A Deterministic Partition Function Approximation for Exponential Random Graph Models

Wen Pu
LinkedIn Corporation
Mountain View, CA, USA
wpu@linkedin.com

Jaesik Choi
Ulsan National Institute of
Science and Technology, Ulsan, Korea
jaesik@unist.ac.kr

Yunseong Hwang
Ulsan National Institute of
Science and Technology, Ulsan, Korea
yunseong@unist.ac.kr

Eyal Amir
University of Illinois at Urbana-Champaign
Urbana, IL, USA
eyal@illinois.edu

Abstract

Exponential Random Graphs Models (ERGM) are common, simple statistical models for social network and other network structures. Unfortunately, inference and learning with them is hard even for small networks because their partition functions are intractable for precise computation. In this paper, we introduce a new quadratic time deterministic approximation to these partition functions. Our main insight enabling this advance is that subgraph statistics is sufficient to derive a lower bound for partition functions given that the model is not dominated by a few graphs. The proposed method differs from existing methods in its ways of exploiting asymptotic properties of subgraph statistics. Compared to the current Monte Carlo simulation based methods, the new method is scalable, stable, and precise enough for inference tasks.

1 Introduction

Social networks are becoming central components to many aspects of marketing, recruiting, web search, and education programs [Kempe *et al.*, 2003; Cetintas *et al.*, 2011; Mancilla-Caceres *et al.*, 2012]. Careful use of social network analysis in those areas is the key to future advances. For that reason, many researchers and practitioners model their relevant social networks and learn them from data [Carrington *et al.*, 2005]. Many of those social networks are large, and precise modeling of them is difficult. A family of simple models called Exponential Random Graph Models (ERGM) [Robins *et al.*, 2007] is commonly used by researchers and practitioners for this purpose.

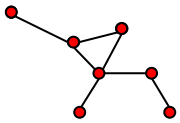
An ERGM defines a distribution over all graphs of n nodes. Coefficients and subgraph statistics, such as number of edges, triangles, and k-stars, are then used to specify ERGM distributions [Robins *et al.*, 2007]. The model captures the correlation of network sub-structures and enables various inferences on complex networks. For example, we can tell whether transitivity is prominent in a network by fitting an ERGM with related subgraphs features, such as triangles.

Learning ERGMs from data is done by Maximum Likelihood Estimation (MLE). Unfortunately, such learning is hard even for networks of modest sizes (e.g. 40 nodes) because calculating normalizing constants (*partition functions*) precisely for such models is intractable. For this reason most current techniques involve sampling using Markov Chain Monte Carlo (MCMC) [Handcock *et al.*, 2003; van Duijn *et al.*, 2009]. This often results in intractable computation or highly imprecise results for these modest-size-or-larger networks [Bhamidi *et al.*, 2008; Snijders, 2002; Handcock, 2003; Hunter *et al.*, 2012].

Recently, Chatterjee and Diaconis [Chatterjee and Diaconis, 2011] derived an analytic approximation to the log-likelihood function of ERGM by applying a new large deviation principle result on Erdös-Rényi models. Chatterjee and Diaconis's theoretical analysis proves a strong convergence result on their approximation, but it only applies to ERGMs with uniformly non-negative/non-positive coefficients for subgraph features.

In this paper, we introduces a new deterministic approximation solution which generalizes Chatterjee and Diaconis's approximation for estimating ERGM partition functions. The two approximations are asymptotically equivalent for large n. However, instead of complex large deviation analysis, we derive the new approximation using much simpler analysis based on weaker convergence results. Moreover, our analysis lifts the unrealistic condition on uniformly non-negative/non-positive subgraph feature coefficients, enabling application of the approximation to the whole subgraph feature coefficient space. Finally, we provide a constructive and pragmatic approach to the problem, which enables us to evaluate the accuracy of the approximation empirically against common stochastic approximation methods in realistic settings.

Specifically, we present a quadratic time (or linear time wrt the number of random variables) deterministic approximation to the log partition function of ERGMs. Asymptotic properties of the subgraph statistics space enable this new approximation. The approximation works as follows: given (coeffi-



feature	count	density
edge	7	0.333 (7/21)
triangle	1	0.029 (1/35)
2-star	11	0.105 (11/105)
3-star	5	0.036 (5/140)
rectangle	0	0 (0/70)

Figure 1: An example network of order $n = 7$ (In ERGM, edges are random variables). Table on the right shows the sufficient statistics (densities) for an ERGM with edge, triangle, 2/3-stars and rectangle as features.

cient) parameters θ of an ERGM, find that edge-count u (between 0 and $\binom{n}{2}$) that maximizes $\tilde{\gamma}(\theta, u) = \theta^T \rho(u) + C(n, u)$ (See (8) for definition), where $\rho(u)$ is a vector of subgraph statistics approximated for graphs with u edges and function $C(n, u)$ approximates the logarithm of the number of graphs with subgraph statistics close to $\rho(u)$. Once the maximizing u is found, we estimate the log partition function $\ln Z(\theta)$ by $\tilde{\gamma}(\theta, u)$. The approximation works because this $\rho(u)$ captures the subgraph statistics of a large (asymptotically) mass of graphs of n nodes. So, in a sense, many graphs would look similar from a subgraph statistics perspective. We also reveal that $\theta = \Theta(n^2)$ is a necessary condition for θ to be asymptotically relevant in an ERGM for network of order n.

Our results show that the new method performs well experimentally comparing to existing sampling methods [Gelman and Meng, 1998; Handcock *et al.*, 2003] on synthetic data and real-world social networks. Our results also show that the new algorithm yields reliable approximation for many models when the size of the network is larger than 30.

The rest of the paper is organized as follows: Section 2 reviews ERGM, Section 3 describes the components of the approximation and key theoretical results, Section 4 describes our experimental evaluation, Section 5 reviews related work and Section 6 concludes the paper.

2 Background

An ERGM defines the following distribution over order-n graphs $g \in \mathcal{G}$:

$$p_\theta(g) = \frac{\exp\left(\theta^T \phi(g)\right)}{Z(\theta)} \text{ and } Z(\theta) = \sum_{g \in \mathcal{G}} \exp\left(\theta^T \phi(g)\right) \quad (1)$$

where $\phi(g)$ is the feature vector for graph $g \in \mathcal{G}$; the parameter θ is a real vector; partition function $Z(\theta)$ is a normalizing constant.

The feature vector $\phi(g)$ may include any network and nodal attributes of g , and the edge statistics is almost always included [Robins *et al.*, 2007]. In this work, we focus on undirected graphs and subgraph statistics features. Specifically, for a set of subgraph structures of interest $\{L_1, \ldots, L_r\}$, the feature vector of undirected graph g can be defined with subgraph densities as below:

$$\phi(g) = \left(\frac{t(g, L_1)}{t(K_n, L_1)}, \frac{t(g, L_2)}{t(K_n, L_2)}, \ldots, \frac{t(g, L_r)}{t(K_n, L_r)} \right) \quad (2)$$

Here $t(g, L_i)$ counts the number of subgraphs in g that are isomorphic to L_i; K_n is the order-n complete graph, therefore $t(K_n, L_i) = \binom{n}{v_i} t(K_{v_i}, L_i)$ is a constant for any L_i of order v_i. Notice that the simplest subgraph K_2, or edge, is almost always included as a feature in ERGMs. Its role in the model is similar to that of the intercept term in most linear regression models [Robins *et al.*, 2007; Hunter, 2007]. For the rest of the paper, we assume K_2 is always included in the feature subgraphs.

Example: Figure 1 illustrates a simple example network of order 7. It has seven edges, one triangle, eleven 2-stars, five 3-stars and no rectangle. The third column shows the subgraph densities of the network. For example, the 7-node labeled graph can have at most $\binom{7}{3} \times 1 = 35$ triangles, therefore the triangle density is $1/35 \simeq 0.029$.

Given a network g, the MLE of parameter vector θ is:

$$\theta^* = \operatorname*{argmax}_\theta \ell(\theta|g) = \operatorname*{argmax}_\theta \left\{ \theta^T \phi(g) - \ln Z(\theta) \right\} \quad (3)$$

In this paper, we are interested in approximation of the log partition function $\ln Z(\theta)$.

3 Approximating Log Partition Functions

In this section, we derive a deterministic approximation to the log partition function $\ln Z(\theta)$ and analyze the behavior of θ for networks of different sizes. We first introduce the counting function for graphs with the same feature vector, which leads to a set of edge-number induced lower bounds of $\ln Z(\theta)$ and its approximation. Figure 2 gives an overview of the new approach.

3.1 Graph counting in the feature space

We introduce the key concept of graph counting function for the feature space of ERGM. Let $\phi : \mathcal{G} \to \mathcal{H}$ be the function that maps a graph in $g \in \mathcal{G}$ to the subgraph density space $\mathcal{H}$ of the graphs. For $\mathbf{h} \in \mathcal{H}$, we define *counting function* $\#(\mathbf{h}) = |\mathcal{G}_h|$ where $\mathcal{G}_h = \{g \in \mathcal{G} | \phi(g) = \mathbf{h}\}$, i.e. the number of graphs in $\mathcal{G}$ having $\mathbf{h}$ as subgraph densities. We re-write the partition function (1) as a compact form using counting function[1]:

$$Z(\theta) = \sum_{\mathbf{h} \in \mathcal{H}} \#(\mathbf{h}) \exp\left(\theta^T \mathbf{h}\right) = \sum_{\mathbf{h} \in \mathcal{H}} \exp\left(\theta^T \mathbf{h} + \ln \#(\mathbf{h})\right) \quad (4)$$

Notice that when θ=0, each term in (4) simply counts the graphs with given subgraph configuration, and the normalizing constant becomes the total number of graphs $|\mathcal{G}|$. Later we will show how the graph counting interpretation helps in computing $\ln Z(\theta)$.

Let $L_1, L_2, \ldots, L_r$ be simple graphs of interest and v_i be the number of nodes for L_i. The following lemma provides an upper bound to $|\mathcal{H}|$. Under the assumption $\forall i,\ n \gg v_i$ and $n \gg r$, the lemma establishes reasonable error bounds for several arguments in the rest of the paper:

Lemma 1. *For* $v* = \max\{v_1, \ldots, v_r\}$*, it holds that* $\ln |\mathcal{H}| \leq r v^* \ln n$.

[1]Note that all isomorphic graphs have the same subgraph densities, but the reverse it not true. Two non-isomorphic graphs may also have the same subgraph densities.

$$\ln Z(\theta) \xrightarrow[\text{ⓐ}]{\geq} \max_u \gamma(\theta, u) \xrightarrow[\text{ⓑ}]{\approx} \max_u \tilde{\gamma}(\theta, u)$$

Figure 2: The algorithm has two approximations: ⓐ $\gamma(\theta, u)$ is an edge-count-u-induced lower bound for $\ln Z(\theta)$, Lemma 3 shows the error is bounded in $O(\ln n)$; ⓑ We propose $\tilde{\gamma}(\theta, u)$ as an approximation to the unknown $\gamma(\theta, u)$, following Lemma 4 and Lemma 5.

Given some set S, and any function $f : S \to \mathcal{R}$, a well known computation trick of computing $\ln \sum_{x \in S} \exp f(x)$ is to use $\max_{x \in S} f(x)$ as an approximation if $|S|$ is small. Specifically, we have the following bounds:

Lemma 2. *Let f be a function on S and $x^* = \operatorname{argmax}_{x \in S} f(x)$, it holds that:*

$$f(x^*) \leq \ln \sum_{x \in S} \exp f(x) \leq f(x^*) + \ln |S|$$

Direct application of Lemma 2 to $\ln Z(\theta)$ yields a sloppy approximation because the huge size of $\mathcal{G}$. Thanks to Lemma 1, the following approximation to (4) has a much tighter error bound:

$$\ln Z(\theta) = \max_{\mathbf{h} \in \mathcal{H}} \{\theta^T \mathbf{h} + \ln \#(\mathbf{h})\} + O(\ln n) \tag{5}$$

Now, we discuss how to estimate the first term of (5).

3.2 Edge-Count Induced Lower Bounds

In this section, we first derive an alternative representation to the approximation in (5), then develop an estimator for the approximation.

Let $\mathcal{G}_u \subset \mathcal{G}$ be the set of graphs with u edges, $\mathcal{H}_u \subset \mathcal{H}$ be the set of subgraph statistics induced by $\mathcal{G}_u$, and $\#_u(\mathbf{h})$ be the restricted counting function which only counts graphs in $\mathcal{G}_u$, i,e. $\#_u(\mathbf{h}) = |\{g \in \mathcal{G}_u | \phi(g) = \mathbf{h}\}|$. For any θ and u, we have the following lower bound to (5):

$$\gamma(\theta, u) = \max_{\mathbf{h} \in \mathcal{H}_u} \{\theta^T \mathbf{h} + \ln \#_u(\mathbf{h})\} \leq \max_{\mathbf{h} \in \mathcal{H}} \{\theta^T \mathbf{h} + \ln \#(\mathbf{h})\}$$

Notice that the equality holds when K_2 is a feature subgraph and $u = \operatorname{argmax}_u \gamma(\theta, u)$, because in this case $\mathcal{H}_u \cap \mathcal{H}_{u'} = \emptyset$ if $u' \neq u$, therefore $\{\mathcal{H}_u\}$ is a partition of $\mathcal{H}$. Specifically, we have the following lemma:

Lemma 3. *Given K_2 is included in subgraph features, the following equation holds:*

$$\begin{aligned}\max_u \{\gamma(\theta, u)\} &= \max_u \left\{ \max_{\mathbf{h} \in \mathcal{H}_u} \{\theta^T \mathbf{h} + \ln \#_u(\mathbf{h})\} \right\} \\ &= \max_{\mathbf{h} \in \mathcal{H}} \{\theta^T \mathbf{h} + \ln \#(\mathbf{h})\} = \ln Z(\theta) - O(\ln n)\end{aligned}$$

Lemma 3 shows $\max_u \{\gamma(\theta, u)\}$ can be treated as a good approximation of $\ln Z(\theta)$ with bounded error. However, the computation of $\gamma(\theta, u)$ is still non-trivial. For the rest of this section, we develop an approximation of $\gamma(\theta, u)$ by exploiting the asymptotic property of $\#_u(\mathbf{h})$ in $\mathcal{G}_u$.

Define $\mathbf{h}'$ and $\mathbf{h}^*$ as the optimum of $\gamma(\theta, u)$ and maximizer of $\#_u(\mathbf{h})$ respectively:

$$\mathbf{h}' \equiv \operatorname*{argmax}_{\mathbf{h} \in \mathcal{H}_u} \{\theta^T \mathbf{h} + \ln \#_u(\mathbf{h})\}$$

$$\text{and} \quad \mathbf{h}^* \equiv \operatorname*{argmax}_{\mathbf{h} \in \mathcal{H}_u} \{\ln \#_u(\mathbf{h})\}$$

The following bounds of $\gamma(\theta, u)$ hold for all θ and u:

$$\begin{aligned}\theta^T \mathbf{h}^* + \ln \#_u(\mathbf{h}^*) \leq \gamma(\theta, u) &= \theta^T \mathbf{h}' + \ln \#_u(\mathbf{h}') \\ &\leq \theta^T \mathbf{h}' + \ln \#_u(\mathbf{h}^*)\end{aligned} \tag{6}$$

Notice that the gap between the upper and lower bounds in (6) is $\theta^T(\mathbf{h}' - \mathbf{h}^*)$. When $\theta = 0$, we have $\mathbf{h}' = \mathbf{h}^*$; As θ deviates from 0, $\mathbf{h}'$ will also stray from $\mathbf{h}^*$ and settle the conflict between the increasing linear term $\theta^T \mathbf{h}$ and a very quickly diminishing $\ln \#_u(\mathbf{h})$ as we will soon discuss in the next section. We argue that $\mathbf{h}^*$ can be used to approximate $\mathbf{h}'$ in terms of estimating the log partition function $\ln Z(\theta)$ if the model of interest is not dominated by a few graphs[2]. Compared to $\mathbf{h}'$ and $\ln \#_u(\mathbf{h}')$, $\mathbf{h}^*$ and $\ln \#_u(\mathbf{h}^*)$ are much easier to estimate, therefore lead to a feasible approximation to $\gamma(\theta, u)$.

Estimating $\mathbf{h}^*$:

$\mathbf{h}^*$ maximizes the counting function $\#_u(\mathbf{h})$ on $\mathcal{G}_u$, therefore for any randomly picked graph $g \in \mathcal{G}_u$, $\mathbf{h}^*$ is the most likely value of $\phi(g)$. If we define a uniform distribution on $\mathcal{G}_u$, then $\mathbf{h}^*$ is the mode of $\phi(\mathcal{G}_u)$.

The process of generating graphs randomly from $\mathcal{G}_u$ is known as as Erdös-Rényi (ER) random graphs model $G(n, M)$ [Erdös and Rényi, 1960]. Here n is the number of nodes in the graph and $M = u$ is the number of edges. An alternative (and popular) definition of ER model is $G(n, p)$ [Gilbert, 1959], in which an order-n graph is constructed by picking each edge independently with probability p. The distribution of subgraph statistics for ER models has been actively studied in probabilistic graph theory[Nowicki, 1989; Döring and Eichelsbacher, 2009].

Nowicki [Nowicki, 1989] proved that $\phi(g)$ is asymptotically normally distributed for $g \in G(n, p)$. The following lemma extends that result to $G(n, M)$ over $\mathcal{G}_u$ using Chebyshev's inequality:

Lemma 4. *Let s_i be the edge count of L_i, define function $\rho_i(u) = (u/\binom{n}{2})^{s_i}$. Given any edge density μ, write the edge count $u = \binom{n}{2}\mu$ as a function of n. Then for any real vector $\mathbf{a} = (a_1, a_2, \ldots, a_r)^T$ and random graph $g \in G(n, M = u)$, the following holds as $n \to \infty$:*

$$P\left(\left|\mathbf{a}^T(\phi(g) - \rho(u))\right| \geq \frac{1}{cn}\right) \to 0 \tag{7}$$

where $\rho(u) = (\rho_1(u), \ldots, \rho_r(u))^T$ and c is some constant.

Notice here $\rho_i(u)$ is the expected density of L_i in $G(n, p = u/\binom{n}{2})$. Lemma 4 suggests that the subgraph densities for

[2] In the cases where a few graphs are dominating the model, $\mathbf{h}'$ will sway away from $\mathbf{h}^*$. Large entries in θ will lead to this scenario as we will show case in Section 4.3.

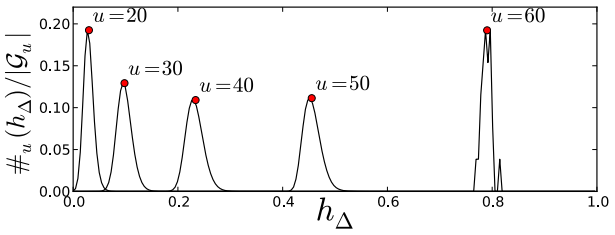


Figure 3: Concentration of triangle density h_Δ conditioned on the number of edges $u \in \{20, 30, 40, 50, 60\}$ for unlabeled graphs ($n = 12$). In this case, there are $\binom{12}{2} = 67$ possible edge counts. Y-axis measures the counting function $\#_u(h_\Delta)$ normalized by $|\mathcal{G}_u|$.

most graphs of $\mathcal{G}_u$ are close to $\rho(u)$! In a sense, graphs in $\mathcal{G}_u$ form a cluster in terms of the subgraph statistics. Figure 3 illustrates the phenomenon using order 12 unlabeled graphs [Brouwer, accessed 2012 Sep 30]. This result suggests that $\rho(u)$ is a good estimation of $\mathbf{h}^*$ for large n.

Estimating $\ln \#_u(\mathbf{h}^*)$:
Eq (7) also hints using $|\mathcal{G}_u|$ to approximate $\#_u(\mathbf{h}^*)$ as $\phi(g)$ concentrates. With the help of Lemma 1, it turns out to be a very good estimation:

Lemma 5. *Given an edge count u, it holds that $\ln \#_u(\mathbf{h}^*) = \binom{n}{2} H(u/\binom{n}{2}) - O(\ln n)$ where $H(x) = -x \ln x - (1-x)\ln(1-x)$.*

Estimating Lower Bound $\gamma(\theta, u)$:
Given estimations of $\mathbf{h}^* \simeq \rho(u)$ and $\ln \#_u(\mathbf{h}^*) \simeq \binom{n}{2} H\left(u/\binom{n}{2}\right)$, we immediately have the following approximation to $\gamma(\theta, u)$:

$$\tilde{\gamma}(\theta, u) = \theta^T \rho(u) + C(n, u) = \theta^T \rho(u) + \binom{n}{2} H(u/\binom{n}{2}) \quad (8)$$

Lemma 4 hints that as $\mathbf{h}'$ deviates away from $\mathbf{h}^*$, $\ln \#_u(\mathbf{h}')$ will diminish rapidly as $\theta^T \mathbf{h}'$ increases linearly. When the gradient of the linear term is small, $\mathbf{h}^*$ tends to be a good approximation of $\mathbf{h}'$. If the gradient is steep, $\mathbf{h}'$ will lean towards the extreme entry in $\mathcal{H}_u$ that maximizes the linear term but leads to a minimal $\#_u(\mathbf{h}')$, i.e. only one graph (or a few graphs) in $\mathcal{G}_u$ has feature vector $\mathbf{h}'$, but it dominates all other graphs.

Before the discussion of the approximate algorithm and its behavior as $n \to \infty$, we first investigate the behavior of fixed θ for networks of different order n.

3.3 Necessary Condition for Asymptotically Relevant Parameters

The analysis in Section 3.2 also reveals that ERGM with fixed θ may have very different behavior for networks of different order n. To see this, let $u^* = \operatorname{argmax}_u \gamma(\theta, u)$, we show that as $n \to \infty$, any fixed θ becomes asymptotically irrelevant in $\max_u \gamma(\theta, u)$, because $u^*/\binom{n}{2}$ will converge towards $1/2$ and $\mathbf{h}'(\theta, u^*)$ will converge towards $\rho(u^*)$:

Lemma 6. *Let $u^* = \operatorname{argmax}_u \gamma(\theta, u)$ with θ fixed, $\mathbf{h}'(\theta, u^*)$ converges to $\rho(u^*)$ asymptotically as $n \to \infty$.*

Lemma 6 implies that the effects of any fixed θ will diminish to a one-dimensional function $\rho(u^*)$ as n increases. The shifting of model behavior for different n is closely related to the discussion of instability of ERGM sufficient statistics [Schweinberger, 2011], and more recently the result of ERGM's inconsistency under sampling [Shalizi and Rinaldo, 2013].

Lemma 6 further suggests that for large n, θ needs to be in $\Theta(n^2)$ to be relevant in the model. This property leads to a functional representation of θ:

$$\theta(n, \xi) = \binom{n}{2} \xi \quad (9)$$

Here ξ is a "scale-free" meta parameter. This representation will give us a reasonable estimation on the scale of θ, which is necessary perform point-wise MLE for our approximation explained in the next section.

3.4 Approximate Algorithm

The estimation of edge-count induced lower bound immediately leads to an approximation of $\ln Z(\theta)$: **Edge Count Search (ECS) approximation:**

$$\mathrm{ECS}(\xi, n) = \binom{n}{2} \max_{0 \le u \le \binom{n}{2}} \left\{ \xi^T \rho(u) + H(u/\binom{n}{2}) \right\} \quad (10)$$

Here ξ is the "scale-free" parameter defined in Eq (9). Algorithm 1 reports a straightforward implementation of (10), which simply searches through all the u to maximize $\tilde{\gamma}(\theta, u)$. Notice that the algorithm requires no extra parameters, which makes the ECS approximation very easy to apply compared to current MCMC sampling methods.

Assume the number of subgraph features $r \ll n$, the time complexity of Algorithm 1 is in $O(n^2)$, which is linear in terms of the number of random variables (i.e. edges) of the model, and quadratic in terms of the order of the network.

A straightforward approximation to the log-likelihood function $\ell(g|\theta)$ is to replace $\ln Z(\theta)$ with $ECS(\xi, n)$:

$$\ell_{\mathrm{ECS}}(\theta \mid g) = \theta^T \phi(g) - ECS(\xi, n)$$

The decision of approximating $\mathbf{h}'$ with $\mathbf{h}^*$ in Section 3.2 leads to a simple algorithm. As $n \to \infty$, ECS approximation (10) converges to another closely related approximation proposed by Chatterjee and Diaconis [Chatterjee and Diaconis, 2011], who show that for certain θ the approximation will converge to the true log partition functions. In this paper, we also show that $\ln Z(\theta)$ will converge to $ECS(\theta, n)$ as $n \to \infty$ in Lemma 6.

In next section, we will resort to experiments to verify the effectiveness of the proposed algorithm.

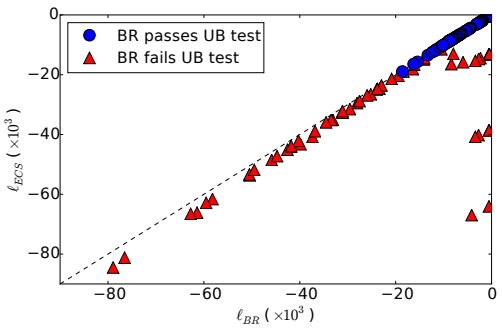


Figure 4: Scatter plot of log-likelihood estimations for ECS and BR on networks of $n = 160$. Many BR estimations fail UB test (11). Otherwise, ECS and BR estimations are very close (top right).

Algorithm 1 Our new ECS Approximation to the log partition function $\ln Z(\theta)$

Input: model parameter θ and number of nodes n
Output: estimation of $\ln Z(\theta)$
Initialize $ECS \leftarrow -\infty$
for $u \leftarrow 0$ **to** $n(n-1)/2$ **do**
  $ECS \leftarrow \max\{\tilde{\gamma}(\theta, u), ECS\}$
end for

4 Experimental Results

In this section, we first use two tasks on synthetic data set to evaluate the performance of ECS approximation: estimating log-likelihood functions and MLE estimation. Then we perform experiments on a real-world social network dataset to show case the quality and stability of our algorithm. We implement the commonly used triad model (edge, 2-star, triangle) for the experiments. To mitigate the degeneration problem in case study, we also used model (edge, altkstar($\lambda = 1.5$)). Here, altkstar is the set of all possible k-stars subgraphs for a network with their weights being set by a function of λ and k. For a fixed n, this feature is mathematically equivalent to including all k-star subgraphs into ERGM, except some constraints on their parameters.

We compare the output of ECS with the state of the art MCMC sampling algorithm for ERGMs: Bridge Sampling [Gelman and Meng, 1998; Handcock *et al.*, 2003; Hunter *et al.*, 2012].[3] The details of our experimental settings are not described here due to lack of space.

To alleviate the interference of the well known stability problem from sampling based methods on ERGMs [Handcock, 2003; Bhamidi *et al.*, 2008], we employ the following upper bound to the log likelihood function as an indicator of bad approximation:

$$\ell(g|\theta) \leq \theta^T \phi(g) - \max\{0, \sum_{i=1}^{r} \theta_i\} \qquad (11)$$

[3]For large $n > 20$, finding the exact ERGM log partition function is intractable because the number of all graphs with n nodes are $O(2^{n^2})$.

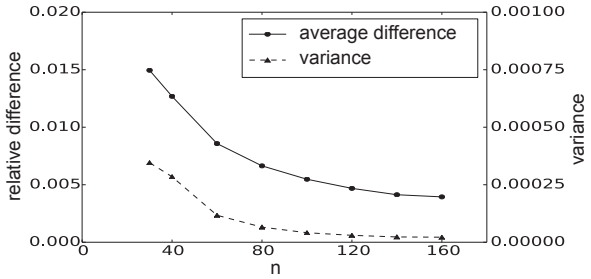


Figure 5: Relative difference between the estimations of ECS and BR for different n, given BR estimation passes the UB test (11). x-axis is the order of the network, y-axis measures the mean and variance of relative differences.

The bound holds for any θ and g, because $\ln Z(\theta)$ must be larger than the log potential of empty graph, which is 0, and of complete graph, which is $\sum_{i=1}^{r} \theta_i$. Notice that by design, ECS will never violate the bound. Because for any θ, we have $\gamma(\theta, 0) = \tilde{\gamma}(\theta, 0)$ and $\gamma(\theta, \binom{n}{2}) = \tilde{\gamma}(\theta, \binom{n}{2})$. We apply this upper bound test (UB test) for all log-likelihood estimations.

4.1 Estimating log-likelihood functions

We sample synthetic networks from a wide range of parameters to evaluate their log-likelihoods. We first generate a $6\times6\times6$ grid of ξ in $[-5, 5]\times[-5, 5]\times[-5, 5]$, and drop the tuples in which all values have the same sign. We ended up with 162 different ξs. Then for each ξ, we sampled networks for different $n \in \{30, 40, 60, 80, 100, 120, 140, 160\}$. The total number of sampled graphs is 1,296. We estimate the log-likelihood for each sampled network using both Bridge Sampling and ECS.

Figure 4 reports the scatter plot of the results of both methods for $n = 160$. Points close to the dashed line suggest ECS and BR produce similar results; points far away from the dashed line suggests the estimation results are very different. For each estimation of BR, we also check whether it exceeds the UB test. If the estimation exceeds the log-likelihood upper bound, we mark the data point with a cross ($\times$); otherwise we mark with a blue circle.

From 4 we can tell when BR estimation fails the UB test, the difference between ECS and BR results are almost negligible. However, there is a significant portion (about 30%) of BR estimation results turn out to be unrealistic, while ECS keeps producing results consistent to (11).

To further compare ECS estimations with the legit BR estimations, we report their relative differences for models that BR estimation pass the upper bound test: reldiff=$|(\ell_{ECS} - \ell_{Bridge})/\ell_{Bridge}|$. Figure 5 reports the mean and variance of the relative difference for networks of which BR estimation passes the UB test. The plot shows both the mean and variance decrease as n increases. As n increases, the estimations become very close.

4.2 MLE estimation

In this section, we use ECS as a sub-routine to perform MLE estimations on network data. Because the number of subgraph features in triad model is $r = 3$, it is practical to per-

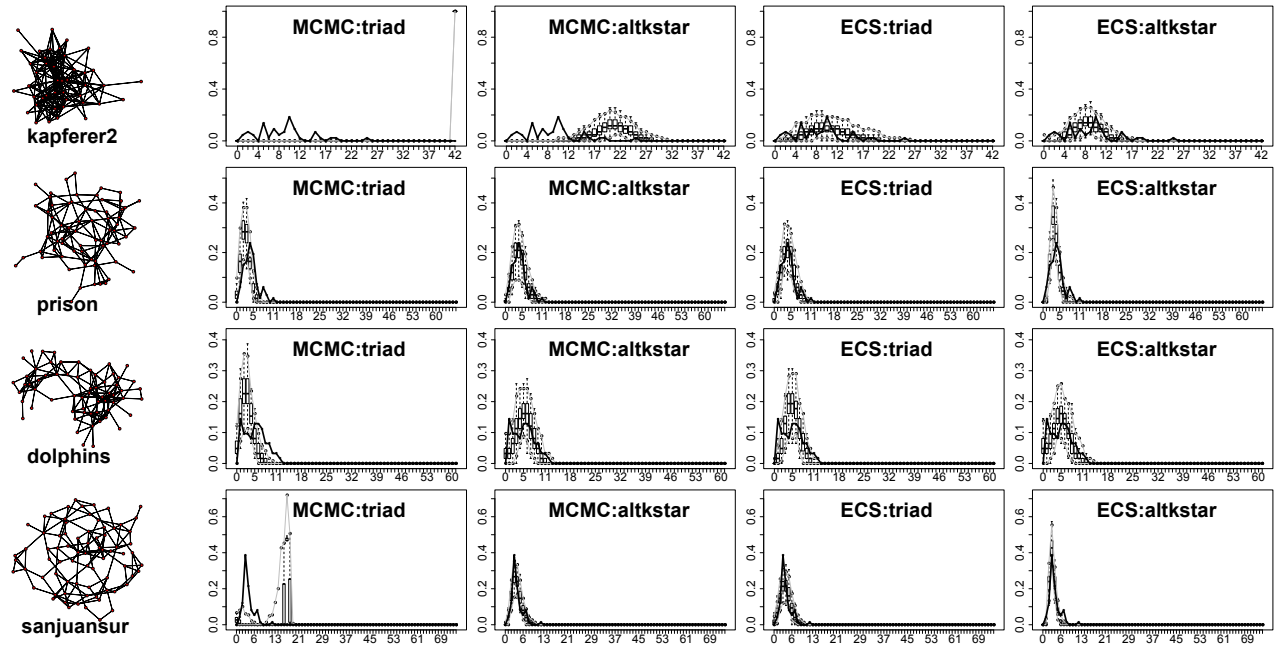


Figure 6: Experiments on four real network data set. The x axis of each figure is the node degree and the y axis is the distribution of the node degree. From the first row, the results of four networks (kapferer2, prison, dolphins and sanjuansur in order) are presented.

form grid search over a restricted sub-space of ξ. Using real social networks, we compare the performance of this simple ECS-MLE with MCMC-MLE [Snijders, 2002], which uses Bridge Sampling as a sub-routine [Handcock *et al.*, 2003].

We first generated a $6\times6\times6$ grid of ξ in in $[-3,3]\times[-3,3]\times[-3,3]$. For each ξ, we sampled one network of $n = 60$. The purpose of this procedure is not to generate samples that represents the underlying ξ well, but to diversify the sampled graphs. Then we fit the triad model with the sampled network using both MCMC-MLE and ECS-MLE. For ECS-MLE, we performed grid search in a slightly enlarged parameter space with finer granularity.

To evaluate, we estimated the log-likelihood of the network on both fitted models using Bridge Sampling. As we observed in Figure 4, BR tends to generate unrealistic estimations. If a BR estimation fails the UB test, we take the upper bound as the log-likelihood estimation instead. Notice that the adjusted value is still valid for the purpose of comparing two MLE algorithms.

4.3 Case Study on Real Network Data

To study the stability and quality of ECS-MLE results, we fit two different models with four networks with more than 40 nodes: one *kapferer2* from statnet; and the other three networks, *prison*, *dolphins*, and *sanjuansur*, from CMU CASOS.

We simulate graphs from the model and compare the node degree distribution to the original network [Hunter *et al.*, 2008]. The quality of the fitted models is measured by the total variation distance[4].

	MCMC		ECS	
	triad	altkstar	triad	altkstar
kapferer2	0.88±0.13	0.62±0.29	**0.39±0.01**	0.43±0.01
prison	0.48±0.09	**0.14±0.01**	0.15±0.00	0.23±0.01
dolphins	0.65±0.28	**0.19±0.01**	0.24±0.01	**0.19±0.01**
sanjuansur	0.61±0.11	0.19±0.01	0.22±0.01	**0.16±0.00**

Table 1: Total variation distance between the degree distribution of real network and simulated networks. Average total variation distances and standard deviations are presented.

Both for triad models and alternating k-stars models, we set the initial value of grid search by using Maximum Pseudo-Likelihood Estimate (MPLE) as in the MCMC-MLE. For triad models, our ECS algorithm searched the grid of two parameters of (edge, 2-star, triangle) ranging the initial value plus offsets in $[-5,5]\times[-5,5]\times[-5,5]$, with granularity of 0.2. For alternating k-stars, the ECS searched the grid of three parameters of (edge, altkstar) ranging, initial value plus offsets in $[-15,-15]\times[-15,15]$, with the same granularity.

Figure 6 presents the node degree distributions of the simulated network models. The first column shows the network plot of the four networks. The following two columns show the node degree distributions of the triad and the altkstar models fitted by the MCMC-MLE. The final two columns show the node degree distributions fitted by our ECS algorithm. The bold line is the degree distribution for the original network. The curve with error bar is the degree distribution for simulated networks. A good match between the two line suggests the simulated networks implies the quality of the fitted

[4] $\|f-g\| = \sup_{A\in\mathcal{B}}(f(A)-g(A))$, $\mathcal{B}$ is a class of Borel sets.

model.

Table 1 shows the total variation distance of node degree distributions of the original and the fitted. We marked the minimum total variation among four approaches for each network. As you can see the ECS-MLE results show better performance compared with the MCMC-MLE results in overall.

5 Related Work

Modeling social network structures has been actively studied in machine learning community. Latent variable models, such as matrix factorization [Hoff, 2008], block modeling [Airoldi *et al.*, 2008; Kemp *et al.*, 2006; Ho *et al.*, 2012] and others [Miller *et al.*, 2009; Lloyd *et al.*, 2012], represent the relational data with latent variables. Among those, Ho et al. [Ho *et al.*, 2012] proposed triangular motifs as network representation, which is closely related to ERGM's subgraph features. In comparison, ERGM posts a simple model with intuitive feature specifications that fits for many network analysis tasks.

Computing normalizing constants for complex and high-dimensional models, such as ERGMs, is intractable. MCMC simulations are arguably among the most effective methods. Gelman and Meng [Gelman and Meng, 1998] proposed the path sampling formulation to unify acceptance ratio method and thermodynamic integration from theoretical physics for estimating the (ratios of) normalizing constants. Annealed importance sampling (AIS) [Neal, 2001], popular in deep learning literature [Salakhutdinov and Murray, 2008], can also be viewed as one form of thermodynamic integration. Although effective in many applications, Bhamidi et al. [Bhamidi *et al.*, 2008] shows the mixing time for any local Markov chain in low temperature regimes of ERGMs is exponentially slow, rendering these methods computationally intractable in many cases. In comparison, ECS approximation is deterministic, therefore avoids the sampling.

ECS approximation is a variational inference algorithm. In this category, there are many other techniques, such as pseudo-log-likelihood [Strauss and Ikeda, 1990], mean field approximation and Bethe approximation [Wainwright and Jordan, 2008]. In the context of ERGM, these methods have been reported to be inferior to sampling based methods [van Duijn *et al.*, 2009], and are usually used to generated initial states for sampling based algorithms [Hunter *et al.*, 2012]. ECS distinguishes from others by exploiting the asymptotic property in the feature space of the model. This macroscopic view goes beyond the conditional independence in local structures of the model, and may be more effective for complex high-dimensional models like ERGMs.

Many studies have empirically shown that the triad model tends to produce degenerated models when fitted by the MCMC-MLE [Hunter *et al.*, 2008; 2012]. We also observe that the MCMC-MLE fitted triad model was completely degenerated in *kapferer2* and *sanjuansur*. However, our ECS-MLE could find an accurate model with no degeneration. To mitigate this problem, the alternating k-stars [Robins *et al.*, 2007; Hunter *et al.*, 2012] have been introduced. In our experiments, the MCMC-MLE fitted altkstar outperforms the MCMC-MLE triad. Although some of the MCMC-MLE result avoided degeneration, ECS-MLE results either with triad and altkstar show the better match to the degree distribution in original network.

6 Conclusions

In this paper, we propose a novel deterministic approximation to the log partition functions of ERGMs. Computing the partition functions (or the ratio of them) is essential in learning ERGMs. Our results show the new method is able to overcome some of the stability issues faced by sampling based methods without losing accuracy. The new algorithm does not depends on extra parameters, making it easy to implement and apply compared to sampling.

We also show that the proposed approximation can be used to build an effective MLE algorithm for ERGMs. In the future, we plan to address various types of MLE problems in EMRGs by using the proposed approximation principles.

Acknowledgments

Wen Pu and Eyal Amir were supported by NSF EAR grant 09-43627, IIS grant 09-17123, IIS grant 09-68552, and a DARPAR grant as part of the Machine Reading Program under AFRL prime contract no. FA8750-09-C-0181. Jaesik Choi and Yunseong Hwang were supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science, ICT & Future Planning (MSIP) (NRF-2014R1A1A1002662), the NRF grant funded by the MSIP (NRF-2014M2A8A2074096). The authors would like to thank David Hunt, Michael Schweinberger, Dan Roth, Gerald DeJong and anonymous reviewers for their constructive suggestions and feedbacks.

A Appendix

A.1 Proof of Lemma 1

Proof. Subgraph count for L_i in any g is bounded by $0 \leq t(g, L_i) \leq t(K_n, L_i) \leq \binom{n}{v_i} v_i!$, therefore

$$\ln |\mathcal{H}| \leq \ln \prod_{i=1}^{r} t(K_n, L_i) \leq \ln \left[\prod_{i=1}^{r} \binom{n}{v_i} v_i! \right]$$
$$\leq r \ln \left[\binom{n}{v^*} v^*! \right] = r \ln \frac{n!}{(n-v^*)!} \leq r v^* \ln n$$

□

A.2 Proof of Lemma 4

Before the proof of Lemma 4, we need some preparations. [Nowicki, 1989] proved that a vector of subgraph counts in $G(n,p)$ are asymptotically normally distributed with a degenerated co-variance matrix with rank 1, as the order of the graph $n \to \infty$. In other words, the subgraph counts are asymptotically linearly dependent on each other. Formally, let $\phi(g') = \{\phi_1(g'), \phi_2(g'), \ldots, \phi_r(g')\}$ be the densities of subgraphs $L_1, L_2, \ldots, L_r$ (i.e. $\phi_i(g') = \frac{t(g', L_i)}{t(K_n, L_i)}$) for $g' \in G(n,p)$, the sizes (number of edges) of these subgraphs are $s_1, s_2, \ldots, s_r$, and $u \sim \text{Bin}(\binom{n}{2}, p)$ is the edge count of g', we have the following theorem:

Theorem 1. *[Nowicki, 1989] For $g' \in G(n,p)$, and real vector $\mathbf{a} = (a_1, a_2, \ldots, a_r)^T$, the following asymptotic property holds:*

$$n^2 E\left[\mathbf{a}^T\Big(\phi(g') - \rho(u,p)\Big)\right]^2 \to 0 \tag{12}$$

where $\rho(u,p) = (\rho_1(u,p), \ldots, \rho_r(u,p))$, and $\rho_i(u,p) = s_i p^{s_i-1} \cdot \frac{u}{\binom{n}{2}} - (s_i - 1)p^{s_i}$.

In theorem 1, if we set $p = u/\binom{n}{2}$, then $\rho_i(u, u/\binom{n}{2}) = \left(\frac{u}{\binom{n}{2}}\right)^{s_i}$, which becomes the expected density of L_i in $G(n, p = u/\binom{n}{2})$.

Next step is to extend the above property from $G(n,p)$ to $G(n,M)$.

Corollary 1. *For $g \in G(n, M = u)$, as $n \to \infty$, it holds that*

$$n^2 E_u\left[\mathbf{a}^T\Big(\phi(g) - \rho(u)\Big)\right]^2 \to 0$$

where $\rho_i(u) = \left(u/\binom{n}{2}\right)^{s_i}$

Proof. Following theorem 1, let $\rho_i(u) = \rho_i(u, u/\binom{n}{2})$, as $n \to \infty$, the following holds for $g' \in G(n, p = (u/\binom{n}{2}))$:

$$\begin{aligned}
&n^2 E\left[\mathbf{a}^T\Big(\phi(g') - \rho(u)\Big)\right]^2 \to 0 \\
\Rightarrow &n^2 E\left[E_u\big[\mathbf{a}^T\Big(\phi(g') - \rho(u)\Big)\ |u\big]^2\right] \to 0 \\
\Rightarrow &n^2 \sum_u p(u)\, E_u\big[\mathbf{a}^T\Big(\phi(g') - \rho(u)\Big)\ |u\big]^2 \to 0
\end{aligned}$$

Because $\sum_u p(u) = 1$ and $p(u) > 0$, the claim holds. □

Let c be some positive constant, apply Chebyshev's inequality to the linear combination $\mathbf{a}^T\phi(g)$, we get:

$$P\left(\left|\mathbf{a}^T\left(\phi(g) - E_u\left(\phi(g)\right)\right)\right| \geq \frac{1}{2cn}\right) \leq 4c^2 n^2 \mathrm{Var}(\mathbf{a}^T\phi(g)) \tag{13}$$

Now we start to prove Lemma 4.

Proof of Lemma 4. We first define function $\varepsilon(u)$:

$$\varepsilon(u) = \mathbf{a}^T(E_u(\phi(g)) - \rho(u)) \tag{14}$$

As we know

$$\begin{aligned}
E_u\left(\mathbf{a}^T(\phi(g) - \rho(u))\right)^2 &\geq E_u\left(\mathbf{a}^T\left(\phi(g) - E_u(\phi(g))\right)\right)^2 \\
&= \mathrm{Var}(\mathbf{a}^T\phi(g))
\end{aligned} \tag{15}$$

The equality holds if and only if $\varepsilon(u) = 0$. We can get the following property after applying it to corollary 1: as $n \to \infty$

$$n^2 \mathrm{Var}\left(\mathbf{a}^T\phi(g)\right) \to 0 \tag{16}$$

Therefore, as $n \to \infty$

$$\begin{aligned}
&n^2 E_u\left(\mathbf{a}^T\left(\phi(g) - \rho(u)\right) - \varepsilon(u)\right)^2 \to 0 \\
\Rightarrow &n^2 E_u\left(\mathbf{a}^T\left(\phi(g) - \rho(u)\right)\right)^2 - n^2\varepsilon(u)^2 \to 0 \\
\Rightarrow &|\varepsilon(u)| < \frac{1}{2n}
\end{aligned} \tag{17}$$

The last step used corollary 1.

We slacks (13) using (15), and rewrite the inner expectation term using (14):

$$\begin{aligned}
&P\left(\left|\mathbf{a}^T\left(\phi(g) - \rho(u)\right) - \varepsilon(u)\right| \geq \frac{1}{2cn}\right) \\
&\qquad \leq 4c^2 n^2 E_u\left(\mathbf{a}^T\left(\phi(g) - \rho(u)\right)\right)^2
\end{aligned} \tag{18}$$

Using (17), we can get

$$\begin{aligned}
&P\left(\left|\mathbf{a}^T\left(\phi(g) - \rho(u)\right) - \varepsilon(u)\right| \geq \frac{1}{2cn}\right) \\
&\qquad \geq P\left(\left|\mathbf{a}^T\left(\phi(g) - \rho(u)\right)\right| \geq \frac{1}{cn}\right)
\end{aligned}$$

Therefore, apply corollary 1, as $n \to \infty$, we get

$$P\left(\left|\mathbf{a}^T\left(\phi(g) - \rho(u)\right)\right| \geq \frac{1}{cn}\right) \to 0$$

□

A.3 Proof of Lemma 5

Proof. Let $\mathcal{G}_u$ be the set of graphs with edge count u, since $\mathbf{h}^*_u$ is the maximizer, we have

$$\#(\mathbf{h}^*_u) \geq \frac{|\mathcal{G}_u|}{|\mathcal{H}|}$$

Together with the trivial $\#(\mathbf{h}^*_u) \leq |G_u|$, we can get:

$$\ln|\mathcal{G}_u| - \ln|\mathcal{H}| \leq \ln\#(\mathbf{h}^*_u) \leq \ln|\mathcal{G}_u| \tag{19}$$

Apply Stirling's approximation on $\ln|\mathcal{G}_u|$:

$$\begin{aligned}
\ln|\mathcal{G}_u| &= \ln\binom{\binom{n}{2}}{u} \simeq \left(\binom{n}{2} - u\right)\ln\frac{\binom{n}{2}}{\binom{n}{2} - u} + u\ln\frac{\binom{n}{2}}{u} \\
&= \binom{n}{2} H(u/\binom{n}{2})
\end{aligned} \tag{20}$$

Therefore the claims hold. □

A.4 Proof of Lemma 6

Proof. By definition $\mathbf{h}'(\theta, u^*)$ is vector of densities ranging in $[0,1]$, therefore the product $|\theta^T\mathbf{h}'(\theta,u^*)|$ is bounded by constant $\sum_i |\theta_i|$.

Let $\alpha = \max_u H(u/\binom{n}{2}) \approx H(1/2)$, Lemma 5 shows that

$$\ln\#_{u^*}(\mathbf{h}^*(u^*)) = \binom{n}{2}\alpha - O(\ln n) \approx \emptyset(n^2)$$

Because Eq (6) implies

$$\begin{aligned}
\ln\#_{u^*}(\mathbf{h}^*(u^*)) - \sum_i |\theta_i| &\leq \ln\#_{u^*}(\mathbf{h}'(\theta,u^*)) \\
&\leq \ln\#_{u^*}(\mathbf{h}^*(u^*))
\end{aligned}$$

as $n \to \infty$ we have:

$$\ln\#_{u^*}(\mathbf{h}'(\theta,u^*)) \to \ln\#_{u^*}(\mathbf{h}^*(u^*))$$

Let $\mathbf{b}$ be some real vector, assume there is some $\epsilon > 0$ so that as $n \to \infty$ we have:

$$|\mathbf{b}^T(\mathbf{h}'(\theta,u) - \rho(u^*))| \geq \epsilon$$

In this case, Lemma 4 implies $\ln\#_{u^*}(\mathbf{h}'(\theta,u^*)) \to 0$. However, given that $\ln\#_{u^*}(\mathbf{h}^*(u^*))$ is in $\emptyset(n^2)$, it contradicts with the definition of $\mathbf{h}'(\theta,u)$. Therefore, $\mathbf{h}'(\theta,u) \to \rho(u^*)$. □

References

[Airoldi *et al.*, 2008] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research*, 9:1981–2014, 2008.

[Bhamidi *et al.*, 2008] S. Bhamidi, G. Bresler, and A. Sly. Mixing time of exponential random graphs. In *Proceedings of the 49th IEEE Annual Symposium on FOCS*, 2008.

[Brouwer, accessed 2012 Sep 30] Andries E. Brouwer. Number of unlabelled graphs with given number of triangles. http://www.win.tue.nl/~aeb/graphs/cospectral/triangles.html, accessed 2012 Sep. 30.

[Carrington *et al.*, 2005] Peter J Carrington, John Scott, and Stanley Wasserman. *Models and methods in social network analysis*. Cambridge University Press, 2005.

[Cetintas *et al.*, 2011] S. Cetintas, M. Rogati, L. Si, and Y. Fang. Identifying similar people in professional social networks with discriminative probabilistic models. In *Proceedings of the 34th ACM SIGIR Conference*, 2011.

[Chatterjee and Diaconis, 2011] S. Chatterjee and P. Diaconis. Estimating and understanding exponential random graph models. *Arxiv preprint arxiv:1102.2650*, 2011.

[Döring and Eichelsbacher, 2009] H. Döring and P. Eichelsbacher. Moderate deviations in random graphs and bernoulli random matrices. *Arxiv preprint arXiv:0901.3246*, 2009.

[Erdös and Rényi, 1960] P. Erdös and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl*, 5:17–61, 1960.

[Gelman and Meng, 1998] A. Gelman and X.L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, pages 163–185, 1998.

[Gilbert, 1959] E.N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, pages 1141–1144, 1959.

[Handcock *et al.*, 2003] Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris. *statnet: Software tools for the Statistical Modeling of Network Data*. Seattle, WA, 2003. Version 2.0.

[Handcock, 2003] M.S. Handcock. Assessing degeneracy in statistical models of social networks, 2003.

[Ho *et al.*, 2012] Qirong Ho, Junming Yin, and Eric Xing. On triangular versus edge representations – towards scalable modeling of networks. In *Advances in Neural Information Processing Systems 25 (NIPS)*, 2012.

[Hoff, 2008] P.D. Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems 21 (NIPS)*, 2008.

[Hunter *et al.*, 2008] D.R. Hunter, S.M. Goodreau, and M.S. Handcock. Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481):248–258, 2008.

[Hunter *et al.*, 2012] David R. Hunter, Pavel N. Krivitsky, and Michael Schweinberger. Computational statistical methods for social network models. *Journal of Computational and Graphical Statistics*, 21(4):856–882, 2012.

[Hunter, 2007] David R Hunter. Curved exponential family models for social networks. *Social networks*, 29(2):216–230, 2007.

[Kemp *et al.*, 2006] Charles Kemp, Joshua B Tenenbaum, Thomas L Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st AAAI Conference*, 2006.

[Kempe *et al.*, 2003] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD Conference*, 2003.

[Lloyd *et al.*, 2012] James Robert Lloyd, Peter Orbanz, Zoubin Ghahramani, and Daniel M. Roy. Random function priors for exchangeable arrays with applications to graphs and relational data. In *Advances in Neural Information Processing Systems 25 (NIPS)*, 2012.

[Mancilla-Caceres *et al.*, 2012] Juan F Mancilla-Caceres, Wen Pu, Eyal Amir, and Dorothy Espelage. Identifying bullies with a computer game. In *Proceedings of the 26th AAAI Conference*, 2012.

[Miller *et al.*, 2009] Kurt Miller, Thomas Griffiths, and Michael Jordan. Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems 22 (NIPS)*, 2009.

[Neal, 2001] Radford M Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

[Nowicki, 1989] K. Nowicki. Asymptotic normality of graph statistics. *Journal of Statistical Planning and Inference*, 21(2):209–222, 1989.

[Robins *et al.*, 2007] G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173–191, 2007.

[Salakhutdinov and Murray, 2008] Ruslan Salakhutdinov and Iain Murray. On the quantitative analysis of deep belief networks. In *Proceedings of the 25th ICML*, 2008.

[Schweinberger, 2011] Michael Schweinberger. Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, 106(496):1361–1370, 2011.

[Shalizi and Rinaldo, 2013] Cosma Rohilla Shalizi and Alessandro Rinaldo. Consistency under sampling of exponential random graph models. *The Annals of Statistics*, 41(2):508–535, 2013.

[Snijders, 2002] T.A.B. Snijders. Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40, 2002.

[Strauss and Ikeda, 1990] D. Strauss and M. Ikeda. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, pages 204–212, 1990.

[van Duijn *et al.*, 2009] M.A.J. van Duijn, K.J. Gile, and M.S. Handcock. A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks*, 31(1):52–62, 2009.

[Wainwright and Jordan, 2008] M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.