Nature Inspired Computing Project

Finding a subset of nodes in a graph that are highly robust

Meenusree Rajapandian
0030164747

I. Introduction

A number of real-world complex systems can be represented as networks. Social structures, power grids, protein interactions, transportation system, internet, ecological food chain are some of the fields in which networks are widely used. In each of these fields, a network, represented by a graph with nodes and edges, is used to model the structure of interactions such that the nodes represent objects or agents that can undergo a change, receive or relay information, or be states of rest/stop and the edges represent some meaningful interaction or connection between these nodes. Depending on the meaning given to these nodes and edges, different methods of analysis are used to understand the topology and the dynamics of these models.

A lot of the analysis however goes in understanding the underlying structure of the graph. The structure of the graph drives key dynamics of the network and a lot of research has gone into quantifying and measuring the structure of the network. The aim is to facilitate comparison between networks and also helps assessing possible changes that can be brought in a network to direct it towards a desirable model. Some of these changes might be to stop an epidemic from spreading in a network, enhance spread of information, maximise influence in a network or make the network more reliable in the events of failures. A measure that is used as a blanket term in all of these applications is Robustness.

The definition of robustness is still vague and open to interpretation. [1] Intuitively, robustness is the ability for the network to continue to perform well when it is subject to failure or attacks. In an epidemic model, the spread of virus is attack and a robust network would prevent the spread of virus to all the nodes. In an information spread or maximising influence model, the attack might be disconnection of nodes or edges and a highly robust network is also one that is reliable under attack. In this work, the latter idea of robustness is taken. Here, a robust network might have higher connectivity, lower diameter, etc. [2], [3]

Although measuring the robustness of a network as a whole may be a measure of interest, not all parts of a network are equally robust or contribute equally to the total robustness of a network. Understanding the structure of the graph is important and finding a sub-graph that is most robust helps in understanding how information might spread or what parts of the network are more prone to epidemics and where initiatives might be taken to control them. [4]

II. Method

The general outline of the project is to use genetic algorithm to find a subset of nodes that are highly robust. The population of the genetic algorithm is a binary representation of the nodes that are elements of the robust subgraph. Each population is a binary string of size n, where n is the number of nodes and a bit being 1 means the corresponding node is selected for the sub-graph. A solution is feasible only when the sub-graph is connected.

Single point crossover and uniform mutation are used in this method. During the analysis, it was found that the solution and the number of fitness function calls is highly dependent on the initial population. Probability of including a node in the sub-graph is taken as a bernoulli random variable p which is also taken as a parameter in the algorithm. A simple tournament selection is done where two elements of the population are the chose and the winner of the tournament is chose to mate. Two such parents are chosen at every iteration and two children are produced.

Since there are no ground truth solution for most networks, we use a robustness measure of interest to find the subset of nodes in a network that maximises this measure and compare it with when all the nodes in a network are used and also other measurements or robustness. This is repeated for some of the robustness measurements found in [1]. This comparison is done a for a number of network models and the results are reported in this project.

Three objective functions of interest are used to find the sub-graph that is most robust and also consequently assess the sub-graphs for comparison. The functions are defined below for an undirected graph $G = (V, E)$ where $V = \{v_1, v_2, \dots, v_n\}$ are the set of nodes and $E = (v_i, v_j)$ are the edges when v_i and v_j are connected by an edge.

The clustering coefficient of a node i is

$$CC_i = \frac{2t_i}{k_i(k_i - 1)}$$

where $t_i = \frac{1}{2}\sum_{j,h \in V}(w_{ij}w_{ih}w_{jh})^{1/3}$ is the geometric mean of triangles around node i. The mean clustering coefficient across all nodes of the sub-graph is considered as one of the objective functions

Hence

$$OBJ1 = \sum_i \frac{CC_i}{n}$$

where n is the number of nodes in the sub-graph and $i \in$ set of nodes of the sub-graph.

Another objective function that is used is the diameter of a (sub-)graph. The diameter of a set of nodes in a graph G is the maximum of the shortest path across all pairs of nodes. Since graphs with larger number of nodes may have a higher diameter, it is normalized with the number of nodes for fair comparison.

$$OBJ2 = \frac{\max_{ij} SP_{ij}}{n}$$

where n is the number of nodes in the sub-graph and SP_{ij} is the shortest path length between a pair of nodes i and j.

Note that although other objective functions are maximised, the normalized diameter of the sub-graph is minimized and hence a negative of the above function is used in the genetic algorithm.

The mean communicability of all pairs of nodes is another objective function that is assessed. Since communicability counts the number of walks between two pairs of nodes, it is a very good measure of robustness. A pair of nodes will have high communicability if there are a number of different walks compared to other pairs of nodes. However, since more nodes increases the chances of higher number of walks of all lengths, the mean communicability across nodes is normalized with the number of nodes n for fair comparison. Hence

$$OBJ3 = \frac{\sum_{ij} C_{ij}}{n^2}$$

where $C_{ij} = \exp A_{ij}$ is the communicability of between two nodes i and j and A is the adjacency matrix.

Every time the genetic algorithm is run on the graph, the population gives a set of solutions that are all unique. Only the best solution is taken at every run. This is done 25-30 times for every graph and the nodes that occur in more than 50% of the best solutions are considered the solutions for this graph. The measurement comparison however, are done as a mean of the best solution across these 25-30 repetitions.

III. Results

Barbell graphs, cycle graphs, star graphs, wheel graphs, small world and Barabasi preferential graphs are used in this method. Along with these networks, common real word networks such as karate club and the dolphin network are also assessed in this project. The table below gives a summary of the different measures when different objective functions are used. It is to be noted that the mean across best solution at every run of the genetic algorithm is reported. The best of the best solutions will have much better results than those reported below.

It can be seen that while the measuring corresponding to the objective function is maximised (or minimized in the case of diameter), the other measures are only marginally minimised. However, when finding a subset of nodes with high communicability, the other measures also improve considerably.

The barbell graph consists of two complete sub-graphs connected by a chain of nodes each with a degree 2. The genetic algorithm consistently finds at least one of the complete graphs as the best solution for the most robust subset of the graph.

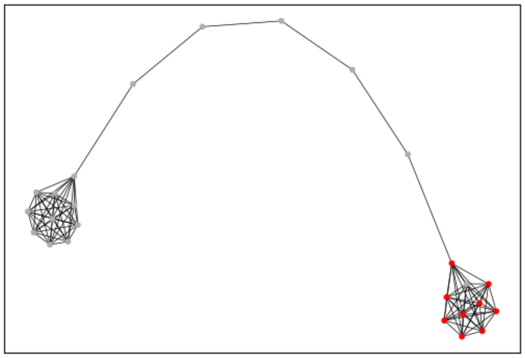


Fig. 1. Barbell Graph. Set of nodes in red is the best solution obtained

For a cycle graph, only two to three nodes that are connected are ever chosen in any of the members of the population. This is because of the very structure of the graph where the possibility of disconnection in the event of attack is very high. This result was found for cycle graphs of size 20-30 nodes and even when starting with a population that has bigger sub-graphs.

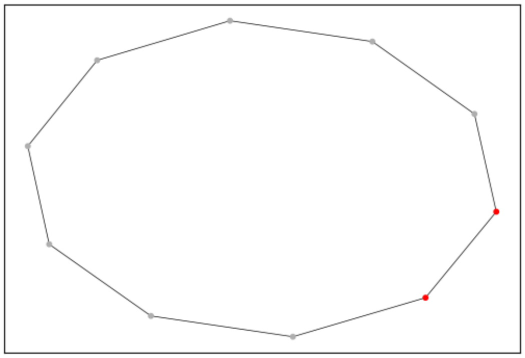


Fig. 2. Cycle Graph. Set of nodes in red is the solution obtained

The Small World and Barabasi-Albert graph models have very interesting results depending on the number of nodes and the parameters corresponding to each model. For the small world graph of size 20 nodes, when attached to 5 neighboring nodes and with probability of 0.5 to be reconnected, almost all the nodes are selected in the sub-graph as can be seen in 7. For a graph of 10 nodes, when attached to 3 neighboring

TABLE I

Summary of fitness values as a result of the genetic algorithm on different graphs. The columns represent the measures of the sub-graphs that were the solution. The rows are for different graphs, when different objective functions are used. Note that the diameter is tried to minimize while the other measures are maximised.

	Objective Function	Clustering Coefficient	Diameter	Communicability
Barbell Graph n = 24	Full Graph	0.784	0.32	10.748
	Clustering Coefficient	0.797	0.318	6.936
	Diameter	0.827	0.280	14.8523
	Communicability	0.6533	0.43443	16.16
Cycle Graph n = 10	Full Graph	0	0.5	0.074
	Clustering Coefficient	0	0.764	0.276
	Diameter	0	0.5	0.67
	Communicability	0	0.5	0.6795
Star Graph n = 11	Full Graph	0	0.18	0.154
	Clustering Coefficient	0	0.344	0.236
	Diameter	0	0.199	0.16
	Communicability	0	0.48	0.69
Small World Graph n = 20	Full Graph	0.225	0.2	0.173
	Clustering Coefficient	0.614	0.457	0.195
	Diameter	0.186	0.202	0.159
	Communicability	0.336	0.429	0.548
Barabasi Albert Graph n = 20	Full Graph	0.685	0.1	113.25
	Clustering Coefficient	0.746	0.321	1.767
	Diameter	0.692	0.106	76.7
	Communicability	0.699	0.115	89.516
Karate Club n = 20	Full Graph	0.5706	0.147	0.527
	Clustering Coefficient	0.792	0.279	0.472
	Diameter	0.547	0.141	0.491
	Communicability	0.529	0.225	1.064

nodes initially and 0.5 probability of change, only two nodes are chosen as part of the sub-graph with higher robustness. The reason for it is obvious in 8.

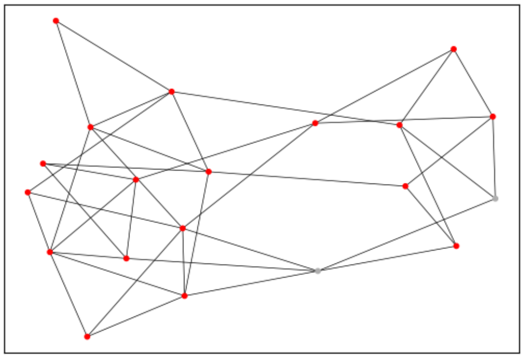


Fig. 3. Small World graph. n=20, k=5, p=0.5 Set of nodes in red is the solution obtained.

A similar solution set can be seen for the Barabasi-Albert model where for 20 nodes and preferential attachment of 8, all the nodes are selected as the sub- graph with highest robustness across all objective functions. However, for the same graph type with 20 nodes and preferential attachment 3, a smaller subset of nodes is chosen which has high robustness.

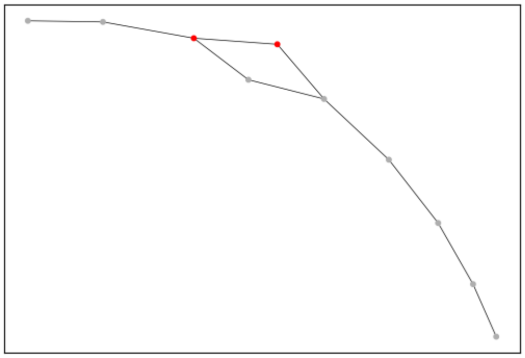


Fig. 4. Small World Graph. n=10, k=2, p=0.5 Set of nodes in red is the solution obtained

The karate club had interesting results. A little background on the model data gives important insights in the network. The members of the karate club, after a feud between two members split into two forming the ‘ground-truth’ communities in the network. However, this also shows that the network of each community is robust and that is the reason for the network being split into two. The genetic algorithm on this model always gives a subset of nodes that belong to only one of the communities. Not all the nodes of a certain community is part of a single solution and this could be because those nodes are not as strongly attached to the community as the other nodes are.

The fitness value distribution and number of fitness function calls made are shown below for maximising only communicability in different networks.

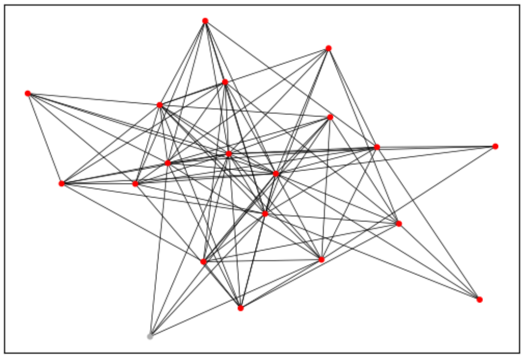


Fig. 5. Barabasi-Albert model. n=20, m=8 Set of nodes in red is the solution obtained.

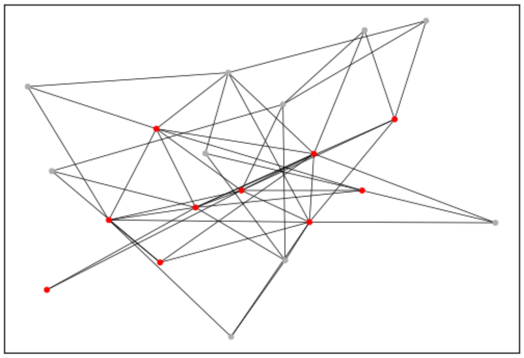


Fig. 6. Barabasi-Albert Model. n=20, m=8 Set of nodes in red is the solution obtained

IV. Did it work?

The solutions are not very consistent and reliable across the different runs and for different parameters of the same model. The variance of fitness value of the best solution is small only for graphs that do not really have a subset of graphs that are robust (eg cycle and star graphs). The fitness function calls are high even for graphs of size 10-20 which gets higher with bigger graphs. The solution space as a result of the number of graphs will vary more as the size of the graph increases. For the barbell graph, only one complete subgraph was ever obtained as the solution when the genetic algorithm was run. An ideal case would be to have at least two members of the population that represent at least most of the two complete sub-graphs.

All results shown above were for a crossover probability of 0.8 and mutation probability of 0.1. A change in any/both of the parameters only results in different range of fitness function calls made and the number of generations it takes for convergence. There is very little to no difference in the actual solution obtained.

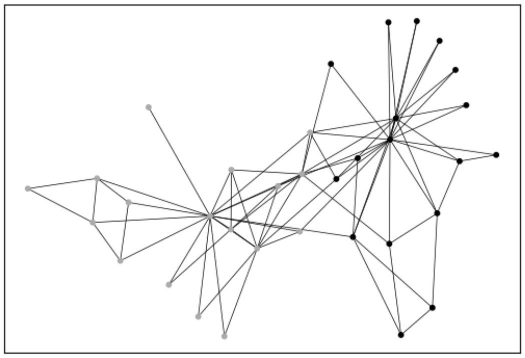


Fig. 7. Karate Club. Node color based on communities that they belong to.

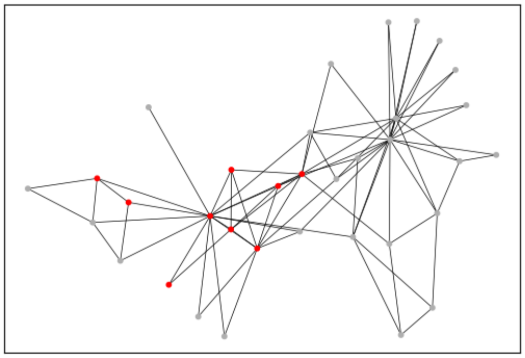


Fig. 8. Karate Club. Set of nodes in red is the solution obtained

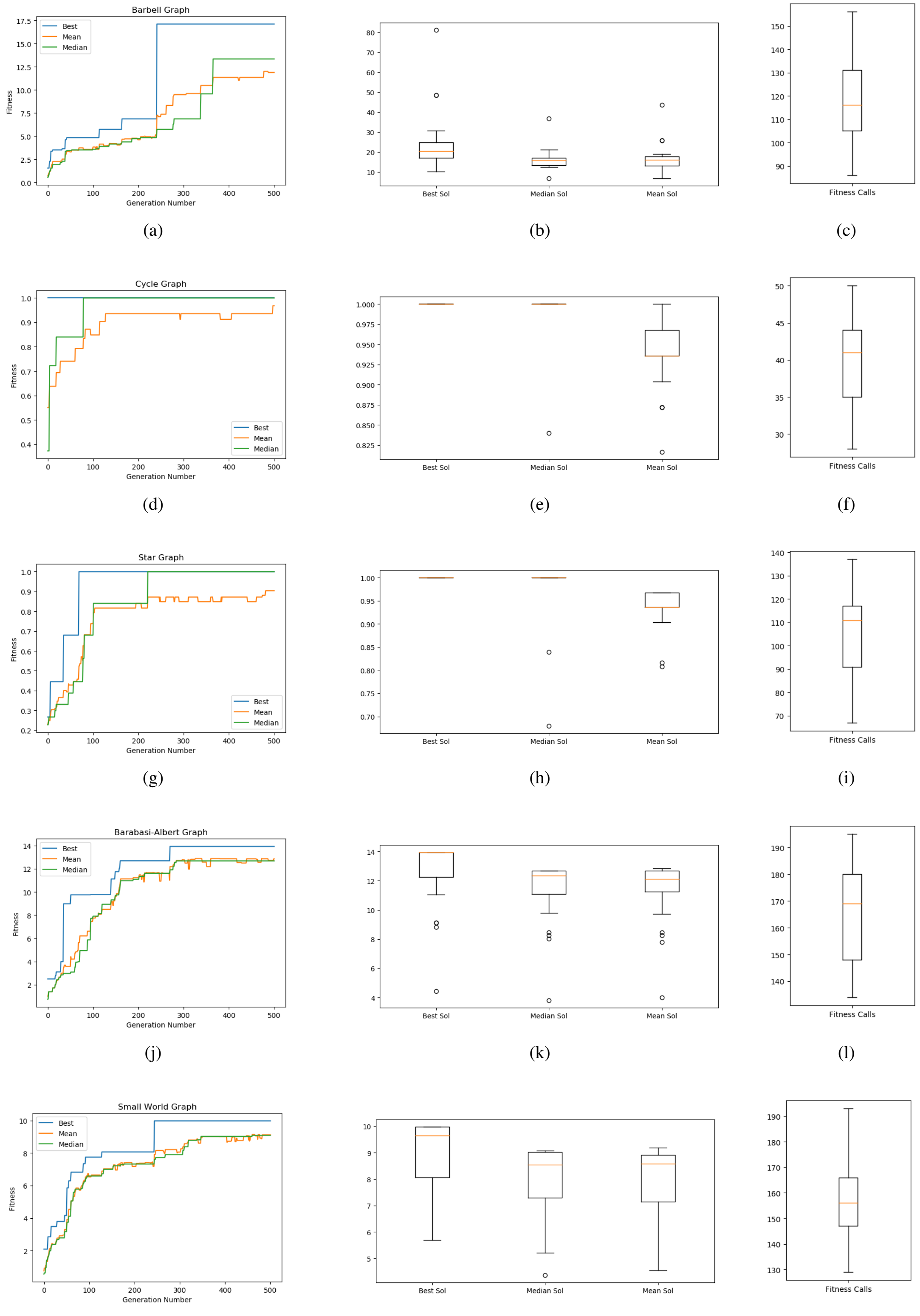
A difference in the population initialization method did not give better results either. Only a change in the fitness function calls made and the generations required for convergence changed.

A diverse exploration of different methods of selection, crossover and mutation methods were not done. Much of the work went into finding objective functions that can be used to find and assess robustness in graphs.

TL;DR Not Really

V. Future Work

Although the results obtained from this assessment are not great, it paves a foundation upon which other variations of heuristic algorithms and even what robustness means in a graph can be explored. Future work may include and is not limited to, using other methods of selection, non-uniform mutation methods and especially different objective function. Since finding a subset of graph using a certain objective function is computationally difficult, genetic algorithms give an easier way to assess and explore different objective functions that can represent robustness in a network.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

(m) (n) (o)

Fig. 9. A summary of fitness function values over generations, solution distribution for best, median and mean over 25 runs and fitness function calls distribution over 25 runs. Each row represents a type of model used as mentioned in the table

Another important aspect of the genetic algorithm that can be explored especially for this problem is how to handle replacement in a population. In this project, a child becomes a part of the population only when that particular representation doesn't always exist in the graph. This is done so that there is no 'convergence' to one particular solution and a preferable result would be a set of sub-graphs that are highly robust. Other methods need to be explored to have a set of solutions without having to repeat the genetic algorithm multiple times.

A number of other ideas came up through the course of this project. Finding communities based on modularity or communicability are computationally intensive problems. Genetic Algorithm can be used to find communities in a network and if a similar implementation is done, it can also be used to find overlapping communities in a network unlike Louvain's or Newman's methods of community detection. [5]–[7] Finding cliques in a graph is another intensive problem that can be explored with genetic algorithm. However such a search might be like finding a needle in a hay stack and might require creative methods of solution representation, crossover and mutation to find the solution.

References

[1] W. Ellens and R. E. Kooij, "Graph measures and network robustness," *arXiv preprint arXiv:1311.5064*, 2013.

[2] M. Youssef, R. Kooij, and C. Scoglio, "Viral conductance: Quantifying the robustness of networks with respect to spread of epidemics," *Journal of Computational Science*, vol. 2, no. 3, pp. 286–298, 2011.

[3] A. Jamakovic, R. Kooij, P. Van Mieghem, and E. R. van Dam, "Robustness of networks against viruses: the role of the spectral radius," in *2006 Symposium on Communications and Vehicular Technology*. IEEE, 2006, pp. 35–38.

[4] H. Chan, S. Han, and L. Akoglu, "Where graph topology matters: the robust subgraph problem," in *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM, 2015, pp. 10–18.

[5] H.-J. Li, H. Wang, and L. Chen, "Measuring robustness of community structure in complex networks," *EPL (Europhysics Letters)*, vol. 108, no. 6, p. 68009, 2015.

[6] M. Gong, L. Ma, Q. Zhang, and L. Jiao, "Community detection in networks by using multiobjective evolutionary algorithm with decomposition," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 15, pp. 4050–4060, 2012.

[7] C. Shi, Z. Yan, Y. Wang, Y. Cai, and B. Wu, "A genetic algorithm for detecting communities in large-scale complex networks," *Advances in Complex Systems*, vol. 13, no. 01, pp. 3–17, 2010.