Linear and optimization Hamiltonians in clustered exponential random graph modeling

J. Stat. Mech. (2011) P08008

(http://iopscience.iop.org/1742-5468/2011/08/P08008)

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

Linear and optimization Hamiltonians in clustered exponential random graph modeling

Juyong Park and Soon-Hyung Yook

Department of Physics, Kyunghee University, Seoul, Korea
E-mail: perturbation@gmail.com (J Park) and syook@khu.ac.kr

Received 18 January 2011
Accepted 6 July 2011
Published 24 August 2011

Online at stacks.iop.org/JSTAT/2011/P08008
doi:10.1088/1742-5468/2011/08/P08008

Abstract. Exponential random graph theory is the complex network analog of the canonical ensemble theory from statistical physics. While it has been particularly successful in modeling networks with specified degree distributions, a naïve model of a clustered network using a graph Hamiltonian linear in the number of triangles has been shown to undergo an abrupt transition into an unrealistic phase of extreme clustering via triangle condensation. Here we study a nonlinear graph Hamiltonian that explicitly forbids such a condensation and show numerically that it generates an equilibrium phase with specified intermediate clustering.

Keywords: random graphs, networks, optimization over networks, statistical inference

J. Stat. Mech. (2011) P08008

1742-5468/11/P08008+13$33.00

Contents

1. Introduction

The study of complex systems found in various disciplines including engineering, biology and sociology that can be represented as networked systems composed of nodes and edges has garnered much interest from statistical physicists in recent years. Building upon a rich and long tradition of studies on many-body systems, they have successfully adapted analytical and computation tools to understand networks [1]–[3].

A network modeling methodology that shows a striking resemblance to the canonical ensemble theory from statistical physics is the exponential random graph (ERG) theory, originally developed in statistics and currently the most actively studied in social network analysis (SNA) circles [4]–[6]. Given that the potential readership of this paper will be composed of statistical physicists, the premise of ERG is perhaps most simply explained using the language of statistical physics. Here, as in the canonical ensemble theory, one considers an ensemble Γ of graph configurations (microstates) G whose probabilities in Γ are given by $P(G) = \sum_{G\in\Gamma} \mathrm{e}^{-H(G)}/Z$, where $H(G)$ is the *graph Hamiltonian*, a function of network characteristics of G, and $Z = \sum_{G\in\Gamma} \mathrm{e}^{-H(G)}$ is the partition function. Both in social network analysis and in statistical physics, the Hamiltonian $H(G)$ is typically set up to be a linear function of *network variables* or *network statistics* such as the number of edges $m(G)$ in the graph. When the network is simple and unweighted (i.e. the number of edges between two nodes is either 0 or 1) it is straightforward to show that $H(G) = \theta m(G)$ generates the so-called Erdös–Rényi random graph in which two nodes are connected with probability $p = 1/(1 + \mathrm{e}^{\theta})$ [5]. The expected number of edges $\overline{m}$ in a network of n nodes is in this case, therefore, given as

$$\overline{m} = \binom{n}{2} p = \frac{n(n-1)}{2} \frac{1}{1+\mathrm{e}^{\theta}}, \tag{1}$$

controlled by the conjugate variable θ. If one then equates $\overline{m}$ from equation (1) with the actual number of edges m in the network data under study, this serves as the *null model* of the network under study with the number of edges as the only observable. Note again that

J. Stat. Mech. (2011) P08008

only the number of edges m is an explicit variable in constructing the network ensemble[1]; whether the model is sufficient (i.e., is a good approximation of network data) is to be judged on the given model's ability to reproduce other (not used as input to the model) network characteristics such as the degree distribution, cluster size distribution, degree–degree correlation, etc. A significant disagreement between the expected characteristic of a model and the data may indicate that the choice Hamiltonian needs to be reformulated; for instance, the ubiquity of scale-free (power-law) networks where the degree distribution is fat-tailed renders the simplest Erdös–Rényi network model (which has a Poissonian degree distribution) inadequate, necessitating the introduction of alternative forms of the graph Hamiltonian. One possibility is to incorporate explicitly the node degrees e$\{k_i\}$ ($i \in \mathcal{N} = 1, \ldots, n$ is the node index) themselves to form the so-called *linear degree Hamiltonian* $H_{\mathrm{LD}}(G)$:

$$H_{\mathrm{LD}}(G) = \theta_1 k_1(G) + \cdots + \theta_n k_n(G), \tag{2}$$

where $\{\theta_i\}$ are the conjugate variables that now control the expected degrees $\{\overline{k_i}\}$ in a manner similar to what θ did to $\overline{m}$ in equation (1).[2] On a historical note, the study of H_{LD} was prompted by the hypothesis that heavily skewed degree distributions such as the power law may cause the observed negative correlation between degrees of connected nodes, while the Erdös–Rényi network produces no such correlation in the thermodynamic limit ($n \to \infty$) [5]. On the other hand, it was shown analytically that power-law networks generated via equation (2) exhibited negative degree correlation, proving the hypothesis, and thus that the skewed degree distribution was indeed responsible for the negative degree–degree correlation. This is, in fact, a typical example of the ERG modeling (also of the general statistical modeling) procedure—identifying 'important' features of the observed system and testing its sufficiency via comparing the model's predictions and real data (i.e. the 'goodness of fit' of the model in the statistical sense; see [9]) and, when a closer agreement is desired, refining the hypothesis and repeating the procedure. This process is presented schematically in figure 1.

Not surprisingly, the development of ERG as a network modeling framework closely follows the study of graph Hamiltonians of increasing complexity. ERG models of historical import include, in addition to the simplest $H(G) = \theta m(G)$, the Holland and Leinhardt model of reciprocity, the Strauss model of clustering, the 2-star model, and the generalized k-star models [5]. We refer interested readers to introductory articles and significant recent work from the SNA community for more detail [6]–[8], [9].

To a statistical physicist, the benefits of such a formalism are obvious: one can utilize appropriate computational (such as the Metropolis–Hastings algorithm) and analytical (such as the Feynman-diagrammatic method) tools to study the properties of the model [5, 10, 11]. It should also be noted that the Hamiltonian need not be linear at all. For instance, when one wishes to construct an exponential random graph model of a network with a specified degree sequence, $H(G)$ only needs to be a function of the node degrees $\{k_i(G)\}$ in G, i.e. $H(G) = H[\{k_i(G)\}]$ where $i \in \mathcal{N} = \{1, \ldots, n\}$ is the node index. This is sufficient to guarantee that two configurations G, G' with an identical degree sequence have

[1] Typically we consider the number of nodes n as given.

[2] Note the absence of the temperature β in equations (1) and (2). Here, one may consider β as having been absorbed into $\{\theta\}$.

J. Stat. Mech. (2011) P08008

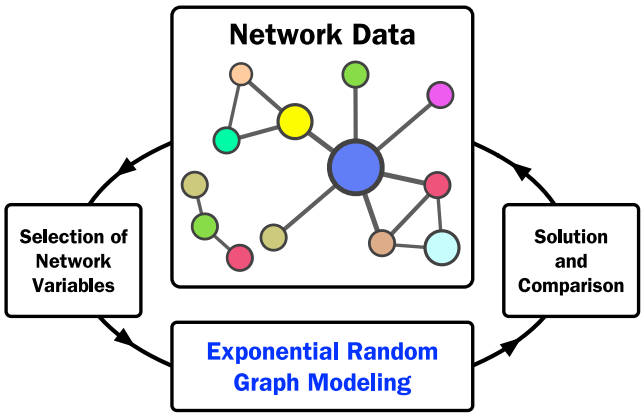

Figure 1. The schematics of the exponential random graph modeling of network data. From the network data of interest (top) one selects network variables such as the node degrees $\{k_i\}$ (left) from which one then forms a Hamiltonian $H(\{k_i\})$, whose solutions and predictions are compared with the network data. Significant disagreements may necessitate a new selection of variables or reformulation of the Hamiltonian.

the same probability in the ensemble, and the aforementioned $H_{\rm LD}$ is one possibility. Thus there is much freedom in choosing the form of the $H(G)$, meaning that there exist ample avenues for exploration of various possible forms of Hamiltonians as one sees fit, not limited to linear forms. In fact, linear forms such as $H_{\rm LD}$ of equation (2) are often not robust in the presence of a perturbation, in the sense that when a composite Hamiltonian $H = H_{\rm LD} + H'$ is used the equilibrium degree distribution may differ significantly from the one specified from $H_{\rm LD}$, defeating the modeler's intention to generate a desired degree distribution using $H_{\rm LD}$. The purpose of this paper is to review the clustering perturbation and compare the characteristics of linear and nonlinear Hamiltonians under it. For simplicity, we here consider only unweighted and undirected graphs.

2. Degree Hamiltonians

Here we briefly review $H_{\rm LD}(G) = \sum_i \theta_i k_i$, equation (2), specifically when the network is *sparse* ($k_i \sim O(1) \ll \sqrt{n}$). In such a case it is well known that the probability p_{ij} that nodes i and j are connected is $\mathrm{e}^{-\theta_i}\mathrm{e}^{-\theta_j}$, leading to the average degree $\langle k_i \rangle$ of node i [5]

$$\begin{aligned}\langle k_i \rangle &= \sum_{j\neq i} p_{ij} = \mathrm{e}^{-\theta_i} \sum_{j\neq i} \mathrm{e}^{-\theta_j} \\ &= (n-1)\mathrm{e}^{-\theta_i} \int_{-\infty}^{\infty} \mathrm{e}^{-\theta} \rho(\theta)\, \mathrm{d}\theta \equiv A(n-1)\mathrm{e}^{-\theta_i}, \end{aligned} \tag{3}$$

where the latter integral form is valid for a large network ($n \gg 1$), $\rho(\theta)$ is the distribution density of θ, and $A \equiv \int_{-\infty}^{\infty} \mathrm{e}^{-\theta}\rho(\theta)\,\mathrm{d}\theta$ is thus a constant. Setting $\langle k_i \rangle = q_i$, the specified (desired) degree of node i and inverting equation (3), we obtain $\theta_i = -\ln(q_i/A(n-1))$.

J. Stat. Mech. (2011) P08008

H_{LD} then becomes

$$\begin{aligned} H_{\mathrm{LD}}(G) &= \sum_{i\in\mathcal{N}} \theta_i k_i(G) \\ &= -\sum_{i\in\mathcal{N}} k_i(G)\ln q_i + \ln(A(n-1))\sum_{i\in\mathcal{N}} k_i(G) \\ &= -\sum_{i\in\mathcal{N}} k_i(G)\ln q_i + 2M(G)\ln(A(n-1)), \end{aligned} \tag{4}$$

where $M(G) = \frac{1}{2}\sum_i k_i(G)$ is the number of edges in G. One can also show that the ensemble generated via H_{LD} is equivalent to the *configuration model*, a popular and useful framework for studying graphs with arbitrary degree distributions [12].

Now, if we restrict the ensemble $\Gamma = \{G\}$ to contain only network configurations G with a fixed number of edges $M(G) = M_0 = \frac{1}{2}\sum_i q_i$ (corresponding to the canonical ensemble of particles), the second term $2M(G)\ln A(n-1)$ becomes a constant. Therefore, we can safely ignore it and use an even simpler form

$$H_{\mathrm{LD}}(G) = -\sum_{i\in\mathcal{N}} k_i(G)\ln q_i. \tag{5}$$

This is particularly useful in edge-conserving Monte Carlo simulations, where the Metropolis–Hastings algorithm would consist of relocating the edge between a randomly selected connected node pair to between a randomly selected unconnected pair with probability 1 if it results in a lower energy, and with probability $\mathrm{e}^{-\Delta H(G)} < 1$ when it results in a higher energy.

It is important to note that it is the ensemble average $\langle k_i\rangle$ of a node that is to be matched with its prescribed degree q_i, and there is no guarantee that $k_i = q_i$ strictly, even at equilibrium: in fact, $P(k_i|q_i)$, the probability that a node with a prescribed degree q_i has degree k_i at equilibrium, is

$$P(k_i|q_i) = \sum_{\{\mathcal{N}_k\}}\Big[\prod_{j\in\mathcal{N}_k}\mathrm{e}^{-(\theta_j+\theta_i)}\prod_{l\in\mathcal{N}'_k}(1-\mathrm{e}^{-(\theta_l+\theta_i)})\Big], \tag{6}$$

where $\theta_i = -\ln q_i/A(n-1)$, and $\{\mathcal{N}_k\}$ is the set of all possible combinations of k nodes from $\mathcal{N}$ excluding i. From this, the total degree distribution $P(k)$ in equilibrium is given as

$$P(k) = \sum_{\{q\}} P(k|q)P(q), \tag{7}$$

where $P(q)$ is the prescribed degree distribution. It is unlikely that $P(k = q|q) \equiv 1$ in equation (6), and thus we cannot expect $P(k) \equiv P(q)$. To find the general characteristics of $P(k)$ from equation (7) in comparison with $P(q)$, we performed a Monte Carlo simulation (using the Metropolis–Hastings method described above) of H_{LD} for a network of $n = 500$ and $P(q = 5) = P(q = 15) = \frac{1}{2}$ for illustrative purposes, whose results are shown in figure 2. In the figure, the prescribed $P(q)$ is shown in gray, and the equilibrium $P(k)$ is shown in blue. While $P(k)$ does exhibit peaks at $k = 5$ and 15, it also shows a fairly wide distribution (although small in comparison with n), and the fluctuation is visibly larger at $k = 15$, resulting in a lower peak. If the specified degree q had been the

J. Stat. Mech. (2011) P08008

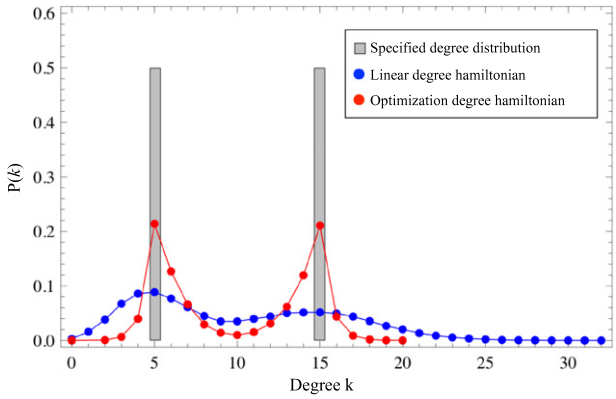

Figure 2. The degree distributions from exponential random graph simulations. For simplicity we set the specified degree distributions to be $P(q = 5) = \frac{1}{2}$ and $P(q = 15) = \frac{1}{2}$ (shown in gray). The linear degree Hamiltonian $H_{\rm LD} = -\sum_{i\in\mathcal{N}} k_i \ln q_i$ generates a smooth distribution over a wide range of degrees with Poissonian-like peaks at $k = 5$ and 15 (blue). The degree distribution from the optimization degree Hamiltonian $H_{\rm OD} = \sum_{i\in\mathcal{N}} |k_i - q_i|$, by contrast, is noticeably closer to the specified one, with sharper peaks of roughly equal heights at $k = 5$ and 10.

same for all nodes (i.e. a q-regular graph) it is well known that $H_{\rm LD}$ would have created an Erdös–Rényi graph with a Poissonian degree distribution, locally not unlike the peaks in figure 2 [5]. Thus we call the peaks we see in figure 2 Poisson-like.

The well-documented success of the configuration model implies that the fluctuations we see in $P(k)$ may not be problematic in general, though in certain circumstances (we see such a case later) a more faithful reproduction of the specified degree distribution may be desired. This means that a graph Hamiltonian is needed that imposes a larger penalty when k_i deviates from q_i than $H_{\rm LD}$ does. It is unclear how $H_{\rm LD}$ can be modified while retaining the linear form. Instead, we introduce a nonlinear Hamiltonian

$$H_{\rm OD}(G) = \sum_{i\in\mathcal{N}} \beta_{\rm d} |k_i(G) - q_i| \tag{8}$$

which we call the *optimization degree Hamiltonian*, being reminiscent of Hamiltonians used in certain optimization problems such as number partitioning [13][3]. The $P(k)$ that results from $H_{\rm OD}$ with $\beta = 1$ for simplicity (the penalty can be controlled via $\beta_{\rm d}$ when necessary) is shown in figure 2 in red, which is indeed a more faithful reproduction of $P(q)$ in comparison with $H_{\rm LD}$, showing sharper peaks at $k = 5$ and 15 of equal heights similar to $P(q)$. The broadening of the peaks around the specified degrees from the $H_{\rm LD}$ in comparison to $H_{\rm OD}$ in figure 2 is persistent in cases of more heterogeneous (thus less artificial) specified $P(q)$, as seen in figure 3 where we compare $H_{\rm LD}$ and $H_{\rm OD}$ for a Poissonian $P(q)$ and a double Gaussian

$$P(q) = \alpha\Phi(q; \mu_1, \sigma_1) + (1 - \alpha)\Phi(q; \mu_2, \sigma_2), \tag{9}$$

[3] Note that the so-called 'curved' exponential random graph model is similar to our formalism in that the graph Hamiltonians are nonlinear functions of network variables [9].

J. Stat. Mech. (2011) P08008

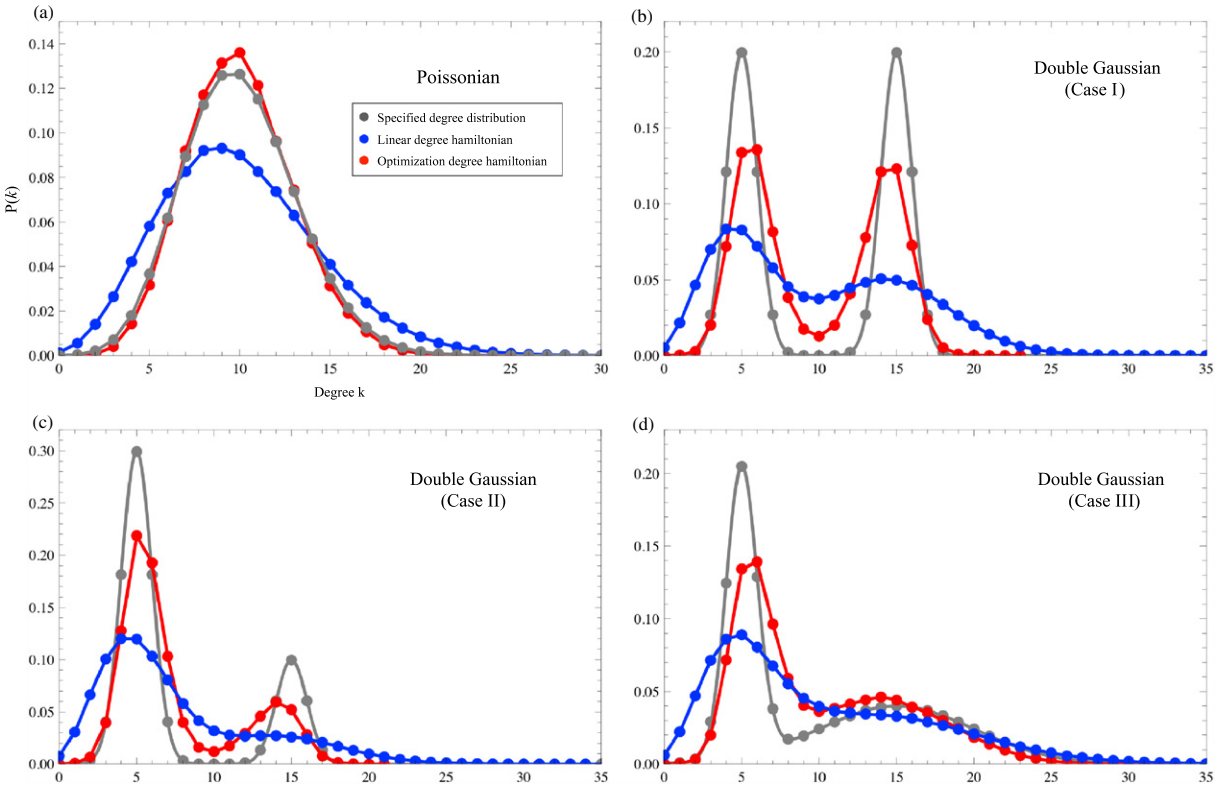

Figure 3. The degree distributions generated via $H_{\rm LD}$ and $H_{\rm OD}$ for heterogeneous specified degree distribution. In (a) $P(q)$ is a Poissonian (with $\langle q\rangle = 10$). In (b)–(d) $P(q)$ is a double Gaussian with peaks at $q = 5$ and 10 with varying relative heights ($\alpha \in [0,1]$ for the peak at $q = 5$, and $1-\alpha$ for the peak at $q = 10$) and variances σ_1, σ_2 of the peaks. (b) $(\alpha,\sigma_1,\sigma_2) = (0.5, 1.0, 1.0)$. This is the most similar to figure 2. (c) $(\alpha,\sigma_1,\sigma_2) = (0.75, 1.0, 1.0)$ and (d) $(\alpha,\sigma_1,\sigma_2) = (0.5, 1.0, 5.0)$. Here, $H_{\rm OD}$ again consistently reproduces $P(q)$ more faithfully.

where $\Phi(q,\mu,\sigma)$ is a Gaussian of mean μ and variance σ^2, and $\alpha \in [0,1]$ sets the relative weights between the two Gaussian peaks. For the Poissonian case we set $\langle q\rangle = 10$ (figure 3(a)), and for the double Gaussian we try three cases of varying weights and variances (figures 3(b)–(d)). The behaviors of $H_{\rm LD}$ and $H_{\rm OD}$ are consistent with what we see from figure 2: in terms of the goodness of fit to $P(q)$ (including the relative heights at the peaks) $H_{\rm OD}$ is superior to $H_{\rm LD}$.[4]

3. Robustness of degree reproduction under perturbation: targeted clustering

Besides degree distribution, a network characteristic that has been widely studied is clustering. Intuitively, a clustered network contains significantly more triadic closures

[4] For the purposes of this paper we are showing some numerical examples. For more systematic studies one could investigate various moments of the degree distributions from the two Hamiltonians, or a difference measure between two distributions P and Q such as $D(P,Q) \equiv \sum_k |P(k) - Q(k)|$.

J. Stat. Mech. (2011) P08008

(triangles) than expected in a random graph with the same number of edges. (A common definition of the strength of clustering of a network is given using the so-called *clustering coefficient*, which we present later.)

In exponential random graph literature, studies have been made on graph Hamiltonians that incorporate the number of triangles T linearly, the simplest case being the Strauss model with $H_{\mathrm{S}}(G) = \theta M + \tau T$ [14]. The motivation for H_{S} is that by controlling θ and τ, one could hopefully generate a network with any desired value of M and T, i.e. a smooth, controllable transition between a non-clustered configuration (small T) and a clustered one (large T). Unfortunately, it has been shown that H_{S} does not show such a behavior: depending on θ and τ, the system undergoes a first-order phase transition from a sparse ER-like phase with vanishing clustering to a nearly fully connected phase [14, 15], while most real networks are neither. More recently, Foster *et al* performed an extensive study of the Hamiltonian $H(G) = \tau T$ on an ensemble of networks of fixed degree sequences (and thus a fixed number of edges), and found that as τ is tuned, T shows a series of jumps consisting of first-order phase transitions [16], each transition indicating the formation of densely connected local cliques.

This pathology renders the linear Hamiltonian for modeling real clustered networks, where the triangles are distributed over the network without such extreme 'condensation' of triangles. The lack of such an intermediate phase in H_{S} stems from the fact that the addition of a single edge in an already densely connected part of the network can lead to a disproportionately large increase in T and decrease in $H_{\mathrm{S}}(G)$, resulting in the condensed phase being energetically favorable. Therefore, it is understood that a Hamiltonian or, more generally, a mathematical formalism is necessary that explicitly discourages such condensation [15, 17].

Before we find such a Hamiltonian in our context of exponential random graphs, let us first review how clustering in networks is quantified. It is often done via the *clustering coefficient* C. In wide use are three versions, one local (node level) and two global (network-wide). On the individual node level, the *local* clustering coefficient is defined as

$$C_i \equiv \frac{t_i}{s(k_i)} = \frac{t_i}{(1/2)k_i(k_i-1)} \qquad (k_i \leq 2), \tag{10}$$

where t_i is the number of triangles of which the node i is at a corner, and $s(k_i) = \binom{k_i}{2}$ is the number of pairs of neighbors of node i, also called two-stars centered on i. C_i is therefore the probability that two neighbors of node i are themselves neighbors. The global measure of clustering is commonly given by two measures. One is the average of C_i which we write as $\overline{C}$, defined as $\overline{C} \equiv \langle C_i \rangle = \sum_{i\in\mathcal{N}} C_i/N$, i.e. the average of the local clustering coefficients. The other, which we call $\widetilde{C}$, is defined as

$$\widetilde{C} \equiv \frac{3T}{\sum_{i\in\mathcal{N}} s(k_i)} = \frac{3T}{\sum_i (1/2)k_i(k_i-1)}, \tag{11}$$

where $T = \frac{1}{3}\sum_{i\in\mathcal{N}} t_i$ is again the number of triangles in the network. Therefore this is the probability that a randomly selected two-star is a part of a triangle (3 exists in the numerator because one triangle contains three two-stars). Although $\widetilde{C}$ and $\overline{C}$ are not identical, $\widetilde{C} = \overline{C}$ when $C_i = C_0$ for all i. In terms of these quantities, the aforementioned behavior of the Strauss Hamiltonian $H_{\mathrm{S}} = \theta M + \tau T$ can be summarized

J. Stat. Mech. (2011) P08008

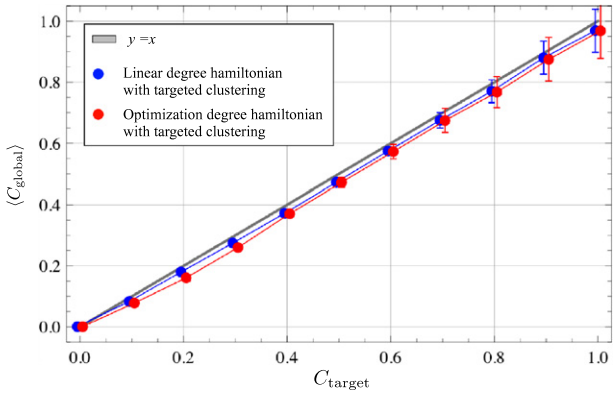

Figure 4. The global clustering $\langle\widetilde{C}\rangle$ of graph ensembles generated from the linear (blue) and the optimization (red) degree Hamiltonians perturbed with the targeted clustering Hamiltonian $H_{\rm C} = \sum_{i\in\mathcal{N}} |t_i - C_{\rm target} s(k_i)|$. $\langle\widetilde{C}\rangle \simeq C_{\rm target}$ is the result of the node-level local clustering coefficients being $\simeq C_{\rm target}$, regardless of the degree distribution.

as the clustering coefficient (local or global) being either $\widetilde{C}$ (or $\overline{C}$) $\simeq 0$ (sparse ER-like phase) or $\widetilde{C}$ (or $\overline{C}$) $\simeq 1$ (condensed phase) or, in other words, $t_i \simeq 0$ or $s(k_i)$ for all i, while in a network of intermediate clustering coefficient C, t_i would be $\sim C s(k_i)$. Taking a cue from the latter and equation (8), we propose the following nonlinear targeted clustering Hamiltonian:

$$H_{\rm C} = \sum_{i\in\mathcal{N}} \beta_{\rm c} |t_i - \gamma_i s(k_i)| = \sum_{i\in\mathcal{N}} \beta_{\rm c} |t_i - \gamma_i \tfrac{1}{2} k_i (k_i - 1)|, \tag{12}$$

where γ_i is now the specified (i.e. targeted) clustering coefficient for node i. The difference between $H_{\rm C}$ and the model of Milo *et al* [18], where $H_{\rm Milo} = |T - T'|$ and T' is the specified number of triangles, is that $H_{\rm C}$ allows us to control local clustering. Similarly to $H_{\rm OD}$ of equation (8), $H_{\rm C}$ explicitly penalizes t_i when it diverges from a prescribed value $\gamma_i s(k_i)$. In studying the effectiveness of $H_{\rm C}$ in reproducing the specified local clustering, we would also like to have the option of controlling the degrees $\{k_i\}$ simultaneously. We have already discussed two Hamiltonians designed specifically for that purpose, $H_{\rm LD}$ and $H_{\rm OD}$. In the remainder of this paper, therefore, we study the following two composite Hamiltonians:

$$H_1 = H_{\rm LD} + H_{\rm C} = \sum_{i\in\mathcal{N}} [-k_i \ln q_i + \beta_{\rm c} |t_i - \gamma_i s(k_i)|] \tag{13}$$

and

$$H_2 = H_{\rm OD} + H_{\rm C} = \sum_{i\in\mathcal{N}} [\beta_{\rm d} |k_i - q_i| + \beta_{\rm c} |t_i - \gamma_i s(k_i)|] \tag{14}$$

to find out whether either is capable of generating network ensembles exhibiting both the specified degrees and local clustering.

A Monte Carlo simulation was performed for a network of size $n = 500$ and $\langle k\rangle = 10$. For simplicity, we again set $\beta_{\rm d} = \beta_{\rm c} = 1$, $P(q) = \delta_{q,10}$ (a regular graph), and $\gamma_i = C_{\rm target}$, a universal value for all i, varied between 0 and 1. First, figure 4 shows

J. Stat. Mech. (2011) P08008

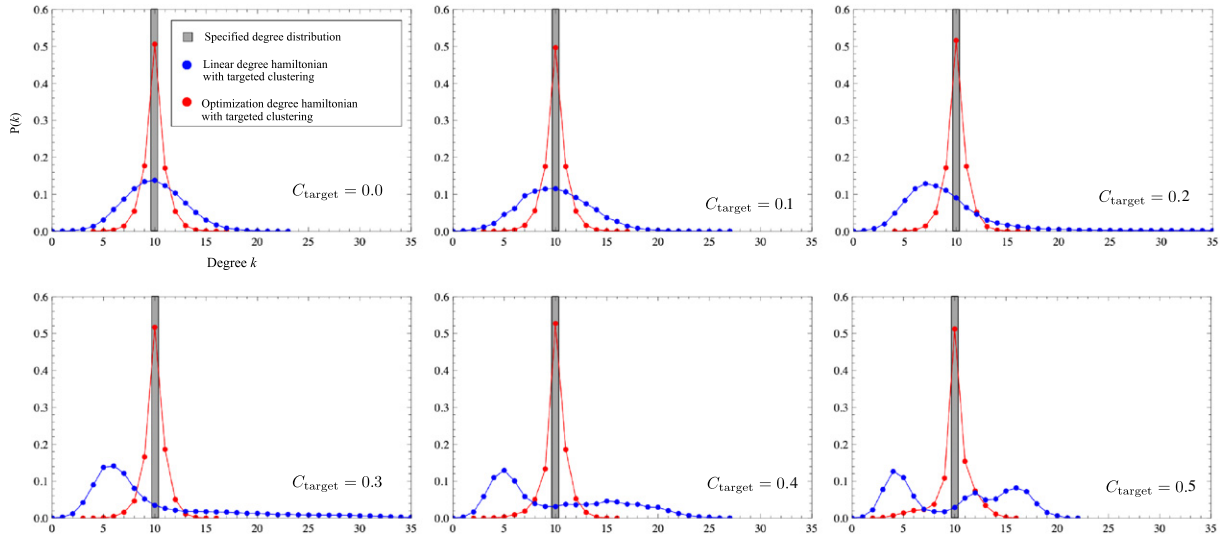

Figure 5. The equilibrium degree distributions $P(k)$ generated from $H_1 = H_{\text{LD}} + H_{\text{C}}$ (blue) and $H_2 = H_{\text{OD}} + H_{\text{C}}$ (red) for various values of C_{target}. The specified degree distribution $P(q) = \delta_{q,10}$ is in gray. When $C_{\text{target}} = 0$, both Hamiltonians generate their natural $P(k)$—a true Poissonian for H_{LD}, and a sharp peak for H_{OD}. As C_{target} is tuned higher, however, $P(k)$ peaks at a smaller k for $H_{\text{LD}} + H_{\text{C}}$ and even exhibits multiple peaks when C_{target} is too large, while it stays virtually unchanged for $H_{\text{OD}} + H_{\text{C}}$ up to $C_{\text{target}} \simeq 0.5$, an unusually high value in real networks.

the mean global clustering $\langle \widetilde{C} \rangle$ from the simulation, which shows us that both H_1 and H_2 $\langle \widetilde{C} \rangle \simeq C_{\text{target}}$ generate networks with the specified clustering. This arises from the fact that $\langle C_i \rangle \simeq C_{\text{target}}$ on the individual level as well (not shown). The difference between the H_1 and H_2, however, is most striking in the equilibrium $P(k)$, shown in figure 5. When $C_{\text{target}} = 0$, perturbation H_{C} is insignificant since the expected clustering without it is 0 anyway, and therefore $P(k)$ is simply as expected—a true Poissonian for H_1, and a sharper peak at $q = 10$ for H_2, similar to the ones we saw in figure 2. When $C_{\text{target}} \neq 0$, on the other hand, the peak in $P(k)$ under H_1 gradually shifts toward a smaller k while high-degree nodes are created in order to compensate for the number of edges M which is a constant. As C_{target} is tuned higher it resembles the specified distribution less and less, and at $C_{\text{target}} \simeq 0.4$ we even observe multiple peaks (at $k = 5$ and 15—the values for which $|t - C_{\text{target}} s(k)| = 0$, meaning the peaks will shift for a different C_{target} and thus are not very meaningful). $P(k)$ under H_2, in contrast, is robust, without noticeable change up to $C_{\text{target}} = 0.5$, already an unusually high value for real-world networks, until it too shows similar (but milder) behavior at a higher value of $C_{\text{target}} \sim 0.6$ and up[5].

Let us now discuss the implications of the findings in figures 4 and 5 on the topology of networks generated from $H_1 = H_{\text{LD}} + H_{\text{C}}$ and $H_2 = H_{\text{LD}} + H_{\text{C}}$. First of all, figure 4 tells us that, unlike the Strauss clustering perturbation τT, H_{C} was able to discourage an extreme condensation of triangles, resulting in $\langle \widetilde{C} \rangle \simeq C_{\text{target}}$ by way of $C_i \simeq C_{\text{target}}$ for both H_1 and H_2. However, it was not enough to completely overcome the cooperative

[5] We performed similar simulations for the four heterogeneous $P(q)$ shown in figure 3 and found similar results.

J. Stat. Mech. (2011) P08008

tendency of triangles under H_{LD}. The telltale sign of this is the creation of high-degree nodes. Figure 5 shows the creation of high-degree nodes in H_1: now many triangles exist between the high-degree nodes, forming a core of densely interconnected high-degree nodes although $C_i \simeq C_{\mathrm{target}}$ as specified. On the other hand, under H_2 where $P(k)$ is sharply peaked at the specified degree $q = 10$ such cores do not exist; with $k_i \simeq q$ and $C_i \simeq C_{\mathrm{target}}$ for all i as specified, H_2 generates a network that truly has a uniform distribution of triangles, lacking any unspecified, accidental local structures.

We check our claim via the following two quantities: the degree–degree correlation r_{deg} (the Pearson correlation between the degrees of adjacent nodes) and the *mean corner degree* of the triangles in the network, shown in figures 6(a) and (b). First, the plot of $\langle r_{\mathrm{deg}}\rangle$ in figure 6(a) indicates that adjacent degrees in the network are highly correlated under H_1, so that high-degree nodes are indeed connected with other high-degree nodes and vice versa, while H_2 shows no such effect. This leads naturally to what we see in figure 6(b): under H_1, the mean corner degree is significantly higher than $\langle k\rangle = q$, unlike H_2 where it is practically equal to q. These observations are presented visually in figure 6(c) (an actual snapshot of an equilibrium configuration of a network with $n = 50$, $P(q) = \delta_{q,5}$, and $C_{\mathrm{target}} = 0.4$; $\langle\widetilde{C}\rangle$ are 0.35 ± 0.02 and 0.31 ± 0.02, respectively). As expected, for H_1 (left) we clearly see that the ten highest-degree nodes (blue, average degree 9.7) form a densely interconnected core (encircled in orange), with the ten lowest-degree nodes (yellow, average degree 1.0) pushed to the periphery with low triangle participation rate. For H_2 (right), no significant difference between highest- and lowest-degree nodes exists, and the triangles are distributed uniformly, expected of a maximally random configuration given the degree and local clustering constraints.

4. Discussion and future directions

Here we have studied two forms of graph Hamiltonian in exponential random graph theory that take node degrees and local clustering as specified input. The tendency of triangles to coalesce in the Strauss model was shown to persist when the linear clustering perturbation was replaced by an optimized clustering form, albeit in a milder fashion, rendering the composite Hamiltonian unable to generate the specified degree distribution[6]. The optimization degree Hamiltonian, on the other hand, was able to satisfy both, exhibiting significant robustness under the same perturbation.

That the optimization Hamiltonian form was able to reproduce both the targeted degree and clustering presents an appealing possibility from the viewpoint of network modeling via exponential random graph theory: given a set of network variables $\Phi = \{\phi_{\mathrm{v}}|v = 1,\ldots,l\}$, it may act as a practical computational method to generate a null model of network data with actual values of the variables $\{\tilde{\phi}_{\mathrm{v}}|v = 1,\ldots,l\}$ using the Hamiltonian [16]

$$H(G) = \sum_{\phi\in\Phi} \beta_{\mathrm{v}}|\phi_{\mathrm{v}}(G) - \tilde{\phi}_{\mathrm{v}}|, \tag{15}$$

[6] The reverse case of $H_{\mathrm{OD}} + \tau T$ was also numerically studied with varying τ. As τ is tuned to be more negative (thus favoring triangles) there also occurs a sudden onset of the emergence of a densely connected cluster of high-degree nodes, signifying a condensation of triangles similar to $H_{\mathrm{LD}} + \tau T$. This demonstrates that to create finite clustering degree and local clustering optimization are necessary.

J. Stat. Mech. (2011) P08008

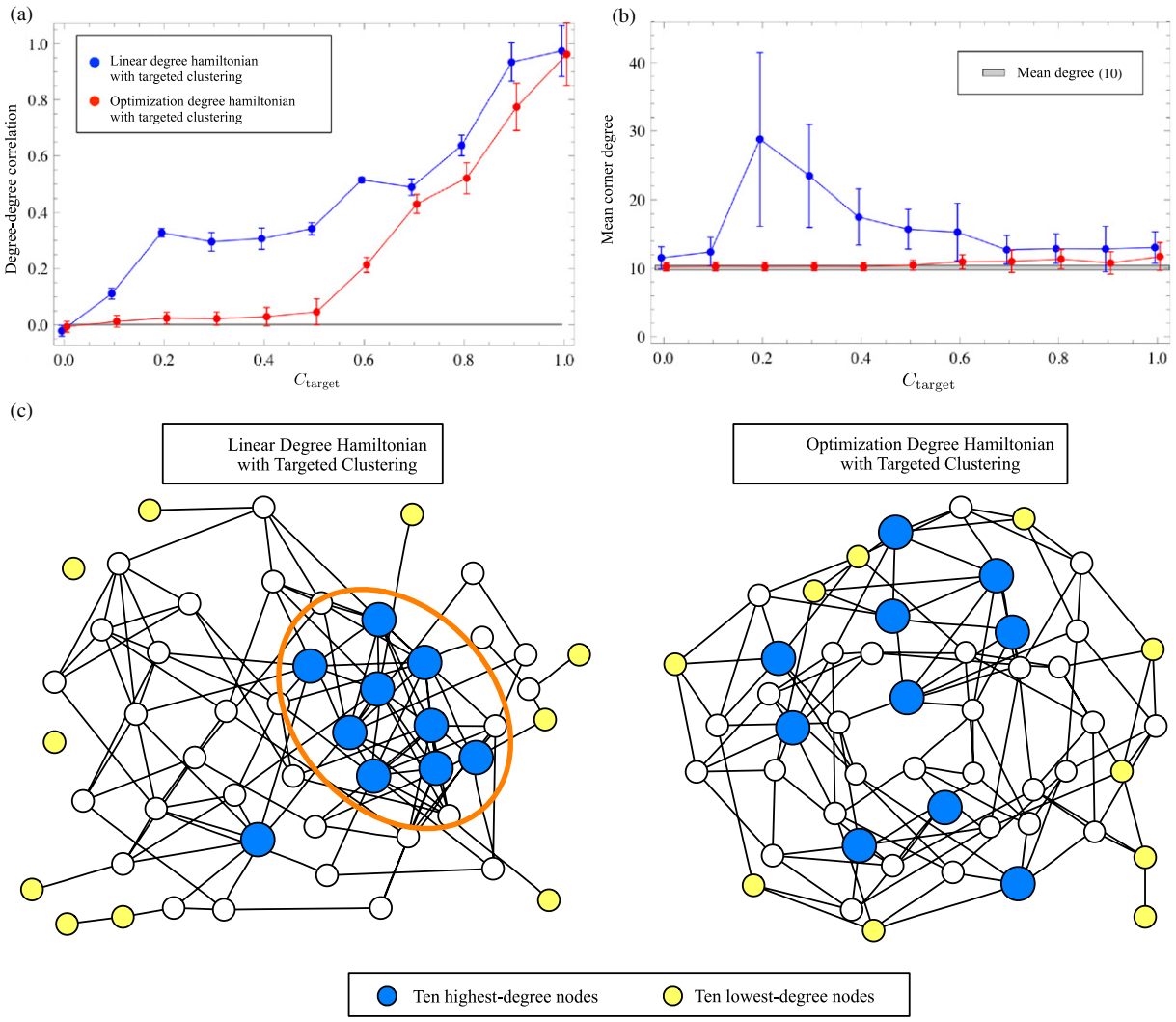

Figure 6. (a) The degree–degree correlation r_{deg} under $H_1 = H_{\text{LD}} + H_{\text{C}}$ and $H_2 = H_{\text{OD}} + H_{\text{C}}$. H_1 generates positive degree correlation for any positive C_{target}, while H_2 exhibits very little correlation up to $C_{\text{target}} \simeq 0.5$. (b) The mean corner degree of triangles contained in the networks in equilibrium. Under H_1 most triangles exist between high-degree nodes, indicating the persistence of the cooperative nature of triangles. (c) Equilibrium topologies of clustered networks under H_1 (left) and H_2 (right). H_1 generates a core of high-degree nodes that are densely connected and share a large number of triangles (enclosed in the orange oval). H_2, in contrast, maintains the specified degree distribution $P(q) = \delta_{q,10}$ while the triangles are distributed uniformly, features expected of a maximally random configuration given the degree and local clustering constraints.

thereby enabling the modeler to assess quickly the sufficiency of the particular set of variables in characterizing the network. An interesting recent application of a related framework was provided by Foster *et al* [19]: specifically, they generated networks with specified global clustering coefficient $\widetilde{C}$ or degree–degree correlation r using the

J. Stat. Mech. (2011) P08008

optimization Hamiltonian and measured their effect on each other and the modular structure of the network (although they kept the degree sequence fixed as the network data). In doing so, they demonstrated the utility of the Hamiltonian of the form (15) in creating network ensembles with desired characteristics. Naturally, more study must be made on the properties of equation (15) in relation to various network variables—global as well as local—in order to establish its general utility. In light of the fact that new, complex measures of network properties are frequently devised and introduced, we hope that the formalism will prove to be a useful tool for network scientists.

Acknowledgments

The authors would like to thank Doochul Kim for helpful comments. This work was supported by Kyung Hee University Grant KHU-20110088 and the Korea Research Foundation Grant KRF-20110005499.

References

[1] Albert R and Barabási A-L, 2002 *Rev. Mod. Phys.* **74** 47
[2] Dorogovtsev S N, Mendes J F F and Samukhin A-N, 2003 *Nucl. Phys.* B **666** 396
[3] Newman M E J, 2008 *Phys. Today* **61** (33) 066117
[4] Frank O and Strauss D, 1986 *J. Am. Stat. Assoc.* **81** 832
[5] Park J and Newman M E J, 2004 *Phys. Rev.* E **70** 066117
[6] Robins G, Snijders T, Wang P, Handcock M and Pattison P, 2007 *Soc. Netw.* **29** 192
[7] Hunter D R, 2007 *Soc. Netw.* **29** 216
[8] Anderson C J, Wasserman S and Crouch B, 1999 *Soc. Netw.* **21** 37
[9] Hunter D, Goodreau S and Handcock M, 2008 *J. Am. Stat. Assoc.* **103** 248268
[10] Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller A, 1953 *J. Chem. Phys.* **21** 1087
[11] Park J, 2010 *J. Stat. Mech.* P04006
[12] Newman M E J, Strogatz S H and Watts D J, 2001 *Phys. Rev.* E **64** 026118
[13] Mertens S, 1998 *Phys. Rev. Lett.* **81** 4281
[14] Strauss D, 1986 *SIAM Rev.* **28** 513
[15] Park J and Newman M E J, 2005 *Phys. Rev.* E **72** 026136
[16] Foster D, Foster J, Paczuski M and Grassberger P, 2010 *Phys. Rev.* E **81** 046115
[17] Newman M E J, 2009 *Phys. Rev. Lett.* **103** 058701
[18] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D and Alon U, 2002 *Science* **298** 824
[19] Foster D V, Foster J G, Grassberger P and Paczuski M, 2009 arXiv:0911.2055

J. Stat. Mech. (2011) P08008