Reaction Paper on Analysing Graph Generative Models

Bhavya Agarwal

Department of Computer Science

Stony Brook University

Late Days Used - 1, Late Days Left - 2

1 Introduction

Mining graph patterns and generating synthetic graphs has been particular area of interest for many researchers over many decades. Predicting the structure of web and discriminating an abnormal network from normal are questions important to many fields.Hence, we try to understand the patterns which exist in a graph and generate them in our efforts for to replicate the real world networks. The graphs which we generate are important for many applications like-

Generating graph based on speculations - For example, if we come up with a model that the advent of semantic web would change the dynamics of web forever, a good model would help us to create a new synthetic graph depicting the future of the web and we can run our analysis on it.

Getting outliers - A graph can be created with the expected pattern and behavior and then can be compared with the real graph to detect the anomalies.

Compression - These patterns can be used in compressing graphs and making a small prototype in order to make our analysis faster.

With the rise of computers in recent times, many tools have been developed for measuring and analysing the underlying topology in a graph. The power of parallel computation has given us the ability to analyse millions if nodes at the same time. Also, the lines between many disciplines have blurred and hence many metrics have come into picture to understand the interactions between various components in the graphs. Due to these reasons, the survey on "Statistical Mechanics of complex networks"(R.Albert, A.Barabasi)[1] has underlined 3 important patterns which are described below-

1. SMALL WORLDS - it implies the minimum distance between 2 randomly selected nodes.
2. CLUSTERING - this pattern helps us in analysing the connectivity of the graph or a part of it .
3. DEGREE DISTRIBUTION - this is the probability distribution of a randomly selected node having a given degree.

These 3 patterns become our base in analysing the differences which occur between Graph Generative models and Real world graphs. They allow us to look deeper into their properties and mark out the areas for improvement in generative models.
Many other patterns like resilience[7] exist which help us out in marking the graphs resilient to attacks. Also, joint distribution is sometimes used instead of degree distribution. However, most researches have focused on the above 3 properties so we'll go ahead with analysing the graph generative models with respect to them.
The 2 graph generative models that will be considered here are -

Random Graph Generators These graphs are made by selecting 2 nodes and then adding an edge between them with a probability P. The Erdos and Renyi model(1960)[4] was the first such widely popular model. It was followed by many of its variants which tried to improve on some of its properties and make it better resemble a real-world graph. Here we also consider it's variant PLRG[3].

Preferential Attachment Graph Generators This model follows that in a network the rich get richer i.e. when a node enters a network it is most probable to be connected to the nodes which are already well-connected. Herbert Simon in 1955 showed that such a graph has a power law tail[2]. A similar model proposed in 1999 by Barabasi and Albert[5] was very critical and hence the model began to be referred as BA model. We also consider its modification Forest Fire model [6].

This paper relates to the topic of graph generators covered in the class and looks at the contribution and extension of a few popular methods. In the subsequent sections we analyse graphs generated by these models and compare them on the patterns defined above with the real graph.

2 Analysis of Different Models

Before exploring graph generative models, we first explore here the properties that real-world graphs exhibit in order to analyse the models better.

SMALL WORLDS- This topic has been of lot of interest in recent times due the widely available data with the increased penetration of Social Networks in the masses. It says that despite huge networks, the maximum distance between 2 nodes is relatively small. It was first discovered by Sociologist Stanley Milgram in 1967 in his manifestation of "Six degrees of separation"[1]. While he described the term only for United States, with the help of new social networks it has been proved to be true for nearly all the humans. Many related concepts like Erdos Number and Kevin Bacon number have been made to analyse the distance of a random member to a significant point in the graph.

CLUSTERING- With the tendency of users to form closed communities in Social Networks, they generally form cliques inside them where everyone knows everyone. This results on a complete graph which serves as the baseline to compute the clustering coefficient of a graph. The formula for clustering coefficient is -

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \tag{1}$$

i.e. the ratio of E_i number of edges that exist between k nodes(whose complete graph would have k(k-1)/2 edges). This term was introduced by Watts and Klogartz in 1998[2].

DEGREE DISTRIBUTION- Degree distribution is the spread in degrees of nodes in a graph. This spread is computed with the help of P(k) i.e. the probability that a randomly selected node would have k edges. In a large real graph like WWW and Internet [7] it has been found that it has a power law tail i.e.

$$P(k) \sim k^{-\gamma} \tag{2}$$

2.1 Paper 1- Erdos-Renyi model

First we consider the ErdosRenyi model[4] which was the first and simplest model for a graph generator. Here the edge is added beteen 2 nodes with a probability p. Hence, the degree distribution comes out to be

$$P(k) = \binom{N}{k} p^k * (1-p)^{N-k} \tag{3}$$

2.1.1 Critique

It can be seen here that this graph follows poisson distribution with peak at around $P < k >$ where $P < k >$ is the probability of a node having degree equal to average degree of the graph. The clustering coefficient in this graph is also equal to p. Hence, the size of the largest component in the graph depends largely on p. It is observed that a giant component with O(N) nodes begins to appear once we have $p > \frac{1}{N}$.

This graph formed the basis of development for graph models. It was shown to have properties like critical probability and phase transitions. While this model has a great historical importance, it is not able to represent any of the real-world graphs very well. The real graphs do not follow poisson distributions shown here. This graph has no sense of community as shown by many real-world graphs. Also, the clustering coefficient is dependent on N(p is actually $\frac{<k>}{N}$) while it is found to be independent of N in real graphs.

2.2 Paper 2- PLRG model

Power Law Random Graph (PLRG) model of Aiello et al. [2000] is a modification to the ER-Model in which the degree distribution is changed to follow power law instead of poisson distribution.

$$P(k) \sim k^{-\phi} \tag{4}$$

It was found that for $\phi < 1$, the graph is surely connected while for $\phi \sim 3.5$ and greater, the graph starts to breakdown into smaller pieces.

2.2.1 Critique

While this graph was shown to have improved on the degree distribution, it still lacked in other properties with respect to a real world graph. It had communities forming inside it and did not display the small world property. Its possible extension can be rewiring the graph to form communities. It does not support directed graphs and also tries to generate a graph with very few parameters which is not compatible with many real-world graphs.

Now we move to the next category of graphs i.e. the preferential attachment model.

2.3 Paper 3- Basic Preferential Attachment

As opposed to other models like random graph models, this model does not assume a fixed number of nodes before starting. It was given by Barabasi et al[5]. The networks grow after starting with a small set of edges. Whenever a new node is added to the graph, the earlier existing nodes with higher degree carry higher probability to be attached to the new node. The probability that an existing node would be an endpoint for the new node is given by

$$P(edge \quad to \quad existing \quad node \quad n) = \frac{degree(n)}{\sum degree(i)} \tag{5}$$

The network is considered undirected. Due to this property, the node with high number of edges end up with even higher number of edges hence showing the ”rich get richer” phenomenon.

2.3.1 Critique

We see here the properties exhibited by this model-

Degree Distribution The degree distribution resulting from the BA model is scale free, in particular, power law of the form

$$P(k) \sim k^{-3} \tag{6}$$

Hence, it has limitation of a fix power-law exponent of 3, as opposed to real-world graphs which display a wide array of exponents.

Average path length The diameter grows as O(logN/loglogN) hence displaying small world phenomenon since the diameter is lot less than the number of nodes in the graph.
It can be seen that diameter increases with increase in N, while many real graphs have shown to have shrinking diameters[6].

Clustering coefficient Clustering coefficient is empirically computed for BA model and was found to be significantly higher than random graphs. It was found to scale with network size by following equation-

$$C \sim N^{-0.75} \tag{7}$$

This result was obtained by Dorogovtsev, Goltsev and Mendes[2].

While the basic BA model does have these limitations, its simplicity and power make it an excellent base on which to build extended models. However it has many limitations like a fixed power law exponent and the constraint on it being directed. It was also found that the older nodes had higher degree, while no such property in present in WWW. The model explained next model shown tries to improve this model.

2.4 Paper 4- The Forest Fire Model

This preferential attachment model was developed by Leskovec et al. [2005][6] and was found to improve on many properties of BA model described earlier. This model has 2 properties, the forward burning probability p and backward burning probability r. The steps for graph generation can be listed as follows-

1. Initially node v picks a node w as its ambassador.

2. A random number x is chosen from a distribution and then node v forms in-links and out-links with x neighbors of w with the given probabilities.

3. The process is then repeated recursively for each of the new neighbors of v.

2.4.1 Critique

The degree distribution of this graph is found to be following power law for both in and out degree. Some nodes end up creating large conflagrations, which forms many out-links before the fire dies out. This also makes the graph display communities since the edges are copied to new node v as the fire spreads. This model was also shown to posses properties like shrinking diameters. However the authors have noted that a rigorous analysis of this model is difficult. These models also cannot cover the property of subgraphs deviating from power law(while the graph as a whole follows power-law) as shown by WWW[6].

3 Discussions

The proposed methods have formed the base of graph generative models and their analysis. While these basic models do have some limitations, the bulk of graph generators in use today can probably trace their lineage back to these models.

The random graphs offer amazing simplicity in their analysis. However they bear a little similarities to the real-world graphs and hence are sparsely used in modern analysis. While models like PLRG have removed the poisson distribution, they are still not able to match any other patterns. The clustering coefficient of most random graphs is found to be inversely proportional to N which implies that it goes nearly to 0 when N approaches infinity. While in the case of many real-world graphs, clustering coefficient in independent of N[Albert and Barabasi[2002]. Random graphs do not show any sense of community in them because every node has a equal probability of having an edge present. There have been some extensions based on probabilistic models which gives weights to endpoints of the node connections. Random graphs are also generally undirected, which needs to be improved in order to match graphs like hyperlinks in WWW. While techniques for Generalized random graph models have improved random graph generation, much further work is needed to accommodate all of the real-world graph patterns in the random graph generation process.

The Basic preferential attachment model also deviates from real-world graphs due to fix power-law exponent of 3. Also the model has not shown the property of shrinking diameters as shown by many real graphs. This property was improved upon by the forest fire model which has shrinking diameters as the number of nodes increases. Also, instead if many isolated components shown in many graphs, these graphs have on big component. The extensions like rewiring of edges also increase the similarity to real world graphs. Many new approaches could be tried here to solve the problem of fixed power-law exponent. Pages with higher PageRank can be given preference. Kumar et al gave a model in 2000[8] which tried to solve the problem of fixed exponent for both in-degrees and out-degrees. Still, most of these models have many limitations like lack of community like structures. The variation in power law is also a big issue. A possible model was given for extending forest fire model in order to resemble WWW's subgraphs deviating from power-laws, however it was marred by implementation issues [2]. Moreover, these models try to control the graph with a fix number of parameters which is not possible to generalise for large networks.Hence some more models are required here which improve on these shortcomings of these algorithms. The Table 1 given by [1] can serve as a benchmark for the models in order to match these real world networks.

References

[1] Albert, Rka, and Albert-Lszl Barabsi. "Statistical mechanics of complex networks." Reviews of modern physics 74.1 (2002): 47.

[2] Chakrabarti, Deepayan, and Christos Faloutsos. "Graph mining: Laws, generators, and algorithms." ACM Computing Surveys (CSUR) 38.1 (2006): 2.

[3] AIELLO, W., CHUNG, F., AND LU, L. 2000. A random graph model for massive graphs. In ACM Symposium on Theory of Computing. ACM Press, New York, NY, 171180.

[4] ERDOS,P. AND RENYI, A. 1960. On the evolution of random graphs. Publication of the Mathematical Institute of the Hungarian Acadamy of Science 5, 1761.

[5] Barabsi, Albert-Lszl, and Rka Albert. "Emergence of scaling in random networks." science 286.5439 (1999): 509-512.

[6] LESKOVEC, J., KLEINBERG, J., AND FALOUTSOS, C. 2005. Graphs over time: Densification laws, shrinking diame- ters and possible explanations. In Conference of the ACM Special Interest Group on Knowledge Discovery and Data Mining. ACM Press, New York, NY.

[7] Albert, Rka, Hawoong Jeong, and Albert-Lszl Barabsi. "Error and attack tolerance of complex networks." Nature 406.6794 (2000): 378-382.

[8] KUMAR, R., RAGHAVAN, P., RAJAGOPALAN, S., SIVAKUMAR, D., TOMKINS, A., AND UPFAL, E. 2000. Stochastic models for the Web graph. In IEEE Symposium on Foundations of Computer Science. IEEE Computer Society Press, Los Alamitos, CA