Evolution of Clusters

Uday Gupta (2019CSB1127)
Aman Palariya (2019CSB1068)
Sagalpreet Singh (2019CSB1113)

November 2020

Abstract

The project aims at simulating the evolution of communities in general social network. We model this on the basis of various parameters including but not limited to how open a person is to make new connections, what effect does time have on friendships/relations, how are people influenced by their connections. We model all of this mathematically and run the simulation for different parameter values.

1 Introduction

Simulation of a social network has always been a field of great interest and research. Information cascading, disease spreading, social segregation, etc. have been deeply investigated by scholars around the world. Not only are these simulations interesting but are also very useful in real life.

We have tried to simulate the evolution of communities in a social network. Like every other simulation, it is very challenging to work with but interesting at the same time. It involves the generation of synthetic graphs and community detection algorithms. The hardest part of the project is to get the mathematical modelling right otherwise the model falls apart on real networks.

A synthetic network was built consisting of communities, nodes and edges. Each vertex was given a feature vector. The feature vectors were allotted according to various distributions schemes (normal distribution, uniform distribution, power distribution). After this initialization the model was put to simulation. In each iteration new edges were added and old edges were removed between nodes depending upon the the similarity of the nodes (calculated from their feature vectors). These new connections affect the feature vectors of the nodes similar to the way humans change (their habits, hobbies, perceptions) depending on the new friends they make. We also studied how the community structure of the entire social network evolved with time.

Communities form an integral part of the society, for each community has some particular features which makes it stand apart. However, a person may be part of more than one society. But, for the sake of simplicity we have ignored this fact although we do not expect output to deviate much from the ones that we present in this report. Detecting clusters in a social networks is a very widely studied topic in the area of social networks and many algorithms have been put forward for this purpose. Girvan-Newman [LV14] algorithm is a beautiful attempt towards solving this task, however the algorithm itself suffers from the following drawback to serve our purpose for this project. On each iteration, it makes a cut in the network so as to have one more component than the original graph. We present forth an approach to make slight modifications to the algorithm to produce outputs accordingly.

1.1 Problem

How do communities evolve?

- Will all the communities eventually merge?
- Does the graph reach a stable state?
- How do the features of a node evolve?

1.2 Literature

The paper on Girvan-Newman algorithm [LV14] inspired us towards community detection in social networks. We then tried to see how to generate and simulate synthetic social networks which led us to the research papers [FA20; RY15]. We also came across different community detection algorithms, specifically the ones discussed in [PP11; NT07].

1.3 New idea

The idea is to run the model on a synthetically generated network which is in close resemblance to the real world network. Clauset Network and LFR Networks are reasonable approximations of real world social networks. [FA20] Since many real world distributions follow power law, so we distributed some features according to this, others with uniform distribution, while others were distributed normally. These features could mean anything, in general, any belief about certain topic and the extent of belief is measured by its magnitude. As a part of experiment, we chose feature vector to be of size 5, arbitrarily but we believe the results could be generalized to any number. It is a common observation to see that new connections in social networks are influenced largely by the commonness amongst them, we present forward an elegant approach, which shows results in close proximity of those which have already been arrived at in social computing like the close world phenomenon and the power law. With appropriate selection of parameters (obtained by hit and trial) we have been able to account closely for the phenomenon like nationalism, partitioning and unification which have occurred in the past. Understanding parameters in more detail may help us in predicting the future societal structure.

2 Method

2.1 Implementation details

This project is made with the following tools

- **Languages:** Python
- **Modules:** `networkx`, `matplotlib`, `python-louvain`, `numpy`

The project is divided into 3 parts

1. Graph generation (Clauset Network ,LFR Model, a custom algorithm explained later in the report)
2. Simulation (Techniques were designed by motivation from Reinforcement Algorithms, they are described later in the report)
3. Community Detection (Girvan-Newman Algorithm and Louvain Algorithm)

2.1.1 Graph Generation

Each node of the graph contains 2 additional properties - *feature* and *vulnerability*

- **Feature**: It is a vector of positive real numbers. It denotes the interests of a node. Each component of the feature vector denotes the extent of belief of node in that interest.
- **Vulnerability**: It is a positive real number between 0 and 1 denoting the vulnerability of the node to change. If vulnerability is 0, it means that the node is not influenced by its neighbors. A vulnerability of 1 means that the node is completely influenced by its neighbors i.e. the node does not have a choice of its own, it always conforms to its neighbors.

Each edge of the graph has 1 additional property - *age*

- **Age**: It is a positive integer which denotes the age of the edge since the beginning of simulation. Age of a newly added edge is initialized to 1 for implementation simplicity, although it has no effect as such on the results as only relative age of edges matters.

There are many ways to generate a synthetic network. As per the article [FA20], Clauset network model and LFR model are best suited. These graphs inherently contain communities in the structure and are hence a good approximation to real world networks and appropriate for the purpose of our simulation. Further, a custom graph is generated, algorithm for which is as follows:

1. Create an empty graph with n number of nodes.
2. Distribute feature vector among nodes according to given probability distribution.
3. Iterate over all pairs of nodes and add an edge if their commonness exceeds the given threshold.

2.1.2 Simulation

The idea for simulation of influence has been taken from general Reinforcement Learning algorithms that are used to simulate learning from experience. We modified the algorithm as follows to suit our need and simulate the spread of influence as if it were a learning from experience step in RL. To run simulation on our graph, we require two parameters - *threshold* and *edge_ strength*

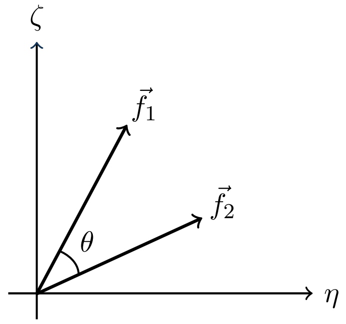
- `threshold`: It is a real number between 0 and 1 which helps in deciding the creation and deletion of edges. More details are given below.
- `edge_strength`: It denotes whether the strength of an edge decays, grows or doesn't change with time.

Steps of simulation

1. For every node, calculate the effect of its neighbors on its feature vector. Take into account the age of the edge and the vulnerability of node towards change.
2. For every node, see if friends of its friends share common interests with the node. If the commonness is greater than the *threshold*, add an edge between the nodes. The process of one step look ahead is motivated from the phenomenon of triadic closure.
3. For every pair of friends, see if they still share similar interests. If their commonness is less than the *threshold*, remove the edge between them. This is an important part of simulation and is motivated from [RY15] paper titled Simulation Method For Social Networks. Although the mentioned paper also suggests adding and deleting nodes, we have avoided this with an assumption of running simulation just over a lifetime, starting with some initial conditions as the former requires a huge amount of computational power.

Finding commonness

How do we find that two nodes share common interests? It is decided by the angle between their feature vectors. If $\vec{f_1}$ and $\vec{f_2}$ are the feature vectors of node a and node b, then the commonness between the nodes a and b is given by the cosine of the angle between $\vec{f_1}$ and $\vec{f_2}$.



$$\text{Commonness} = \frac{\vec{f_1} \cdot \vec{f_2}}{|\vec{f_1}| \cdot |\vec{f_2}|} = \cos\theta \tag{1}$$

Effect of friend

Let node b be the friend of node a and node a has a vulnerability of α. If $\vec{f_1}$ and $\vec{f_2}$ are the feature vectors of node a and node b, then the feature vector of node a is updated as follows

$$\vec{f_1} \longleftarrow (1-\alpha) \cdot \vec{f_1} + \alpha \cdot \vec{f_2} \tag{2}$$

3 Results

Different observations were made for simulations over different initial conditions and on basis of distributions used. Thus, the choice of initial conditions affects the simulation a lot and influences results which was both expected and desired. As mentioned above in the report, due to limited computational power available, we started from initial conditions and ran simulation assuming it were to last over a lifetime because of which we restricted our program not to add new nodes so selection of initial conditions should influence the results. Results of interest were obtained under certain choice of parameters and initial network configuration. Some simulations resulted in observations which were close to how the societal structure used to be in the past, when society comprised of tribal groups and communities lived in complete isolation from each other. Another selection of initial conditions and parameters suggested how communities formed out of pre-existing communities.

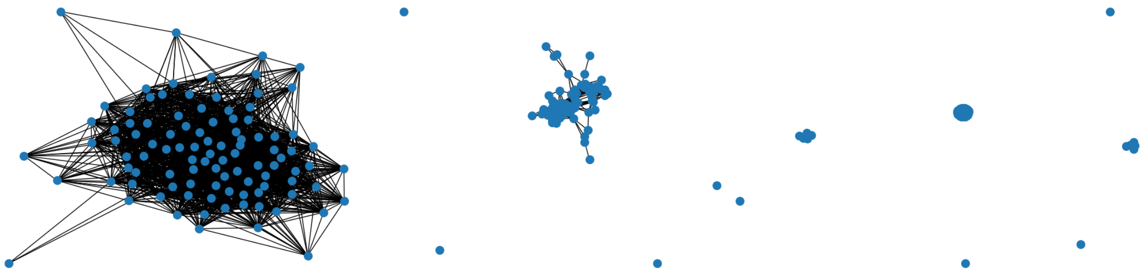
3.1 Experiment findings

Experiment 1

When `features` and `vulnerability` are distributed uniformly in a network and the edge strength decays with age. With `threshold` ranging from 0.8 to 0.95, the following results are obtained.

- The average commonness in a community gradually increases to 1.0.
- The number of communities first decrease (during first few time-steps), then increase (during next few time-steps) and finally become constant.

- There are no inter-community edges.



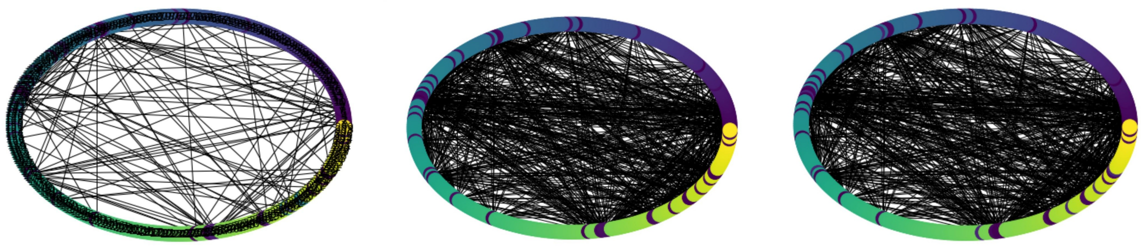
(a) Initial (b) Mid-iteration (c) Final

Figure 1: Evolution of network in experiment 1

The network finally obtained is thus a collection of strongly knit communities which are completely isolated from each other. The feature vectors of nodes belonging to a certain group evolve and converge to a common value i.e commonness of 1. However, feature vectors for all the nodes in a certain community resemble each other very closely. The plots for the experiment at different time steps of simulation are shown in fig 1.

Experiment 2

The initial network was created using the Clauset algorithm with 300 nodes. Feature vectors were allotted to the graph using uniform distribution.



(a) Initial, 200 edges (b) 1000 iterations, 745 edges (c) 2000 iterations, 752 edges

Figure 2: Evolution of network in experiment 2

Then the simulation was run for 2000 iterations in which nodes having features close to each other above a certain threshold ranging from 0.8 to 0.95 developed connections and nodes, commonness of whose feature vectors was below threshold got disconnected.

The number of connections increased and reached a saturation level.

3.2 Interpretation of findings

Experiment 1

The key results from the simulation are

1. The graph reaches a stable state where communities are fully connected.
2. There are no cross-community edges in the stable graph obtained after simulation.

3. The feature vector for nodes of community are very close to each other and converge.

The initial conditions chosen were a very good approximation of how life would have started on earth for humans with all the feature vectors being random. The initial links in the graph were made based on the commonness. The results on this simulation show society to be divided sharply between different communities with no edges in between and this is exactly what the social structure of tribal communities used to be. Hence, the results were very close to expectation for this experiment.

The results are surprising as a mathematical model have been able to predict the outcomes and changes in social structure that are in resonance with the ones we are aware of.

These results are quite promising. They are a result of mathematical modelling based on well-thought heuristics. The results not only provide a valid output for the given initial graph but also validates the initial graph itself, assuming the algorithms used are good approximation to how communities evolve with time. We may have been able to see the change of structure to inculcate inter-community edges with time, if our parameters were allowed to vary in some reasonable fashion and new nodes created while older ones deleted, however it was not the aim of experiment as we simulated only over one lifetime.

Experiment 2

The key results from the simulation are

1. The number of communities increased rapidly with time.

2. They attained a saturation level with no significant net increase in the connections.

The initial network was created using the Clauset algorithm with 300 nodes. It was observed that the number of edges in the communities increased up to a saturation level after which the number of new edges did not increase significantly with time.The number of connections develop very rapidly with time and reach a saturation level.

The results match up very closely with the real world phenomenon of community development and nation forming. If one observes the record of history one can observe how large number of separate communities and tribes developed into small political units. Then these units which had close cultural similarity developed stronger bonds and aligned themselves into bigger and close-knit social and political units. With time these clans grew into bigger empires. These constantly evolved, breaking up and realigning with other groups to form the first structure of nation states. In the post world-war period the situation has come to a more or less stable state with number of nations reaching a stable state. We still had breaking up of nations like the erstwhile USSR and unification of certain nations like East and West Germany, but by and large the nations have formed communities based on strong commonality of features amongst themselves.

4 Conclusion

As a part of this project, we learnt the following things

- Synthetic networks like Clauset model and LFR benchmark graph
- Different types of community detection techniques
- How Reinforcement Learning algorithms simulate the process of learning

We practically understood that mathematical modeling is an iterative process of improving the model inculcating finer and finer details, arguing for their functionality.

4.1 Team Work

We first searched related papers and articles on the internet individually. Then, we shared our ideas of what we found on the internet, the challenges that we might face. Everyone presented a summary of papers, articles or blogs they had gone through. We stayed through hours-long conferences, reasoning about the mathematical modelling - taking into account the important details of real social networks. We argued a lot over the selection of parameters for modelling, why are those factors important to have, how will they affect the simulation, what would their values be bound in between and how would values of different parameters affect the results.

In terms of implementation, the work was divided equally. Everyone was working on different files but monitored every pull request and suggested changes. In case of report, we took responsibility of different sections of the report. After a rough draft, we communicated the changes to each other and modified the report accordingly.

Although, we were able to achieve the goal of the experiment, we are of the opinion that, had we got access to higher computational power, we could have been able to simulated for "more than a lifetime" by introducing new nodes and deleting the existing ones. However, there exist a lot of challenges to it including but not limited to how to decide the feature vector for next generation from the previous one and how to simulate population growth.

References

[FA20] Arman Ferdowsi and Abdolreza Abhari. “Generating High-quality Synthetic Graphs for Community Detection in Social Networks”. In: *SpringSim'20*. 2020. `https://doi.org/10.22360/SpringSim.2020.CNS.001`.

[LV14] Tanja Vojkovic Ljiljana Despalatovic and Damir Vukicevic. “Community structure in networks: Girvan-Newman algorithm improvement”. In: *37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 2014. `https://doi.org/10.1109/MIPRO.2014.6859714`.

[NT07] Isabelle Stanton Nina Mishra Robert Schreiber and Robert E. Tarjan. “Clustering Social Networks”. In: *Lecture Notes in Computer Science*. 2007. `https://doi.org/10.1007/978-3-540-77004-6_5`.

[PP11] Giacomo Fiumara Pasquale De Meo Emilio Ferrara and Alessandro Provetti. “Generalized Louvain Method for Community Detection in Large Networks”. In: (2011). `https://doi.org/10.1109/ISDA.2011.6121636`.

[RY15] Quan Z. Sheng Rui Zeng and Lina Yao. “A Simulation Method for Social Networks”. In: (2015). `https://doi.org/10.1007/s13278-015-0246-4`. `https://crpit.scem.westernsydney.edu.au/confpapers/CRPITV144Zeng.pdf`.