Synthetic Social Network Generator

Authors
Raunak Kumar
Satish Kumar Oraon

Computer Science and Engineering

May 4, 2023

1 Abstract

Due to privacy issues, obtaining access to user attributes for network generation might be challenging. Users are growing increasingly hesitant to provide their personal information with third-party services, particularly social networking platforms, as data privacy concerns gain more attention. Social media firms struggle with how to get and use user data while still protecting users' privacy. When users' personal characteristics, like age, gender, and location, are essential for creating useful networks, this problem is particularly acute. To create networks that are efficient and considerate of user privacy, businesses must strike a balance between the necessity for user data and the right of users to privacy.

So, we have proposed a network generator that can not only create random attributes per node, but also it can take attributes from real-world datasets and try to mimic the real-world network. We try to evaluate our generator by comparing the real-world network with the network generated corresponding to the attributes of that node

2 Introduction

Online social networking platforms like Facebook, Twitter, Reddit, etc have become very popular . These platforms allow users to communicate with each other quite conveniently in addition to the ease it provides its user to share photos, and videos.

These online social networks have in them an immense amount of data associated with them. This data can be used by scientific researchers, organizations, psychologists, and commercial houses to study and understand human behavior. Researchers can analyze social interactions, identify trends and opinions, predict behavior, study the diffusion of information, and investigate the impact of social networks on behavior.

The problem associated with studying those data is that it is not easily available due to privacy concerns. Here is when the need of synthetic social network generators come in the picture. Generators that can mimic the real world network can be of great use .

Before moving further , we would like to first give a brief introduction on the typical features of a real-world social network . The first one is the **Scale Free property**. The real world social networks have typical scale free property associated with it . By scale free we mean that there are few nodes in the network which are very influential i.e they have millions of connections . These nodes are also known as the **Hubs** .We can say that the node degree distribution follows a **power law distribution**. In a network with a power-law degree distribution, a few nodes have many connections **(high degree)** while many nodes have relatively few connections **(low degree)**.Power-law distributions are also known as **"heavy-tailed" distributions**, since they have a long tail that extends far beyond the majority of the data.The second important property of real world social networks is the **Small World** property which means real-world social networks tend to have a high level of clustering, meaning that nodes tend to form tightly knit groups or communities. At the same time, the average path length between any two nodes in the network is relatively short. This property is known as the small-world property.The next important property is the **Community structure** in the real-world social networks meaning that nodes tend to be organized into clusters or groups that are more densely connected within themselves than with the rest of the network.

Homophily is also one of the important properties of the real-world social networks where nodes tend to connect to those nodes which are similar to them in some respect. We have tried to target this property to generate links between nodes which will be explained in detail in the coming sections.

In the field of network science, the synthetic network generators are denoted by a graph which is a set of nodes and set of links that exist between the nodes , The network can be of two types,**directed** and **un-directed**. Nodes can also have attributes associated with them. Attributes can be any thing . If we are trying to generate an intra-college Facebook network, then the attributes can be **major, year , gender, hometown and etc**

The validity or usability of any synthetic social network generator is determined on number of real-world features which it has in common with the real world network. We try to see if the network generated is follwing the power law or not , is it having the small world property or not , is it displaying the community structure or not . Hence a good generator is the one which generates large scale synthetic network which mimics the properties of real-world social network.

3 Related Work

Barabasi and Albert Model

In order to produce graphs that match the **power law distribution** for the node degrees, Barabasi and Albert created a model in 1999 that is popularly referred to as the BA model. Two techniques are used to create the network structure:

•**1 Slowly expanding the network** by adding nodes over time; • Attaching nodes to existing nodes with preference in accordance with the "rich get richer" theory.

2 Starting with only a few nodes, let's say m0, the network is being built. A new node with m edges, where m $\leq$ m0, and connected to m nodes is added in each timestep. The network is generated at random with **t+m0** vertices and **m * t** edges after **t** timesteps. The produced graphs are self-organized into a scale-free network that replicates the observed distribution.

However, the network generated shows low clustering coefficient as the model does not take into account other properties such as attribute similarity, and friends of friend concept .

Watts Strogatz model

This model was proposed to generate the networks which follow the small world property of the real networks . Initially there are N nodes which are connected in a ring like lattice . Each node in the neighbor is connected to **K** neighbors. A node and the edge incident to it are later randomly selected from the network with probability **p**, and the selected edge is then reconnected to a different node.

RMAT model

In this model , the adjacency matrix is divided into 4 parts . Each quadrant is given some probability.By recursively splitting the graph into four quadrants and probabilistically inserting edges between them, the RMAT model creates graphs. In more detail, the model chooses one of the quadrants and with a specific probability adds an edge between two randomly selected nodes in that quadrant at each step of the recursive process. Until the desired number of edges or nodes is reached, this process is repeated.The more detail can be found in the paper

Synthetic generators for simulating

social networks by Awrad Mohammed Ali University of Central Florida

In this model , the author is generating a baseline graph based on preferential attachment in the first phase . In the second phase , the generator is generating node attributes randomly .The links are created between nodes which are having **high feature similarity** .The statistics of the node and its attribute is calculated and is checked with a target statistics(of the network which which we want to mimic) . The generated statistics is passed through some fitness function and the attributes are tuned using **particle swarm optimization** or **genetic algorithm** . Once the termination threshold is reached , the final network is generated.

4 Proposed Method

Some of the above generated generators(except by the Awrad Mohammed Ali University of Central Florida) do not consider the node level attributes while generating the synthetic networks. But in most networks, we cannot ignore the role of node level attributes in formation of links. Most of the previous models are trying to just mimic the scale free network property , the small world phenomena etc .

Whenever a synthetic network is generated , the nodes represent the users and the attributes represent the characteristic of a node . Suppose , if a Facebook network of Indian users is formed , the attributes can be the political beliefs , age , gender, state in which the user is residing and many more such attributes.

So we propose to model a network generator which takes into account the **Jaccard similarity** between nodes alongside the preferential attachment . **Jaccard Similarity** is used to calculate the degree of resemblence between two sets . It is the ratio of intersection of two sets by the union of the two sets . So when the ratio is 1 , it means there is high resemblence between the two sets . When the ratio id 0 , it means the two sets are completely different.The motivation behind the use of Jaccard simlarity is that whenever a new user joins the network , he/she tends to find all those nodes which are influential(preferential attachment) and are also having some similarity in attributes between them . Let us understand this through a simple example . If suppose a new node joins the network and it is a huge football fan . So what the new node will do , it will check all the celebrities and form nodes with nodes such as **Cristiano Ronaldo , Lionel Messi**. Why will the node chose to form a link with **Virat Kohli**. So this is the basic idea which we have tried to incorporate in our network generator alongside also considering the preferential attachment model.

Our network generator also considers the fact that whenever a new node comes , it has high entropy , so there is a chance that it may form link with the nodes which are not having any attributes in common . So we have kept the threshold jaccard similarity low initially considering the observed tendency .

The network is generated in two phases. In the first phase we start with 2 nodes which are connected . We start adding node one by one in the network. The new node being added scans the whole network and calculates the jaccard similarity score with each remaining node in the network . It keeps all the nodes which are satisfying the threshold jaccard similarity score in a list called **Candidate nodes**. So these candidate nodes are those which show some resemblence to the new node . Now among these candidate nodes , we calculate the probability of connecting to these candidate nodes on the number of links they have . So a candidate node which has more links will be chosen with high probability . So , the first phase ends when the desired number of nodes are added in the network.

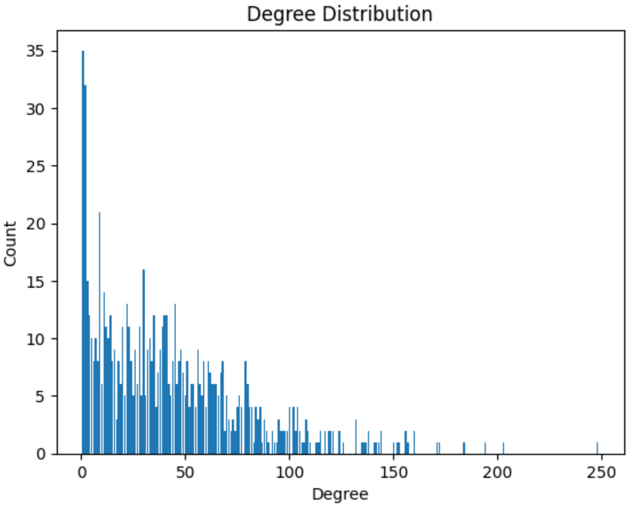
In the second phase of the network , we now consider forming links with friends of friend who are showing high similarity . The motivation behind the second phase is that, once the nodes become old in the network , their entropy reduces . Now there is tendency to form links with all the nodes which are very simialr to each other . Let us understand this with the persepective of the node which formed link with **Cristiano Ronaldo** when it was new . Suppose the node also went to XYZ University and was studying Engineering. So , as time passes it will have the tendecy to make links with nodes which are also having these attributes .Theres is a strong tendency to make links with friends of friend . The other models are just making connection with friend of friend without considering the similarity . We have tried to incorporate jaccard similarity when making connection with friend of friend . But unlike the first phase , we have increased the Jaccard threshold.

5 Discussion and Experimental Results

We have incorporated both the preferential attachment and the concept of Jaccard similarity while generating links. We also have considered making links with friends of friends if they pass the threshold. So expect the genrated network to follow both the scale free property and the high clustering coefficient property of the real world network

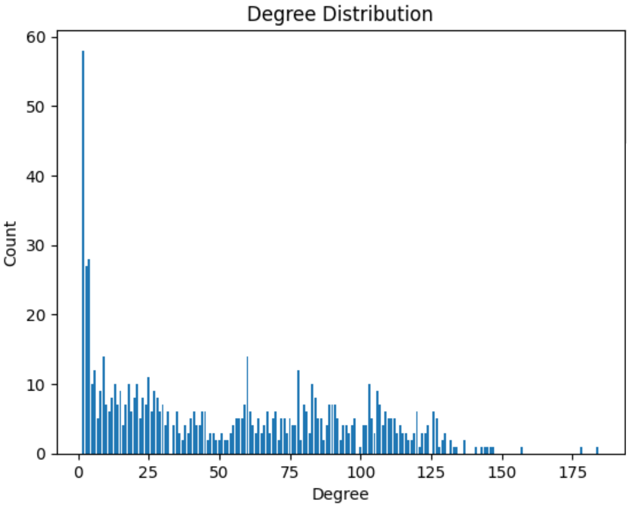
We have tested our generator on an intra university facebook network. The nodes are students . The attributes of nodes include **Year , major , gender , high school , dorm etc** .The real network is available which has 769 students and 16655 edges . So we have given the nodes the same attributes as given in the dataset and generated the network . We have also compared our model with earlier model for the same average degree and the number of nodes

The result images are attached below

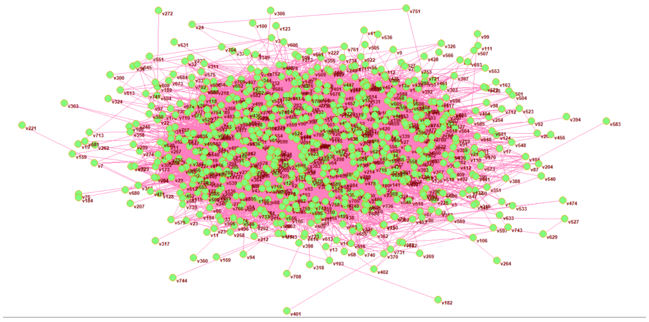


The above image is the image of the degree distribution of the original facebook network . The highest degree in the original network is around 250.

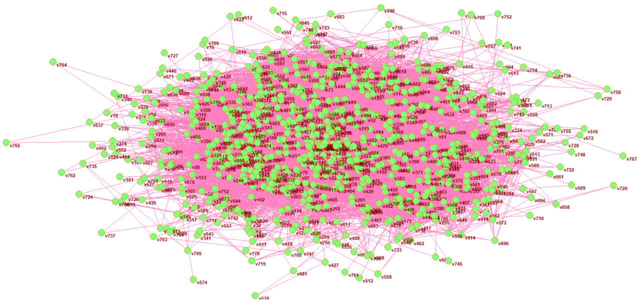
The image attached below is the degree distribution of the network generated using the same number of nodes and the same attributes used in the real network



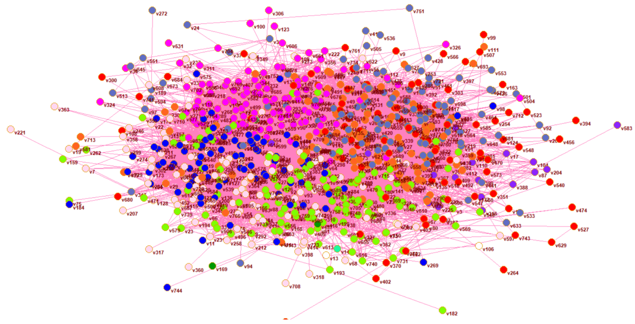
The original network as seen in pajek using Kamada Kwai layout is given below



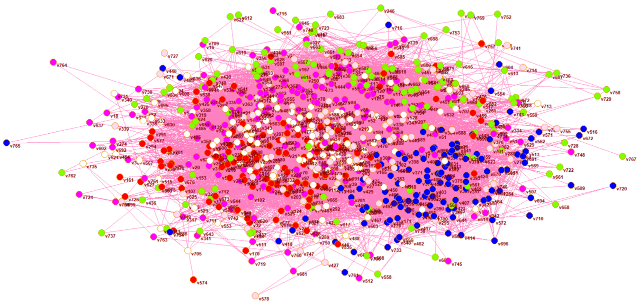
and the generated one is given below



We find that the number of edges in the generated graph is comparatively more than the number of edges in the original network . There are 20639 edges in the generated graph as compared to 16655 edges in the original graph. The **clustering coefficient** in the original network is 0.40929439048517247 and in the generated network is 0.4349093561836689 . Hence we observe that the clustering coefficient is showing promising results. There were 11 communities in the original network as shown below



However , the number of communities decreased in the generated network to 6 as shown in the picture below



6 Conclusion

We observed that our network generator is partially following the power law degree distribution and it is also not identical to the degree distribution of the original network. The clusteing coefficient is showing promosing results . This is because we are making links with friends of friend using jaccard similarity.

As far as the degree distribution not being similar is concerned, we are in conclusion that there are several other factors apart from Jaccard similarity which also influence the link formation like the concept of **heterophilicity** which is forming links with the completely different node when situation of some kind arise.

As far as degree distribution not fully following the power law is concerned, we are in conclusion that very few new nodes see Jaccard similarity as a reason to make links . Most of the nodes tend to make nodes to hubs without even considering the similarity of attributes. So we conclude that when a new node which is a **football fan** comes in picture, it not only makes link to **Cristiano Ronaldo** but also makes links to **Virat Kohli , Lebron James , Micahel Schumakar , Rohit Sharma , Sharukh Khan etc** because they are the hub and it is the tendency of new node to explore various field no matter if the node has common attributes or not.

For the future work,we think of incorporating **heterophilicity ,transitivity** while link formation so that our network is much similar to real world network

References

1 D. Chakrabarti, Y. Zhan and C. Faloutsos, "R-MAT: A recursive model for graph mining", Proc. SIAM Int. Conf. Data Mining, pp. 442-446, Apr. 2004.

2 A.L.Barab´asi and R. Albert. Emergence of scaling in random networks. Science, 286(5439):509–512, 1999.

3 D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. Nature,393(6684):440–442, 1998.

4 Synthetic Generators for simulating Social Networks by AWRAD MOHAMMED ALI B.S. University of Mosul, 2005