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: 24 February 2017)

We study the structural characteristics of power-law networks generated by soft-constrained ensemble methods using a given degree sequence. Comparisons between 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. We also show that the Chung-Lu Model for generating power-law networks cannot work where $\gamma < 3$ as $N \to \infty$.

I. INTRODUCTION

When studying complex systems we often model them networks, in which a set of connections (edges) link a set of points (nodes) [cite newman, barabasi]. The number of edges which link to a node is known as its degree. Often the empirical study of networks is impeded by incomplete knowledge about it, which stems from a myriad of factors including size, non-static nature, and privacy concerns. Ideally researchers would like to generate networks which conform to the known information about the system. A common case occurs when the degree, k, of each node in the network is known, i.e. the degree sequence, but the manner in which they are connected is unknown [citation pertaining to sexually transmitted disease]. In such instances, it is possible to utilize network sampling methods to create networks which conform to the expected degree sequence.

Methods including Markov Chain Monte Carlo (MCMC), Chung-Lu and Exponential Random Graph Model (ERGM) are widely used in graph sampling and fall into two main categories: soft-constrained (SC) and hard-constrained (HC) [cite charo, CL and ERGM]. Both categories require that some constraint, such as the degree sequence, utilizing the known information about the system be placed upon the subset of all possible graphs. HC will meet these constraints with each and every network generated, while SC methods will meet such constraints on average over an ensemble of many networks, perhaps thousands [cite newman and charo].

Despite the inaccuracy of individual networks, SC are preferred due to their relative simplicity and speed compared to HC which can be slow and difficult to implement. One advantage of HC is that by definition they introduce little to no bias towards networks which do not accurately reflect the original system, put simply it is possible to garner other information about the possible configuration of the system using only a single constraint. The effects of SC methods on their resulting ensemble are not well understood. They can suffer from degeneracy issues in which although the prescribed value is met on average the individual values are not distributed around the mean instead they separate into two or more micro-states none of which reflect the actual expected value [cite deg paper].

Our research focused on network sampling methods using power-law distributed systems, or those whose degree distribution conforms to, $\rho = Cx^{-\gamma}$. Systems which are known to exhibit such behavior include the internet, some social networks, protein-protein interactions and disease transmission [cite]. Of particular interest is the region where several major power-law systems notably the internet exist, $2 < \gamma < 3$ [newman]. In this region there are various structural constraints placed upon any networks which form resulting in networks which are far denser than would be expected of scale-free networks [cite all scale free networks]. Furthermore, it has been shown that a critical point in the graphically of power-law networks exists where no sequences are graphical when $0 < \gamma < 2 \; and \; N \to \infty$ [all scale free sparse].

It was previously unclear what bias SC methods introduced to the structural characteristics of generated ensembles when used to generate power-law networks in the range $2 < \gamma < 3$ i.e. local clustering coefficients and number of loops lengths 4. We show that a major soft-constrained method, the random exponential graph model (ERGM) is biased toward more clustered networks in this region and discuss the reasons and implications of these findings [newman].

II. METHODS

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, $\langle x_i \rangle = \sum_{G=\zeta} P(G)\, x_i(G)$ [newmans book]. In the case where a prescribed degree sequence k is given, the constraints on the ensemble become $\langle k_i \rangle = \sum_{G=\zeta} P(G) k_i(G)$. 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,

$$p_{ij} = \frac{1}{1 + e^{-}(\beta_i + \beta_j)}, \tag{1}$$

[newmans]. Therefore the average degree a node i will be,

$$\langle k_i \rangle = \sum_j \frac{1}{1 + e^{-}(\beta_i + \beta_j)}, \tag{2}$$

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 Eq. 3 lead to Eq. 4,

$$e^{-(\beta_i+\beta_j)} \gg 1, \tag{3}$$

$$p_{ij} \simeq e^{\beta_i} e^{\beta_j}, \tag{4}$$

enabling the β values to be estimated using,

$$\beta_i = log(\frac{k_i}{\sqrt{\sum_j k_j}}). \tag{5}$$

The Chung-Lu model is a simplification of the ERGM which utilizes Eq. 4 to show that,

$$p_{ij} = \frac{\langle k_i \rangle \langle k_j \rangle}{\sum_w \langle k_w \rangle} \tag{6}$$

where

$$k_{max}^2 < \sum_w \langle k_w \rangle.[citations?newmanCL] \tag{7}$$

This simplification allows one to avoid the non-trival task of solving the non-linear system of equations involving β which is required by ERGM.

The hard-constrained method we utilized was the Markov-Chain Monte Carlo Method (MCMC). – probably should ask weibin to help write this because he knows about this

III. SOFT-CONSTRAINED ENSEMBLES

In the context of power-law networks there were several issues concerning the soft-constrained methods, ERGM and Chung-Lu. Ideally it would have been preferable to use Chung-Lu to generate the ensembles of networks, however we have found that it is incapable of producing networks where $2 < \gamma < 3$ a problem acknowledged by few others when dealing with real-world networks, many of which lie in this critical region [cite pg2 winlaw cite for newman internat gamma value]. The percentage of sequences which conform to Eq. 6 approaches zero as $\gamma \to 0$ see figure 1. This occurs because there is a significant transition in power-law networks in the region where $\gamma < 3$.

For scale-free network, the probability of a node having degree k, and where N represents the sequence size and the maximum possible degree, is $P(k) = \frac{k^{-\gamma}}{H_{N-1,\gamma}}$, where $H_{a,b}$ is the a^{th}

generalized harmonic number of exponent b, $H_{a,b} = \sum_{t=1}^{a} t^{-b}$. When $N >> 1$ and $\gamma > 1$,

$$H_{N-1,\gamma} = \sum_{t=1}^{N-1} t^{-\gamma} \approx \int_{1}^{N} t^{-\gamma} dt = \frac{N^{1-\gamma} - 1^{1-\gamma}}{1-\gamma} \approx \frac{1}{\gamma - 1} \tag{8}$$

The expected maximum degree of a scale free sequence is?

$$\hat{k} = max\{x : N \sum_{i=x}^{N-1} \frac{i^{-\gamma}}{H_{N-1,\gamma}} \geq 1\} \tag{9}$$

When $N >> 1$ and $\gamma > 1$

$$N \int_{i=x}^{N-1} \frac{i^{-\gamma}}{H_{N-1,\gamma}} di = 1 \tag{10}$$

$$\frac{N}{(1-\gamma)H_{N-1,\gamma}} [(N-1)^{1-\gamma} - x^{1-\gamma}] = 1 \tag{11}$$

$$x = [\frac{N}{(1-\gamma)H_{N-1,\gamma}}]^{\frac{1}{\gamma-1}} \sim N^{\frac{1}{\gamma-1}} \tag{12}$$

When $N >> 1$ and $\gamma > 2$,

$$\sum_{i=1} k_i = N\bar{k} \approx N\frac{\gamma-1}{\gamma-2} \sim N \tag{13}$$

Chung-Lu model requires $k_{max}^2 < \sum_{i=1}^{N} k_i$. Since when $N >> 1$ and $\gamma > 2$, we have $\hat{k} \sim N^{\frac{1}{\gamma-1}}$, $\sum_{i=1}^{N} k_i \sim N$, the Chung-Lu model requires,

$$N^{\frac{2}{\gamma-1}} < N \tag{14}$$

which is true only when $\gamma > 3$.

Considering the implications of Eq. 14 we instead relied upon the ERGM, which requires solving the non-linear system of equation laid out by Eq. 2. Often non-linear system solvers require an initial guess, expecting the β values to conform to the predictions laid out by Eq. 5, however as $\gamma \to 2$ convergence upon the solutions to the system becomes more difficult. Based upon our results see figures 2, 3, and 4, the β values radically diverge from the predicted values made using Eq. 5. Naturally, we expect that as $N \to \infty$ no graphical sequences will exist, however there are distinct structural transitions which occur between $2 < \gamma < 3$ where networks become denser and can no longer be considered sparse [cite scalefreenetworks]. Because Eq. 5 relies upon the assumption made in Eq. 3 that such networks are sparse we cannot expect Eq. 5 to produce the correct results.

Although the resulting β values differ from the expected values, the ERGM is still able to produce an ensemble which meets the prescribed degree sequence see figure 5. Therefore, while there may yet exist other viable solutions, those which we obtained are viable. Of particular interest is whether the ensemble produces structural characteristics such as local and global clustering coefficients (LCC and GCC) and number of squares which converge to a stable value with increasing ensemble size. Figures 6, 7 and 8 indicate that the ensemble produces such values which quickly converge to a stable value–within a few hundred networks.

– include distributions of the gcc, lcc and of squares