

An Investigation of the Distributional Characteristics of Generative Graph Models

Sebastian Moreno
Purdue University
Department of Computer Science
smorenoa@cs.purdue.edu

Jennifer Neville
Purdue University
Departments of Computer Science and Statistics
neville@cs.purdue.edu

1. Introduction

Graphs and networks are a natural data representation for analysis of a myriad of domains, ranging from systems analysis (e.g., the Internet) to bioinformatics (e.g., protein interactions) to psycho-social domains (e.g., online social networks). Researchers in these fields are developing models to analyze the properties of the network, algorithms to propagate information throughout the network, and methods to predict the characteristics of nodes and edges in the network. However, evaluation of this work is restricted by our ability to collect a sufficient number of network samples, with which to evaluate claims about the performance of various approaches.

To assess the performance difference between two algorithms A and B with a measure m , statistical tests generally compare the sampling distributions for m_A and m_B and assess whether the difference between the two distributions is significant. However, to estimate the sampling distributions for m_A and m_B , we need to measure performance on different samples of data. With networks, we no longer have a population of independent instances from which we can easily draw a set of samples. Instead, we generally have a *single*, large graph of interconnected objects. For example, there is only one World Wide Web, only one social network on Facebook, and only one protein-protein interaction network for a species. Since there are complex dependencies among the nodes of these graphs, as well as heterogeneous graph structure, it is difficult to devise unbiased network sampling methods to evaluate the general properties of network models, algorithms and protocols. One solution is to use statistically sound methods for generating *similar* networks that belong to the same *population* as the observed network.

Due to the recent surge of interest in small-world graphs and networks with power-law degree distributions, there has been a flurry of work on generative models of graphs that can reproduce skewed degree distributions, short average path length and/or local clustering (see e.g., [2, 11, 1, 6,

10, 7]). However, these methods typically focus on preserving either global graph properties (e.g., average path length) or local graph properties (e.g., transitive triangles), but not both. There has been little research focusing on the connections between local and global properties and the accuracy/efficiency tradeoffs between matching on one or the other.

Moreover, analysis of these proposed methods has generally centered on empirical validation that the properties of interest are adequately preserved in generated graphs. Specifically, the empirical analysis has consisted of visual comparison of properties of the input and a small number of generated graphs. There has been little work investigating the characteristics of the *distribution* of graphs that are generated with the formulated models. One exception is the recent discovery that learning ERGM models with only local features can lead to *degenerate* global models (i.e., the estimated distribution places most of its probability mass of either the empty or complete graphs) [3]. This indicates that there are long-range dependencies in social networks that cannot be accurately captured by local features alone. On the other hand, graph generation methods that use fractal patterns to preserve global characteristics [7] have not been analyzed to determine the extent to which they preserve local properties—since they are based on a fractal formulation, it is likely that the generated graphs do not exhibit sufficient variance at the local level.

In this work, we study the distributional properties of two competing generative models—comparing the Kronecker model [7], which was designed to capture global graph properties, to the ERGM model [10], which was designed to capture local graph properties. We consider networks drawn from two *real-world* populations—grade school social networks from the Adolescent Health Survey and undergraduate social networks in the Purdue Facebook network. First, we measure the empirical distribution of graph properties in these two sets of networks. Next, we learn models from a representative sample in each domain. Then we generate multiple networks from each of the learned models to inves-

tigate the distributional properties of models. We evaluate how well the generated datasets reflect the properties of the observed network—both in terms of accuracy on global and local properties, and to see whether the models produce sufficient variance in the generated graphs.

2. Models

For this work, our aim is to study the distributional properties of probabilistic models of graph structure. We consider two representative models that can be learned from an observed network data—one that aims to preserve local structure and another that aims to preserve global structure.

2.1. Kronecker Model

The first model is the Kronecker graph model, a fast and scalable algorithm for learning models of large-scale networks with power-law degree distributions [7]. The Kronecker multiplication model of [7] is a fractal method, which starts from a small adjacency matrix with specified probabilities for all pairwise edges, and uses repeated multiplication (of the matrix with itself) to grow the model to a larger size. To generate a sample graph from the model, the algorithm independently samples each edge according to its associated probability of existence in the model. This results in a graph with self-similar structure at different levels of granularity. It has been shown empirically, that this approach successfully preserves a wide range of global properties of interest, including degree distributions, eigenvalue distributions, and path-length distributions.

More specifically, the model generates self-similar graphs, in a recursive way using Kronecker multiplication. The algorithm starts with a initial matrix $X_1 = \Theta$ with N_1 rows and columns, where each cell value is a probability. Typically $N_1 = 2$ or 3 , e.g.:

$$X_1 = \begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{bmatrix}$$

To generate graphs of a larger size, the Kronecker product of X_1 is taken K times with itself to generate a matrix $\mathbf{P} = X_K = (X_1)^K = X_{K-1} \otimes X_1$, with $(N_1)^K$ rows and columns. Each cell value in the matrix \mathbf{P} represents to the probability that an edge exists between node i and j . A new graph with $(N_1)^K$ nodes can be generated from P_{ij} by randomly determining the existence of each edge e_{ij} according to the associated probability P_{ij} .

To estimate a Kronecker model from an observed graph G , the learning algorithm uses maximum likelihood estimation to determine the values of Θ that have the highest likelihood of generating G :

$$l(\Theta) = \log P(G|\Theta) = \log \sum_{\sigma} P(G|\Theta, \sigma) P(\sigma)$$

where σ defines a permutation of rows and columns of the graph G . The model assumes that each edge is a binomial random variable, given $\mathbf{P} = (X_1)^K$. Therefore the likelihood of the observed graph $P(G|\Theta, \sigma)$ is calculated as:

$$P(G|\Theta, \sigma) = \prod_{(u,v) \in E} \mathbf{P}[\sigma_u, \sigma_v] \prod_{(u,v) \notin E} (1 - \mathbf{P}[\sigma_u, \sigma_v])$$

where σ_u refers to the permuted position of node u in σ .

With this formula, the maximization algorithm searches for the MLE parameters $\hat{\Theta}$. See Leskovec and Faloutsos [7] for more details on the learning algorithm (including how to efficiently scale the estimation procedure for large graphs).

2.2. ERGM Model

The second model is the exponential random graph model (ERGM) [10, 8] from the mathematical sociology community (also known as the p^* model). ERGMs represent probability distributions over graphs with an exponential linear model that uses feature counts of local graph properties considered relevant by social scientists (e.g., edges, triangles, paths).

The approach is focused on modeling the local structure of networks such as triangles or k -stars. For a graph G , we will denote its features with a vector $\mathbf{f}(G)$. ERGM models define a probability distribution of $P(G)$ over the set of possible graphs \mathcal{G} :

$$P(G|\theta) = \frac{1}{Z(\theta)} \exp(\theta^T \mathbf{f}(G))$$

where $Z(\theta) = \sum_{G' \in \mathcal{G}} \exp(\theta^T \mathbf{f}(G'))$ is the standard partition function.

As with all exponential models, P can be estimated using the maximum entropy principle: Choose the parameter $\hat{\theta}$ which maximizes $l(\theta)$ the log-likelihood of the input graph G^* . It is easy to show that this corresponds to the model that matches the expected values of features for graphs in \mathcal{G} to the features of the input graph G^* (i.e., $\mathbf{f}(G^*) = E_{P(\mathcal{G})} \mathbf{f}(\mathcal{G})$). The feature constraints ensure that the local properties of the graph are preserved and the partition function Z ensures global consistency. However, it has recently been discovered that estimation with ERGMs using only local features can result in *near-degenerate* models that place all their probability mass on either the complete or empty graph, and in which the original graph is very unlikely [3]. Sociologists have alleviated extreme degeneracy problems by manually specifying a larger number of features to use in ERGM models, such as alternating k -stars and alternating k -paths [9]. Although, this work has begun to explore a more expansive set of features, it is driven primarily by sociological motivations—the researchers are more focused on identifying subgraph patterns that are semantically meaningful (e.g., alternating k -paths measure

the connections among a pair of nodes that are not directly linked) rather than identifying local patterns that will improve preservation of the global graph structure.

3. Experimental Analysis

The goal of our experimental evaluation was to explore the distributional properties of the ERGM and Kronecker models. More specifically, our aim was to test the following three hypotheses:

1. The ERGM models preserve local clustering more accurately than the Kronecker models.
2. The Kronecker models preserve global path lengths more accurately than the ERGM models.
3. The ERGM models generate graphs with more variance than the Kronecker models, particularly with respect to local patterns of clustering.

3.1 Data

We evaluated the models with two different real-world social network datasets. The first set is drawn from the public Purdue Facebook network. Facebook is a popular online social network site with over 150 million members worldwide. We considered the set of 56061 Facebook users belonging to Purdue University network in March 2008. The public Purdue network comprised more than 3 million public friendship links among the members. Users had an average and median degree of 46 and 81 respectively. Within the larger Purdue network there are subnetworks for faculty, staff, and students. To estimate the variance of real-world networks, we considered the six networks corresponding to each of the undergraduate classes from 2006-2011.

The second dataset is comprised of a set of social networks from the National Longitudinal Study of Adolescent Health (AddHealth) [5]. The National Longitudinal Study of Adolescent Health consists of survey information from 144 middle and high schools, collected (initially) in 1994-1995. The survey questions queried for the students’ social networks along with myriad behavioral and academic attributes, to study how social environment and behavior in adolescence are linked to health and achievement outcomes in young adulthood. In this work, we considered the social networks from six schools with roughly equivalent structure—each network has 1100-1600 nodes with a density in the range [0.004-0.005].

3.2 Methodology

To compare the models, we consider three graph properties: (1) degree, (2) clustering coefficient, and (3) path

lengths. The degree of a node d_i is simply the number of nodes in the graph that are connected to node i . Degree is a local property of nodes, but since many networks have heavy-tailed degree distributions, the overall degree distribution is often considered a global property to match. Clustering coefficient is calculated for a node i as: $c_i = \frac{2|\delta_i|}{(d_i-1)d_i}$, where δ_i is the number of triangles in which the node i participates and d_i is the number of neighbors of node i . Clustering coefficient measures the local clustering in the graph. For path length, we consider the hop plot distribution in the graph, which refers to the number of nodes that can be reached with h “hops” in the graph: $N_h = \sum_v N_h(v)$, where $N_h(v)$ is the number of neighbors that are $\leq h$ edges away from node v in G . The hop plot measures the global connectivity of the graph.

For the six Facebook networks and the six AddHealth networks, we calculated the cumulative distributions for degree, clustering coefficient, and hop plot. From these distributions, we investigated the empirical variance of real-world networks that are conceivably *drawn* from the same *distribution*. Figures 1-2 plot the median (solid lines) and interquartile range (dashed lines) for the two sets of networks.

To learn the models, we selected a single network from each dataset to use as a training set. To control for variation in the samples, we selected the network that was closest to the median of the degree distributions in each dataset. In the Facebook data, we selected the class of 2008 network, with 2464 nodes and 12900 edges. In this case, the selected network was also very close to the median of the clustering coefficient and hop plot distributions. In the AddHealth data, we selected the network from school 117, with 1278 nodes and 7939 edges. This network is close to median of the degree and hop plot distribution but has the highest clustering of the six network samples.

Using the selected network as a training set, we learned a Kronecker and ERGM model. For the Kronecker model, we set $N_1 = 2$ and generated the initial values for $\hat{\Theta}$ from a uniform distribution. The learning algorithm searches for the MLE $\hat{\Theta}$ and ends when the estimated values stabilize. (We ran the learning several times and it produced very similar parameters for $\hat{\Theta}$ each time). For the ERGM model, we used the Statnet implementation in R [4]. For features, we used the geometrically weighted edgewise shared partner distribution (gwesp) and alternating k-star (altkstar) parameters (Facebook: $\alpha = 0, \lambda = 5$, AddHealth: $\alpha = 0, \lambda = 0$). To learn the models, we set the minimum number of MCMC samples to 1 million. Note that, since the aim of the work was to compare the relative performance of the two models, we did not spend a great deal of effort on fine-tuning the features of the ERGM model or the size of the Kronecker initiator matrix to fit the distributions of both data set more closely. However, these results, reflect the best fit over a

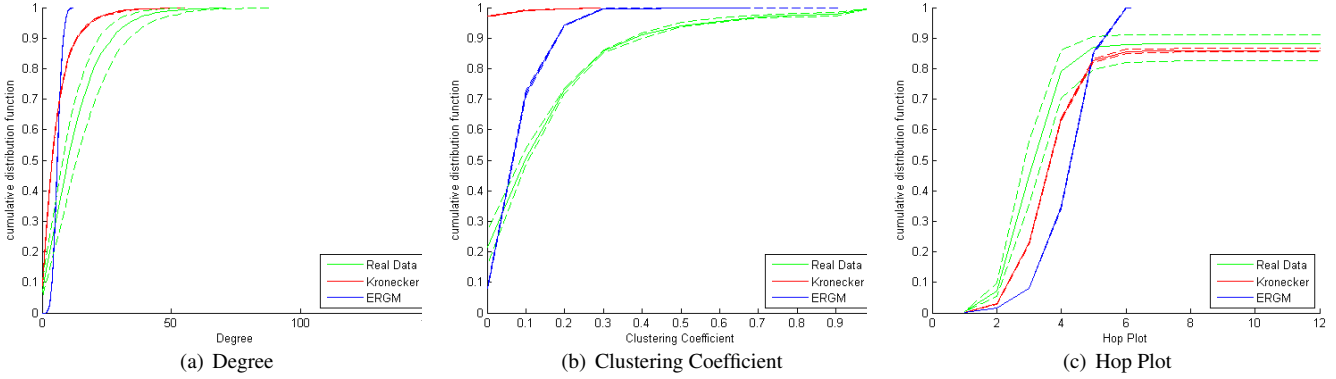


Figure 1. Variation of graph properties in Facebook networks—both real and generated networks.

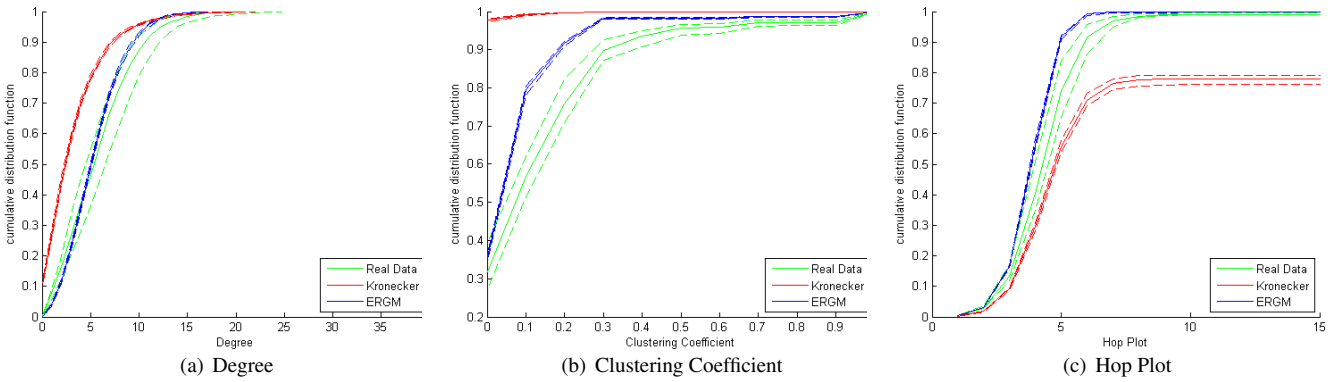


Figure 2. Variation of graph properties in AddHealth networks—both real and generated networks.

limited number of variations, including Kronecker graphs with $N_1 = 3$ and ERGM models with standard local features (e.g., density, number of triangles).

3.3 Results

For each learned model, we generated thirty graphs (five for each network size in the original samples). From the thirty samples we estimated the empirical sampling distributions for degree, clustering coefficient, and hop plots. In Figures 1 and 2 we plot the median and interquartile range for the graphs generated from the Kronecker and ERGM models. Solid lines correspond to the median of the distributions; dashed lines: the 25th and 75th percentiles. The ERGM graphs were each generated using ≥ 10 million MCMC samples, from a different (randomly-chosen) initial graph. The Kronecker graphs were generated using the estimated matrix $\mathbf{P} = X_K = (X_1)^K$. Since the size of \mathbf{P} is generally larger than the target network size ($(N_1)^K \geq N$), we simply drop the extra $(N_1)^K - N$ entries from \mathbf{P} before generating the edges of the graph.

Figure 1 shows the results for the Purdue Facebook data.

The Kronecker graphs were generated using the parameters estimated from the class of 2008 network:

$$\hat{\Theta} = \begin{bmatrix} 0.95 & 0.52 \\ 0.52 & 0.35 \end{bmatrix}$$

The ERGM graphs were generated using the following estimated parameters: $g_{wesp} = 2.183$ and $altkstar = -0.966$. The results show two things. First, the models match on global and local properties as expected. The Kronecker model captures the global degree distribution and hop plot distribution more accurately than the ERGM model, but the reverse holds for clustering coefficient, which the Kronecker model does not capture at all. Note that the ERGM models will match the degree distribution more closely if a separate degree parameter is included for each node in the network, however this approach is not scalable to networks of more than a few hundred nodes. Second, neither model reflects the amount of variance exhibited in the real networks. Moreover, it is surprising that the variance of the generated graphs is so slight that it is almost not apparent in the plots. Although our conjecture was that the Kronecker model would not generate graph with enough lo-

cal variance, we still expected that both models would produce graphs with sufficient variation in the global properties.

Figure 2 shows the results for the AddHealth data. The Kronecker graphs were generated using the parameters estimated from the network of school 117:

$$\Theta = \begin{bmatrix} 0.93 & 0.43 \\ 0.43 & 0.47 \end{bmatrix}$$

The ERGM graphs were generated using the following estimated parameters: $gwesp = 1.444$ and $atklstar = -2.933$. As with the Facebook data, the local properties of the AddHealth networks are not captured by the Kronecker model. The clustering coefficient distribution indicates that almost all nodes have a clustering coefficient close to zero. Although the variance of the Kronecker graphs is again very low, it is however slightly higher than the on the Facebook data. We conjecture that this is due to the smaller number of Kronecker multiplications needed to generate the graphs—since the AddHealth networks are smaller than the Facebook networks. For the AddHealth data, the ERGM graphs appear to preserve the local and global properties better than the Kronecker graph. However, this is not entirely unexpected, since the degree distributions in the AddHealth data are not as skewed as the Facebook networks. Again, the most notable characteristic of the generated ERGM graphs is this lack of variance compared to the observed samples.

There are two potential reasons for the lack of variance in the graphs generated from the ERGM models. Either the estimated probability distribution is sharply peaked over a small number of graphs that most closely matches the input network, or it is difficult for the Markov chain Monte Carlo sampling method to fully explore the space of possible graphs. Our experiments used 10 million steps in the MCMC sampling process for each network, so it seems that the former is a more likely explanation. However, we are currently investigating this in more detail. For the Kronecker model, it is likely that the use of independent edge probabilities and fractal expansion contribute to the small variation in generated graphs. The reported experiments use 2X2 initiator matrices. Although we observe the same effect with 3X3 matrices, it is possible that larger initiator matrices may help to alleviate the problem.

4 Conclusion

Overall, our empirical investigation helps to illustrate the distributional properties of current generative graph models learned from real-world data. The Kronecker models appear to capture the global network properties more accurately than local properties such as clustering. On the other hand, the ERGM model appears to capture local patterns more accurately than global patterns such as path lengths.

Most notably however, is the lack of variance in the generated graphs from both models—when compared to the variation in network samples drawn from the same distribution. Since both our datasets consist of social networks, it would be interesting to see if similar levels of variation are present in networks drawn from other domains (e.g., physical networks, biological networks).

Overall, these results are an initial indication that there is a need for alternative generative models that can (1) tradeoff between capturing local and global patterns efficiently, and (2) generate samples with sufficient variance to reflect real-world domains. More importantly it highlights the need to investigate the distributional characteristics of networks more carefully—both in real-world networks and in the network samples generated from our models.

References

- [1] A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.
- [2] O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, 81:395:832–842, 1986.
- [3] M. S. Handcock. Assessing degeneracy in statistical models of social networks. Working Paper 39, Center for Statistics and the Social Sciences, University of Washington, 2003.
- [4] M. S. Handcock, D. R. Hunter, C. T. Butts, S. M. Goodreau, and M. Morris. *statnet: Software tools for the Statistical Modeling of Network Data*. Seattle, WA, 2003. Version 2.0.
- [5] K. Harris. The National Longitudinal Study of Adolescent health (Add Health), Waves I & II, 1994 1996; Wave III, 20012002 [machine-readable data file and documentation]. *Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill.*, 2008.
- [6] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proceedings of the 42st Annual IEEE Symposium on the Foundations of Computer Science*, 2000.
- [7] J. Leskovec and C. Faloutsos. Scalable modeling of real graphs using Kronecker multiplication. In *Proceedings of the International Conference on Machine Learning*, 2007.
- [8] G. Robins, T. Snijders, P. Wang, M. Handcock, and P. Pattison. Recent developments in exponential random graph (p^*) models for social networks. *Social Networks*, 29:192–215, 2006.
- [9] T. Snijders, P. Pattison, G. Robins, and M. Handcock. New specifications for exponential random graph models. *Sociological Methodology*, 36:99–153, 2004.
- [10] S. Wasserman and P. E. Pattison. Logit models and logistic regression for social networks: I. An introduction to Markov graphs and p^* . *Psychometrika*, 61:401–425, 1996.
- [11] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–42, 1998.