Generating Power-Law Networks with Soft Degree Constraints

Erich McMillan,[1] Weibin Zhang,[1] and Kevin E. Bassler[1]

Department of Physics, 617 Science and Research 1, University of Houston, Houston, Texas 77204-5005, USA

(Dated: 10 January 2017)

We study the structural characteristics of power-law networks generated by soft-constrained ensemble methods using a given degree sequence. Comparisons of the random exponential graph model (ERGM), a soft-constrained method to hard-constrained methods, which by definition form correct structural associations, show that ERGM is not effective at accurately reproducing clustering coefficients and quantity of loops 4 in power-law networks whose exponents lie between 2 and 3. Our results reveal that the ERGM is biased toward more tightly clustered networks indicating that care should be used when utilizing ERGM and possibly other soft-constraint methods.

I. INTRODUCTION

Complex systems can be, and often are, modeled as networks, in which a set of connections (edges) link a set of points (nodes) [citation newman barabasi]. The number of edges which link to a node is known as the node's degree, k, in many cases a defining property of the network is the distribution of the degrees of its nodes. Of particular interest are networks which are power-law distributed in which the degree distribution has the following relationship, $\rho_k = Ck^{-\gamma}$ [citation newman]. Examples of power-law networks include the Internet, social systems, protein-protein interactions, however the study of such networks is often impeded by incomplete knowledge of the system stemming from its size and non-static nature, or elusiveness. A common case occurs when the degree of each node in the network is known, but the manner in which they are connected is unknown [citation pertaining to sexually transmitted disease]. If the distribution of the system is known to exhibit power-law behavior there are several methods of generating plausible networks which meet the prescribed degrees. Two major categories of so called network generation models are hard and soft-constrained, the former, by definition forms networks which exactly meet the prescribed degree sequence and the latter forms such ensembles using monte-carlo methods to generate an ensemble which meets the prescribed degree sequence on average over an ensemble of networks[newman and charo's hard const method]. Soft-constrained methods are preferred to hard-constrained methods due to their simplicity and speed, however it was previously unclear what bias they introduced to the structural characteristics of generated ensembles i.e. local clustering coefficients and number of loops lengths 3 and 4 due to their stochastic nature. We show that a major soft-constrained method, the random exponential graph model is biased toward more clustered networks particularly in the region where, $2 < \gamma < 3$ [newman].

As our soft-constrained method we used the well known exponential random graph model (ERGM). The ERGM allows for great flexibility in the application of constraints to the set of possible graphs, ζ, such that, $< x_i >= \sum_{G=\zeta} P(G)\ x_i(G)$ [newmans book]. In the case where a prescribed degree sequence k is given, the constraints on the ensemble become $< k_i >= \sum_{G=\zeta} P(G) k_i(G)$. Maximizing the entropy, $S = -\sum_{G=\zeta} P(G) ln P(G)$, of the set of possible graphs minimizes the bias in the resulting ensemble. The mathematics behind the derivation of the such a minimization are better explained in the following [][], nonetheless the resulting

derivation includes a set of Lagrange mulitpliers, β, for each of the constraints applied to the system. Applying a prescribed degree sequence k as a constraint, and restricting the model to simple, undirected networks with no self-links gives the following for the probability of an edge occurring between nodes i and j, $p_{ij} = \frac{1}{1+e^{-(\beta_i+\beta_j)}}$ [newmans]. Therefore the average degree a node i will be, $< k_i >= \sum_j \frac{1}{1+e^{-}(\beta_i+\beta_j)}$, which will result in a non-linear system of equations with a number of equations equal to the number of nodes. Fortunately, assumptions concerning sparse networks where $e^{-(\beta_i+\beta_j)} \gg 1$, meaning that $p_{ij} \simeq e^{\beta_i} e^{\beta_j}$, enable the β values to be estimated using the following, $\beta_i = log(\frac{k_i}{\sum_j k_j})$.

The well known Chung-Lu model for generating power-law distributed networks, which we had hoped to use in our research, takes advantage of this knowledge and further simplifies the process: $\rho_{ij} = \frac{k_i * k_j}{\sum_h k_h}$, avoiding the difficult process of solving the non-linear system of equations involving β. We have found that the Chung-Lu model's constraint, $max(k)^2 \leq \sum_i k_i$, prevents the model from being used in the region of interest, $2 \leq \gamma \leq 3$, because the percentage of sequences which conform to the constraint approaches zero as $\gamma \to 0$ (include figure of percentage of figures which are CL compliant). This occurs near a significant graphically transition for power-law networks $\gamma = 2$, beyond this point as $\gamma \to 0$ few large graphical sequences may be formed. Additionally near this limit, as we have shown in previous work, [all scale free networks are spares] power-law networks in this region cannot be considered sparse violating the principle assumption upon which the Chung-Lu model relies. In this case, the full ERGM model, $p_{ij} = \frac{1}{1+e^{-(\beta_i+\beta_j)}}$, must be used and the resulting non-linear system of equations must be solved. **in this limit gamma between 0 and 2 no graphical networks may be constructed in the limit of large N

insert weibin's proof for graphically transition of CL

Further proof that the assumptions of sparseness are incorrect are shown by comparing the β values solved for in ERGM to those predicted by $\rho_{ij} = \frac{k_i * k_j}{\sum_h k_h}$. The following figures show that in the region where $\gamma > 3$ the predicted values and ERGM values align closely. As $\gamma \to 2$ the configuration of the β values diverges radically from the predicted values. Although, it remains a distinct possibility that other solutions exist for individual sequences, it is impossible that any system of equations which diverged from the predicted values would also exhibit the predicted values as a solution because we utilized the predictions as initial guesses for the system of equations. Additionally, efforts to find other possible solutions

were unsuccessful, however due to the difficulty of visualizing and predicting solutions to such large non-linear systems there may still exist alternative solutions.

insert discussion on DD3M and hard constrained methods

Utilizing the ERGM the closeness of the convergence of the degree sequence over the ensemble toward the prescribed sequence is a critical component toward determining whether the structural components of the ensemble on average will be correct when utilized in research. Figure shows that the ensemble degree sequence quickly approaches the prescribed values quickly as the ensemble size increases, utilizing this data we chose to use 10000 networks in our ensembles. Of acute importance is what occurs to the average value of the structural characteristics of the ensemble as the number of networks increases; ideally these values should converge toward a value with increasing accuracy. Common key structural characteristics of networks include: local clustering coefficient (LCC), global clustering coefficient (GCC), and the number of squares (NOS). As seen by the data shown in figure

REFERENCES

W. T. V. William H. Press, Saul A. Teukolsky, *Numerical Recipes in C, Second Edition*, 2nd ed. (Cambridge University Press, 2002) pp. 347–383.