algorithm runs, some sections slowly continue to expand. These portions of the manifold represent the discovery of new network topologies that better meet the dynamic constraints imposed by the genetic algorithm. Although a solution generated by a genetic algorithm can never be considered 'optimal', it is reasonably sure that some areas of the manifold will remain static, allowing us to place our computational resources on areas that are more likely to grow. We do this by initializing the algorithm not as a population of random networks, but rather as networks located near the border we wish to expand. By 'seeding' these potential networks, while lowering the mutation rate and increasing number of iterations, our generic algorithm slowly is able to expand these boundaries and continue its exploration of the manifold.

5 Future Research and Conclusion

Using genetic algorithms, we have developed a method of generating network topologies with a set of desired characteristics. This methodology has enabled us to begin understanding which combinations of network characteristics are possible. However, the manifold continues to expand. Meanwhile, our own generic algorithm is evolving to more efficiently search this parameter space for a diminishing number of new topological solutions that better meet the constraints imposed by the fitness function.

We have shown that using genetic algorithms to maximize the entropy of a network while keeping the number of links constant generates an exponential degree distribution, similar to many real networks. We have also introduced the concept of the Network Manifold and shown how it characterizes optimal network topologies and constrains possible network characteristics. It is our hope that this work will inspire additional research on network design and optimization.

Acknowledgements

The authors would like to thank the organizers of the SFI Summer School, especially Jonathan Shapiro and Tom Carter. They also thank Aaron Clauset for the original code and Mickey for moral support.

References

[1] M.E.J. Newman, *The structure and function of complex networks.* SIAM Review **45**, 167-256 (2003).
[2] Reka Albert, Albert-Laszlo Barabasi, *Statistical mechanics of complex networks.* Reviews of Modern Physics **74**, 47 (2002). 5.
[3] S.N. Dorogovtsev and J.F.F. Mendes, *Evolution of networks.* Advances in Physics **51**, 1079-1187 (2002).
[4] Milgram, S., Psych. Today *The small world problem*. **2**, 60 (1967).
[5] Watts, D. J. and S. H. Strogatz, *Collective dynamics of 'small-world' networks.* Nature **393**, 440 (1998).
[6] Albert, R., H. Jeong and A.-L. Barabasi, *Diameter of the world-wide web.* Nature **401**, 130 (1999).
[7] Faloutsos, M., P. Faloutsos and C. Faloutsos, *On Power-Law Relationships of the Internet Topology.* Proc. ACM SIGCOMM, Comput. Commun. Rev. **29**, 251 (1999).
[8] Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai and A.-L. Barabasi, *The Large-Scale Organization of Metabolic Networks.* Nature **407**, 651, (2000).
[9] Holland, J.H. Hidden Order: how adaptation builds complexity. Addison Wesley. (1995)
[10] Walters, G.A., Lohbeck, T. *Optimal layout of tree networks using genetic algorithms.* Engineering Optimization **22**, 27-48. (1993).

[11] Alcaraz, J., Maroto, C., Ruiz R. *Solving the multi-mode resource-constrained project scheduling problem with genetic algorithms*. Journal of Operational Research Society. **54**, 614-626. (2003)
[12] Lohn, J.D., Colombano, S.P. *A circuit representation technique for automated circuit design*. IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION. **3**, 205-219. (1999)
[13] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin, *Principles of Statistical Mechanics of Random Networks*., cond-mat/0204111.
[14] R. Guimera, L. Danon, A. Dýaz-Guilera, F. Giralt and A. Arenas, *Self-similar community structure in organizations.* cond-mat/0211498.
[15] P. M. Gleiser, L. Danon, *Community structure in Jazz* cond-mat/0307434.