

Estimation for Dyadic-Dependent Exponential Random Graph Models

Xiaolin Yang*, Alessandro Rinaldo, Stephen E. Fienberg

Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Abstract. Graphs are the primary mathematical representation for networks, with nodes or vertices corresponding to units (e.g., individuals) and edges corresponding to relationships. Exponential Random Graph Models (ERGMs) are widely used for describing network data because of their simple structure as an exponential function of a sum of parameters multiplied by their corresponding sufficient statistics. As with other exponential family settings the key computational difficulty is determining the normalizing constant for the likelihood function, a quantity that depends only on the data. In ERGMs for network data, the normalizing constant in the model often makes the parameter estimation intractable for large graphs, when the model involves dependence among dyads in the graph. One way to deal with this problem is to approximate the likelihood function by something tractable, e.g., by using the method of pseudo-likelihood estimation suggested in the early literature. In this paper, we describe the family of ERGMs and explain the increasing complexity that arises from imposing different edge dependence and homogeneous parameter assumptions. We then compare maximum likelihood (ML) and maximum pseudo-likelihood (MPL) estimation schemes with respect to existence and related degeneracy properties for ERGMs involving dependencies among dyads.

2000 Mathematics Subject Classifications: 05C82, 62F10

Key Words and Phrases: Dyadic dependence, ERGMs, Markov random graph models, Maximum likelihood estimation, Pseudo-likelihood estimation, Social networks

1. Introduction

Graphs provide a natural way to represent relational or network data in which nodes represent individuals and edges represent relationship among individuals. Network data are of special interest in many different scientific fields such as social science, biology and epidemiology. For an overview of such models and their analyses, see [18].

Among the common descriptive statistics used to describe network data are: counts of motifs such as edges, stars, triangles, density, centrality and cohesive subsets, etc. A

*Corresponding author.

Email addresses: xyang@stat.cmu.edu (X. Yang), arinaldo@stat.cmu.edu (A. Rinaldo), fienberg@stat.cmu.edu (S. Fienberg)

model incorporating these statistics may give descriptive “explanations” for those structural effects. In what follows we study several members of the class of network models whose probability distributions are an exponential family, describe their interconnections, and focus on the estimation of their parameters.

1.1. Types of Network Models

For simple network settings, in which the data form an adjacency matrix for the graph, [15] focus on two basic classes of network models: Exponential Random Graph Models (ERGMs), and Bayesian hierarchical models (see also [10]). ERGMs exhibit familiar exponential family form and they are often amenable to approaches linked to algebraic statistics because the likelihood functions involve polynomials, e.g, see [21, 11]. In addition the minimal sufficient statistics (MSS), which offer a lower dimensional representation of the data, possess in many cases interesting geometric properties that can be exploited for formal inference. Typical examples include Erdős-Rényi models, dyadic independence models such as the β -model [7, 24] and p_1 model [17], and Markov random graph models [13] more generally. Bayesian hierarchical models can involve ERGMs as partial building blocks but then they lose their simple exponential family form through the hierarchical assumptions, and almost always have no simple minimal sufficient statistics. The number of parameters is reduced by integrating over all parameters in lower levels of the hierarchy. In this paper, we mainly focus on the first type of model and study the properties and estimation methods of those models.

Here we focus on the following special subclasses of ERGMs:

- Undirected Bernoulli graphs with mutually independent and identically distributed edges (Erdős-Rényi model).
- Undirected Bernoulli graphs with node-dependent edges (e.g., β -model).
- Directed dyadic independence graphs with node-dependent parameters (e.g., p_1 model [17]).
- Undirected Markov graphs where we assume two edges are conditionally independent if they don’t share a node. [13] used ERGMs with number of edges, triangles and different degree of stars to model Markov graphs.
- Realization independent models: two tie variables share two relationships and form a circuit (e.g., see [25] and [31]).

Exponential Random Graph Models (ERGMs) have also been described as “ p^* ” models, e.g., by [36], [26]. All of these models, like those in the special classes mentioned above, happen to characterize the network through descriptive statistics, e.g., the number of edges, stars and triangles, etc. Each statistic represents a social relation pattern or motif that can occur in the graph. The most common form of inference for such models involves Maximum Likelihood Estimation (MLE) and Maximum Pseudo-Likelihood Estimation (MPLE).

1.2. Issues Associated With Parameter Estimation in ERGMs

While maximum likelihood estimation for the Erdős-Rényi, β and p_1 models is relatively straightforward, once one moves beyond the settings involving dyadic independence, problems arise due to the partition function that is needed to enumerate all graphs with the same number of nodes and the inferential degeneracy property of such models, e.g., see [16] and [22]. [32] systematically studied a more general class of statistical models with interacting points and talked about the degenerate behavior that can occur.

The term **degeneracy** is typically used to describe an array of seemingly pathological behaviors of some ERGMs. In this paper, we mean that as the number of nodes n increases the normalizing constant of the model becomes infinite so that there do not exist valid models that can be fit to the data. Typical phenomena include network data settings where (1) the probability distribution corresponding to the estimated parameters has mass only over few network configurations, for instance the empty or fully connected graph, or (2) the maximum likelihood estimation (MLE) does not exist and the Fisher Information matrix is singular or nearly so. [22] characterized the existence of MLE from the perspective of the geometry structure of discrete exponential families, and illustrate near-degeneracy behavior for some small graphs. [28] introduced the notion of “sensitivity” and used it to characterize the property of degeneracy. He attempted to explain the phenomenon from the perspective of fitting the model using MCMC and showed what kind of sufficient statistics make the model unstable.

The normalizing constant for dyadic-dependent ERGMs is often computationally intractable because its calculation requires the enumeration of all graphs with the same number of nodes. One way to estimate model parameters is to convert the joint likelihood function into a product of conditional likelihood to eliminate the normalizing constant. We refer to this as the method of Maximum Pseudo-Likelihood Estimation (MPLE) and [33] proposed its use for ERGMs and following the work of [35, 20] the method found widespread adoption in the social science literature because MPLE can be accomplished by fitting a logistic regression model.

The idea of MPLE goes back to the work by [4] who applied it in modeling the spatially interacting random variables using lattice systems, where each lattice point is conditionally independent of the other points given its nearest neighbors. Thus that the joint probability distribution can be factorized into a product form of conditional probabilities. [5] proposed pseudo-likelihood estimation for Gaussian random fields. These ideas then worked their way into a wide range of applications such as image processing and computer vision, e.g., [19] and [14]. The problem setting here is the estimation of the conditional relationships between random variables given the replicated observations of each random variable. The consistency of MPLE under some regularity conditions were proved by [2]. [9] gave the consistent confidence intervals for MPLE when independent and identically distributed observations are available. [8] also considered pseudo-likelihood estimation constructed from marginal densities and investigated the consistency property in this scenario. In ERGMs, the random variables are defined in terms of edges linking nodes and, when dyads are no longer independent, the existing asymptotic theory on pseudo-likelihood

estimation does not apply.

The properties of MPLE for ERGMs were not well understood and questions about them arose after [6] and [30] explained how maximum likelihood estimates could be produced directly via Markov chain Monte Carlo (MCMC) methods. When MCMC methods came into use, questions remained about the usefulness of MPLE methods, which seemed to work even in near-degenerate situations. [34] proposed a framework to compare the MLE and MPLE of ERGMs and investigated through simulations bias and efficiency in terms of mean-squared error of the natural and mean valued parameters.

1.3. Our Contributions

In this paper, we reconsider the comparison of MLE and MPLE methods for dyadic-dependent ERGMs. We give some theoretical properties, describe how they differ for small graphs, for which we can do full enumeration of possibilities, and discuss the situation for large graphs when the MLE can not be computed directly. We examine the asymptotic performance of the MLE and MPLE as the number of nodes increases, and report an empirical study regarding the consistency of MLE and MPLE.

The paper is organized as follows: we introduce the basic form of ERGMs and the relationships among different subclasses of ERGMs in section 2. We show the theoretical and empirical results about the comparison in section 3 and provide an overview of ERGM estimation properties in section 4.

2. Exponential Random Graph Models (ERGMs)

In this paper, we mainly focus on static or cross-sectional network settings and we consider different assumptions about the edge structures within exponential family framework. Thus we represent a network as a graph $G = (V, E)$, where the vertex set V corresponds to the individuals comprising the network and the edge set E corresponds to the relations linking the individuals. We represent the observed network by an $n \times n$ adjacency matrix, y , with entries that are 0's or 1's, where a 1 represents the presence of an edge between a pair of nodes and 0's represent absence. Note that the number of binary undirected

graphs with n nodes is $2^{\binom{n}{2}} = 2^{\frac{n(n-1)}{2}}$ because there are $\binom{n}{2}$ possible edges and each edge takes value 0 or 1. The observed adjacency matrix y is a realization of a random variable Y , whose distribution is assumed to have the exponential family form:

$$P_{\theta, \mathcal{Y}}(Y = y) = \frac{\exp[\theta \cdot T(y)]}{c(\theta, \mathcal{Y})}, \quad y \in \mathcal{Y}, \quad (1)$$

where

$$c(\theta, \mathcal{Y}) = \sum_{y \in \mathcal{Y}} \exp(\theta^T T(y)),$$

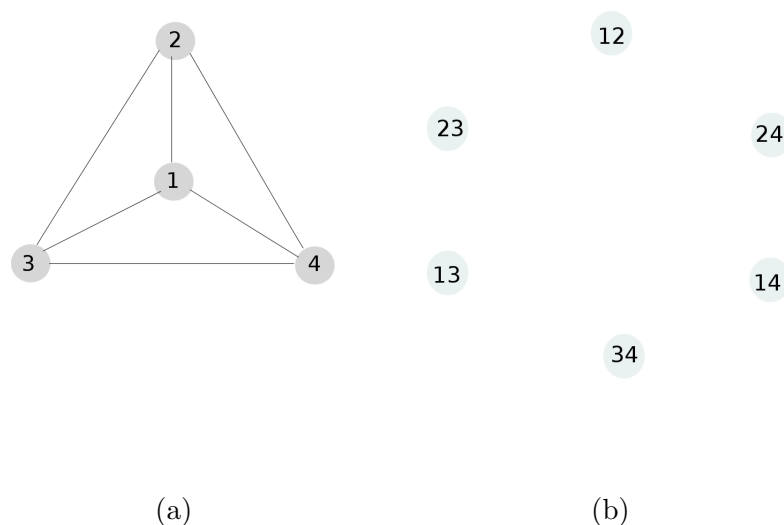


Figure 1: (a) 4-node complete graph (b) dependence graph with the assumption of complete independence among edges

$\theta \in \Theta \subseteq R^q$ is the parameter, and $T : \mathcal{Y} \rightarrow R^q$ are statistics based on functions of Y . These statistics may include counting quantities such as the number of edges, the number of reciprocated or mutual edges, the number of triangles, or the number of k -stars for $k = 2, 3, \dots$, etc. Other statistics or network motifs that might also be of interest are functions of Y that represent transitivity, centrality and clustering. The term in the denominator $c(\theta, \mathcal{Y})$ is the normalizing constant or *partition function* and computing it requires enumerating all possible graphs with n nodes. This is a barrier to do both simulation and inference for many specific ERGMs when the number of vertices is large.

We first describe some special classes of ERGMs where the edges or pairs of edges between pairs of nodes or dyads are independent, and then we focus on the dependent setting.

2.1. Erdős-Rényi Model

The Erdős-Rényi Model assumes edges are independent and identically distributed with common probability p . Thus we can view the edges in the graph as a sample from the binomial distribution:

$$P(Y = y) = \prod_{i < j} p^{y_{ij}} (1 - p)^{1 - y_{ij}}.$$

The 4-node complete graph and its independence graph are shown in Figure 1.

2.2. Bernoulli Graph Model

To generalize the Erdős-Rényi model, we relax the assumption that the edges are identically distributed. We call such a model as Bernoulli model and such graphs as Bernoulli graphs. Let

$$P(Y_{ij} = 1) = \frac{\exp \alpha_{ij}}{1 + \exp \alpha_{ij}},$$

then

$$P(Y = y) = \prod_{i < j} \left(\frac{\exp \alpha_{ij}}{1 + \exp \alpha_{ij}} \right)^{y_{ij}} \left(\frac{1}{1 + \exp \alpha_{ij}} \right)^{1 - y_{ij}}.$$

Much of the probabilistic and statistical physics literature focuses on variants of this model and on settings in which statistics of interest involve counts of connections among nodes, e.g., see [15]. The *degree* of a node in a network is the number of connections it has to other nodes and the *degree distribution* is the distribution of these degrees over the entire network. For graphs with directed edges we often consider *in-degrees* (counts of edges coming into nodes) and *out-degrees* (counts of edges going out from nodes) separately. Depending on specifications for α_{ij} the degree distributions or the sequences of degrees (or in-degrees and out-degrees) may be sufficient statistics.

2.2.1. β -model for undirected graphs

The simplest Bernoulli graph model for undirected graphs represents α_{ij} as a sum of β parameters corresponding to the two nodes being linked by an edge, i.e., the probability of the edge between node i and j depends on the sum of the parameters β_i and β_j :

$$P(Y_{ij} = 1) = \frac{e^{\beta_i + \beta_j}}{1 + e^{\beta_i + \beta_j}}.$$

Thus the probability of observing a graph with each edge having the above probability is

$$P(Y = y) = \exp \left\{ \sum_{i=1}^n d_i \beta_i - \phi(\beta) \right\},$$

where d_i 's are the degree sequence of the observed graph. The d_i 's are the sufficient statistics for this model and correspond to both the row and the column totals of the adjacency matrix y . [7] studied properties of this model and [23, 24] characterized the conditions for the existence of the MLE for the β -model.

2.2.2. Holland-Leinhardt p_1 model for directed graphs

The p_1 model of [17] essentially provides variants on a directed version of the β -model and proposes that three factors affect the outcome of a dyad (involving a pair of nodes)

with directed edges: (1) the propensity for individual outgoing ties, α , (2) the propensity for individual incoming ties, β of an individual, (3) and “reciprocity” or mutual linkage between pairs of nodes comprising a dyad, ρ . If we add a parameter for the overall density of edges θ , the form of the joint likelihood for p_1 is

$$\log P(X = x) \propto \theta x_{++} + \sum_i \alpha_i x_{i+} + \sum_j \beta_j x_{+j} + \rho \sum_{ij} x_{ij} x_{ji}, \quad (2)$$

with $K(\rho, \theta, \alpha, \beta)$ as the ERGM normalizing constant. The minimum sufficient statistics are the in-degree and out-degree for each node and the number of dyads with reciprocated edges.

[17] presented an iterative proportional fitting method for maximum likelihood estimation for this model, and discuss the complexities involved in assessing goodness-of-fit. [12] provided a contingency table and log-linear representation of this simple version of p_1 and extend the model to allow for node specific reciprocation where we replace ρ by $\rho + \rho_i + \rho_j$. [23] also characterized the conditions for the existence of the MLE for p_1 , and [21] and [11] studied the algebraic statistical aspects of these models.

2.3. Dependence among edges or dyads

[13] introduced a formal approach towards the study of the dependence structure among edges and proposed the class of Markov random graph models.

The Markov property is usually associated with stochastic processes where the conditional probability distribution of future states of the process (conditional on both past and present values) depends only upon the present state. Graphs display the Markov property when two nodes are conditionally independent of one another (not linked by edges) given the edges linked to the nodes that separate them in the graph. Frank and Strauss introduced the notion of the (dual) dependence graph of a given network graph, in which there is an edge between two nodes if they share a common node in the original graph. Figure 2 shows the Markov dependence graph.

According to the Hammersley-Clifford theorem, the probability function of a general random graph G can be factorized according to its dependence structure D , i.e.,

$$P(G) = c^{-1} \exp \sum_{A \subseteq G} \alpha_A,$$

where c is the normalizing constant and A is a clique of D . This produces a new factorization for graphs with a dependence structure described by the Markov dependence graph.

2.4. Markov Random Graph Model

The Markov random graph model of [13] is built on the notion that two edges are conditionally dependent if they share one common node. The dependence graph thus

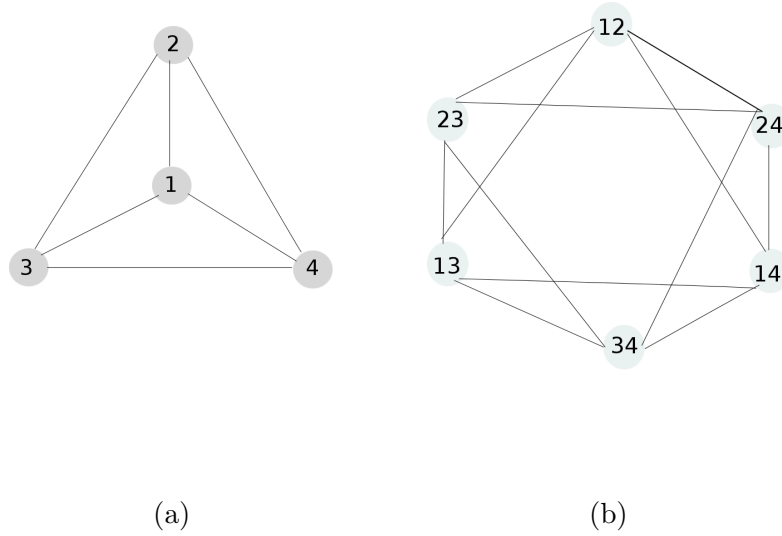


Figure 2: (a) 4-node complete graph (b) Markov dependence graph

characterizes the dependence structure among all possible $\binom{n}{2}$ pair of nodes, either edges or non-edges.

In the dependence graph of the Markov graph the cliques correspond to edges, triangles and different degree of stars in the original network as shown in Figure 3 and 4. By the Hammersley-Clifford theorem, the probability function of graph G under the Markov graph assumption can be written as:

$$P(G) = c^{-1} \exp\left[\sum \tau_{uvw} + \sum_{k=1}^{n-1} \sum \sigma_{v_0 \dots v_k} / k!\right]. \quad (3)$$

Frank and Strauss focus on the homogeneous representation of this model so that each kind of structure has the same parameter.

Let d_k be the number vertices having degree k , then we have the following relationship between the degree distribution and stars, and the relationship between their parameters,

$$s_k = \sum_{j \leq k} \binom{j}{k} d_j \text{ and } \delta_j = \sum_{k \leq j} \binom{j}{k} \sigma_k.$$

Equation (3) becomes

$$P(G) = c^{-1} \exp\left(\tau t + \sum_{j=1}^{n-1} \delta_j d_j\right). \quad (4)$$

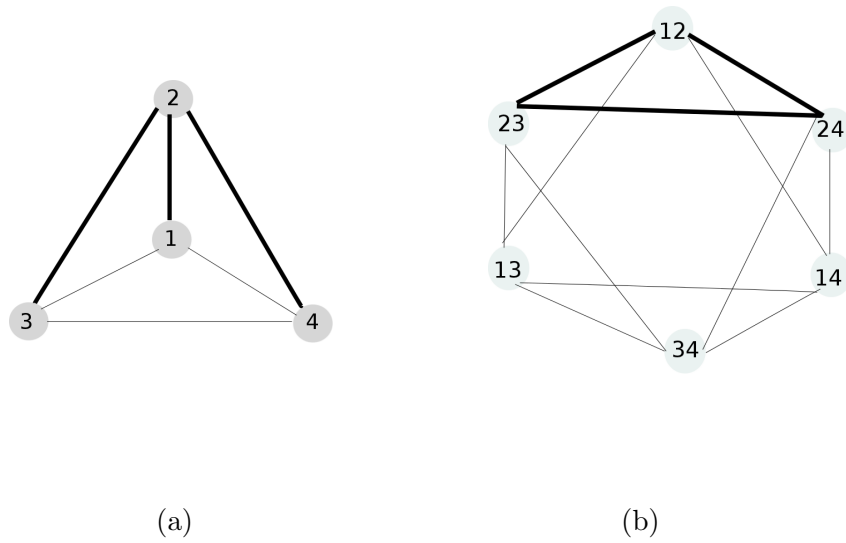


Figure 3: (a) The 3-star highlighted in the complete graph (b) The corresponding triangle in the dependence graph

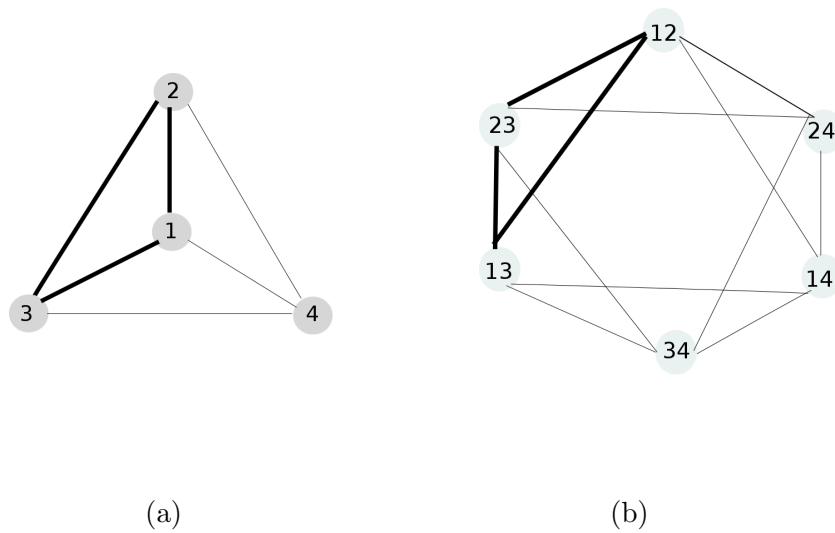


Figure 4: (a) The triangle highlighted in the complete graph (b) The corresponding triangle in the dependence graph

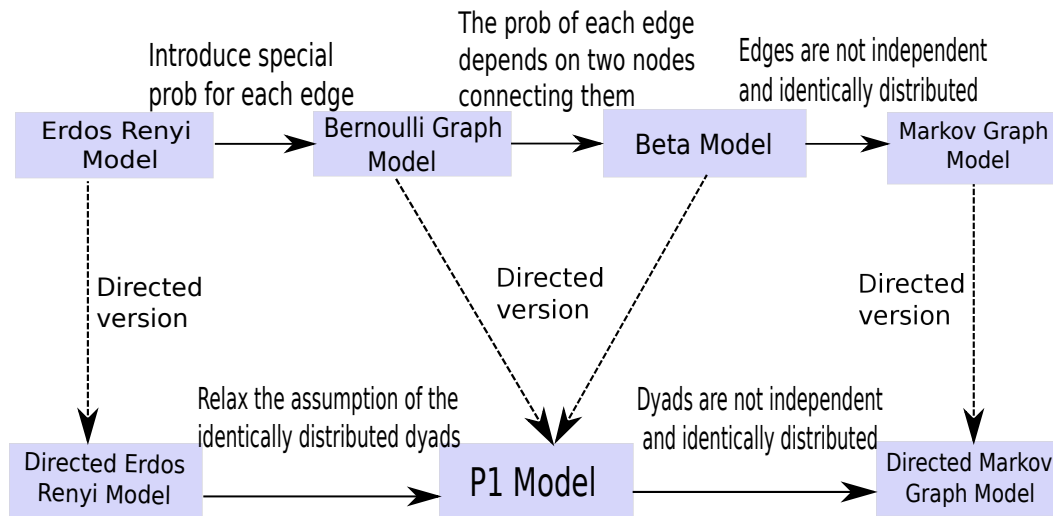


Figure 5: The relationships among subclasses of ERGMs

The dependence structure makes the distribution difficult to factorize, so that the computation of normalizing constant c becomes a major barrier for parameter estimation.

2.5. Connections among subclasses of ERGMs

The special subclasses of ERGMs we have described are related by layers of assumptions on the dependence structure of the distributions over edges or dyads. We summarize their relationships in Figure 5. On the top row are models for undirected graphs and on the bottom row are models for directed graphs. Imposing different assumptions and following the arrows, we can see the increasing complexity of the subclasses of ERGMs.

3. Estimation Methods

3.1. Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) has been a mainstay of the inferential statistics toolkit, and it is based on the maximization of the likelihood function with respect to the parameters given the observed data.

In detail, given the distribution of a random graph Y as in Equation 1, we can write the likelihood function as:

$$l(\theta; y) = \log(P_{\theta, \mathcal{Y}}(Y = y)) = \theta \cdot T(y) - \kappa(\theta), \quad (5)$$

where $\kappa(\theta) = \log c(\theta, \mathcal{Y})$ and $c(\theta, \mathcal{Y}) = \sum_{y \in \mathcal{Y}} \exp(\theta^T T(y))$. The maximum likelihood estimator (MLE) $\hat{\theta}$ of the parameters θ is the (unique) maximizer of the log-likelihood

function (5), if well defined. Formally

$$l(\hat{\theta}, y) = \sup_{\theta \in \mathbb{R}^q} l(\theta, y).$$

The MLE is said to be nonexistent when the above supremum is not attained at any vector in \mathbb{R}^q . For ERGMs nonexistence implies that the log-likelihood function is maximized along certain sequences of parameter vectors with norm diverging to infinity. See [22].

For ERGMs the calculation of MLEs is complicated by the normalizing constant, especially in settings where the edges have a dependence structure. In the complete independence of edges case, the probability function factorizes in a nice way so that we don't need to enumerate all graphs when estimating the normalizing constant. There is no such nice property, however, for the Markov random graph models. This is precisely why the Maximum Pseudo-likelihood Estimation (MPLE) was introduced in the late 1980s, to estimate parameters in ERGMs.

3.2. Maximum Pseudo Likelihood Estimation

Now we consider the problem in another way. As before, we denote Y_{ij} as the edge connecting node i and j . Let $P(Y_{ij} = 1|Y_{ij}^c)$ be the conditional probability where Y_{ij}^c is the graph after removing edge ij . Then, we have

$$P(Y_{ij} = 1|Y_{ij}^c) = \frac{P(Y_{ij} = 1, Y_{ij}^c)}{P(Y_{ij}^c)} = \frac{P(Y_{ij} = 1, Y_{ij}^c)}{P(Y_{ij} = 1, Y_{ij}^c) + P(Y_{ij} = 0, Y_{ij}^c)} = \frac{\exp[\theta \cdot \delta(y_{ij}^c)]}{1 + \exp[\theta \cdot \delta(y_{ij}^c)]},$$

where $\delta(y_{ij}^c) = T(y_{ij}^+) - T(y_{ij}^-)$ is the change of sufficient statistics when y_{ij} changes from 0 to 1. Y_{ij}^+ and Y_{ij}^- represent graphs by setting $Y_{ij} = 1$ or 0 with the remainder of the network Y_{ij}^c fixed. Note that this has a logistic regression form and we can estimate the parameters by fitting a logistic regression model using the observed network and the change of sufficient statistics as shown above. The pseudo-likelihood is:

$$l_P(\theta; y) = \theta \cdot \sum_{ij} \delta(y_{ij}^c) y_{ij} - \sum_{ij} \log(1 + \exp(\theta^T \delta(y_{ij}^c))), \quad (6)$$

and the MPLE maximizes this pseudo-likelihood.

Let us consider the case where edges are independent of one another. Then $P(Y_{ij} = 1|Y_{ij}^c) = P(Y_{ij} = 1)$. This indicates that the pseudo-likelihood is the same as the likelihood and the MPLE is the same as the MLE in this scenario. But for Markov random graph models where the independence assumption doesn't hold, the conditional likelihood is no longer the same as the likelihood. How the MPLEs and MLEs differ is of our interests for the edge dependent case.

[3] showed the MLE exists if and only if $t(y_{observed}) \in rint(C)$ where C is the convex hull formed by all possible sufficient statistics. By $rint(C)$ we mean the relative interior of convex hull C . Similarly $rbd(C)$ is the relative boundary of C .

As we noted above, we can compute the MPLE by fitting a logistic regression. For each possible edge Y_{ij} of a graph, we get the difference between the sufficient statistics by

adding (when $Y_{ij} = 0$) or removing this edge (when $Y_{ij} = 1$). For example consider the ERGM for an undirected graph where the sufficient statistics are the number of edges, triangles and 2-stars. Then we count the number of edges, triangles and 2-stars by adding or removing one of the $\binom{n}{2}$ edges. Then we treat each possible edge as the response variable and the sufficient statistics difference as the covariates. In this sense, the existence of MPLE is equivalent to the existence of MLE for logistic regression. According to [1] and [27], we have: A necessary and sufficient condition for the MPLE to exist is $\forall \alpha \in R^q, \exists i, j$ such that $(2y_{ij} - 1)\alpha^T \delta(y_{ij}^c) < 0$, which is equivalent to the fact that the MPLE exists unless a separating hyperplane exists between the scatterplot of ties and non-ties in the space defined by the $\delta(y_{ij}^c)$ (assuming a full-rank design matrix).

More specifically, these conditions for the existence of MPLE can be characterized as follows:

1. **Complete separation.** There exists a solution a and b to the following linear programming problem,

$$ax - b > 0, ay - b < 0$$

where x is the matrix whose rows are of the vectors $\delta(y_{ij}^c)$ corresponding to the realized edges in the graph and y is the matrix whose rows are the vectors $\delta(y_{ij}^c)$ corresponding to non-edges. In another word, there exist a hyperplane separating the vectors $\delta(y_{ij}^c)$ corresponding to edges and non-edges. In this case, the MLE of the logistic regression parameters (the MPLE) does not exist.

2. **Quasi separation.** There exists a solution a and b to the following linear programming problem,

$$ax - b \geq 0, ay - b \leq 0$$

and the equality has to hold for at least one data point. In this case, there exists a hyperplane separating data points belonging to two classes, however, both classes have data points on the separation hyperplane and the other data points are completely separated. In this case the MPLE does not exist either.

3. **Overlap.** There is no solution to the above linear programming problem. In this case, the MPLE exists.

These three conditions are mutually exclusive and exhaustive, that is, all observations will fall into one of the three categories.

4. Computational Results

In this section, we compare the MLE and MPLE for the dependent models with the number of edges, triangles and 2-stars as minimal sufficient statistics. The goal of this study is to give a general idea on how these two estimation methods differ for models with dependence among edges.

4.1. Results on small graphs

It is possible to enumerate all graphs if n is small and we can get the exact estimation of the partition function and MLE. The availability of the convex hull formed by all MSSs helps us to understand and characterize the existence of MLE. Our first experiment is based on all 7, 8 and 9-node graphs.

Table 1 summarizes the number of graphs and number of MSSs for 7, 8 and 9-node graphs. Table 2 shows the proportion of MSSs inside and on the boundary of the convex hull. Those on the boundary are cases whose MLEs don't exist. As the number of nodes increases, the number of graphs increases exponentially and the number of different sufficient statistics increases proportionally. But the proportion of them on the boundary of the convex hull decreases.

Table 1: Summary of MSSs for 7, 8 and 9-node graphs

	# graphs	MSS	# diff MSS
7-node	$2^{21} = 2M$	E, Tri, 2-stars	390
8-node	$2^{28} = 268M$	E, Tri, 2-stars	1274
9-node	$2^{36} = 68B$	E, Tri, 2-stars	3746

Table 2: Some statistics of the convex hulls for 7,8 and 9-node graphs

	insideboundary	onboundary	prop onboundary
7-node	252	138	0.354
8-node	1,003	271	0.213
9-node	3,239	507	0.135

We show the two dimensional and three dimensional MSSs plots for 7-node graphs in Figure 6. [16] and [22] use this plot to illustrate the geometry of sufficient statistics for ERGMs. Those graph realizations on the boundary of the convex hull correspond to the non-existence of MLEs.

4.1.1. Existence of MLEs and MPLEs

The following theorem captures the relationship between the existence of MLE and the existence of MPLE:

Theorem 1. *The existence of the MPLE implies the existence of the MLE.*

Proof. We will show the equivalent statement that the MLE doesn't exist implies the MPLE doesn't exist. The MLE exists if and only if $T(y) \in \text{ri}(C)$ where C is the convex hull formed by all sufficient statistics (see [3]).

For any sufficient statistic t on the $\text{rbd}(C)$, there exists a vector α such that $\alpha^\top(t - t') \leq 0$ for any $t' \in C$. When computing the MPLE, we are actually fitting a logistic regression

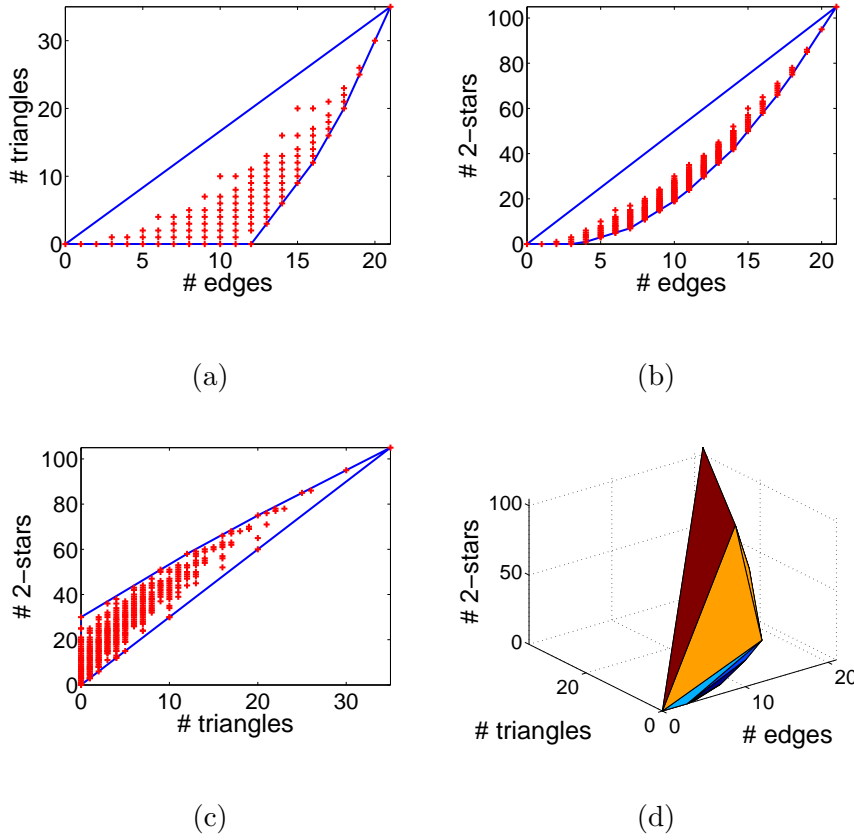


Figure 6: Convex hull formed by sufficient statistics of 7-node graphs (a) number of edges vs number of triangles (b) number of edges vs number of 2 stars (c) number of triangles vs number of 2 stars (d) number of edges, number of triangles vs number of 2 stars

with 1's corresponding to the observed edges and 0's to the non-edges. For every pair of nodes (i, j) , $\delta(y_{i,j}^c) = t - t(y_{i,j}^-)$ if there is an edge between them and $\delta(y_{i,j}^c) = t(y_{i,j}^+) - t$ otherwise. We also have, for every pair (i, j) ,

$$\alpha^\top (t - t(y_{i,j}^-)) \leq 0 \text{ and } \alpha^\top (t - t(y_{i,j}^+)) \leq 0$$

where $t = t(y_{i,j}^+)$ if there is an edge between i and j and $t = t(y_{i,j}^-)$ otherwise. This implies that $\alpha^\top \delta(y_{i,j}^c) \leq 0$ if there is an edge between i and j and $\alpha^\top \delta(y_{i,j}^c) \geq 0$ otherwise. This shows that the 1's and 0's are quasi-completely separated and the MLE for the logistic regression does not exist in this case.

As a consequence of this theorem, we can fit the MPLE before fitting the MLE. If we know that the MPLE exists, then we are certain that the MLE exists and we can proceed to fit the MLE using MCMC. In the near degeneracy case, more than one run of the MCMC sampler may be needed to get reasonable estimates because it is difficult to sample enough

graphs with different observed sufficient statistics but the MCMC should converge with a long enough chain. Unfortunately instability near the boundary also means that the variance associated with the MLE will have some increasingly large components.

We further show the relationship between the MPLE and MLE empirically in Table 3 and 4. All graphs with the same sufficient statistics form a fiber. It is obvious that the MLEs are the same for graphs in a fiber. The MPLEs could be different. We evaluate the existence of the MLE for each fiber and the MPLEs for each graph in all fibers. The cases of (E)xistence and (N)on-existence are shown in the contingency tables where “(E)”, “(N)” and “(EN)” represent existence, non-existence and cases when the estimates both exist and non-exist. For example, the number of different MSSs is 390 for 7-node graphs: 129 of them are cases when both the MLEs and MPLEs exist, 22 of them are cases when the MLEs exist but the MPLEs don’t exist, and 101 of them are cases when the MLE exists but the MPLE exists in some cases but not in others. We find similar patterns for the 8-node graphs. The results illustrate Theorem 1.

Table 3: The contingency table regarding the existence of MLEs and MPLEs for fibers of 7-node graphs

	(E).MPLE	(N).MPLE	MPLE-(EN)
(E).MLE	129	22	101
(N).MLE	0	138	0

Table 4: The contingency table regarding the existence of MLEs and MPLEs for fibers of 8-node graphs

	(E).MPLE	(N).MPLE	MPLE-(EN)
(E).MLE	190	66	741
(N).MLE	0	250	0

4.1.2. Estimation of the Model Parameters

We study the difference between MLE and MPLE with respect to parameter estimation. We denote the two different parameter estimates by $\theta_{MLE} = \hat{\theta}$ and $\theta_{MPLE} = \tilde{\theta}$, and the distributions under θ_{MLE} and θ_{MPLE} are $p_{\hat{\theta}}$ and $p_{\tilde{\theta}}$. We compute the MLE of the Markov random graph models for each sufficient statistics and compute the MPLEs for all the graphs within that fiber. Figure 7 displays the number of different MPLE estimates in each fiber of 7-node graphs for the cases in which both the MLE and MPLE exist. For readability we sorted these different values in the ascending order. The x -axis labels the indices of the sorted sufficient statistics from 1 to 129. We have obtained the values displayed on the y -axis by computing the L_1 distances between the estimates $p_{\hat{\theta}}$ and $p_{\tilde{\theta}}$ and counting the number of different values of these distances in each fiber.

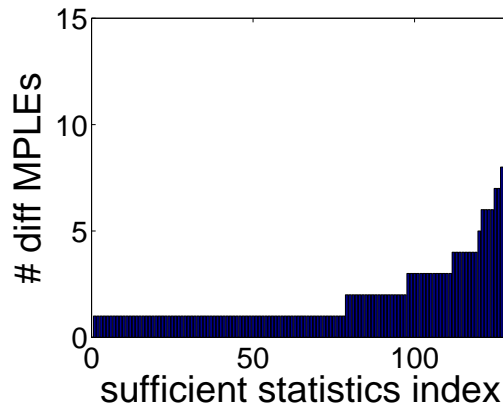


Figure 7: The number of different MPLEs estimated from graphs in the same fiber for 7-node graphs

When both the MLE and the MPLE exist, we compare the L_2 distance between $\hat{\theta}$ and $\tilde{\theta}$, the L_1 distance between $p_{\hat{\theta}}$ and $p_{\tilde{\theta}}$ and the entropy of distributions $p_{\hat{\theta}}$ and $p_{\tilde{\theta}}$, where the MPLEs $\tilde{\theta}$ are estimated from all graphs in the fiber. Figure 8 shows these comparisons for 7-node graphs. Again, to give a better illustration of the differences, we sort the sufficient statistics according to the minimum difference between the MLE and MPLE in each sub-figure. Note that the MLEs are the same for graphs with the same sufficient statistics. The MPLEs might differ, however, as we show in Figure 7. It is easy to see that both the parameter estimates and the distributions under the two estimates differ considerably in some cases. In particular, 8(b) show the L_1 (total variation) distance between the distributions $p_{\hat{\theta}}$ and $p_{\tilde{\theta}}$ based on the sufficient statistics (and their ordering) considered in Figure 7. For most fibers, the values of the distances are greater than 0, indicating that in those cases the MLE and the MPLE parametrize different distributions. Remarkably, in a number of cases, the distance gets as large as 2 which means the two distributions are supported on disjoint sets, even if the MLE $\hat{\theta}$ and the MPLE $\tilde{\theta}$ are estimated using the graphs in the same fiber. Figure 8 (c) and (d) display the values of the entropies. Note that the distributions with (nearly) 0 entropy are degenerate as it only has support on a single or a small number of values. In particular, we see degenerate behaviors for the MLE even when the MPLE exists.

4.2. Results on large graphs

We now compare MLEs and MPLEs on large graphs. Because it is essentially infeasible to enumerate all graphs for more than $n = 10$ nodes, we use an MCMC sampling algorithm to obtain the MLE approximately, instead of estimating the MLE directly.

Our study of the two approaches is structured as follows. Given a ground truth parameter $\theta = (0.0384, 0.2378, -0.0853)$ for edge, triangle and 2-star which corresponds to

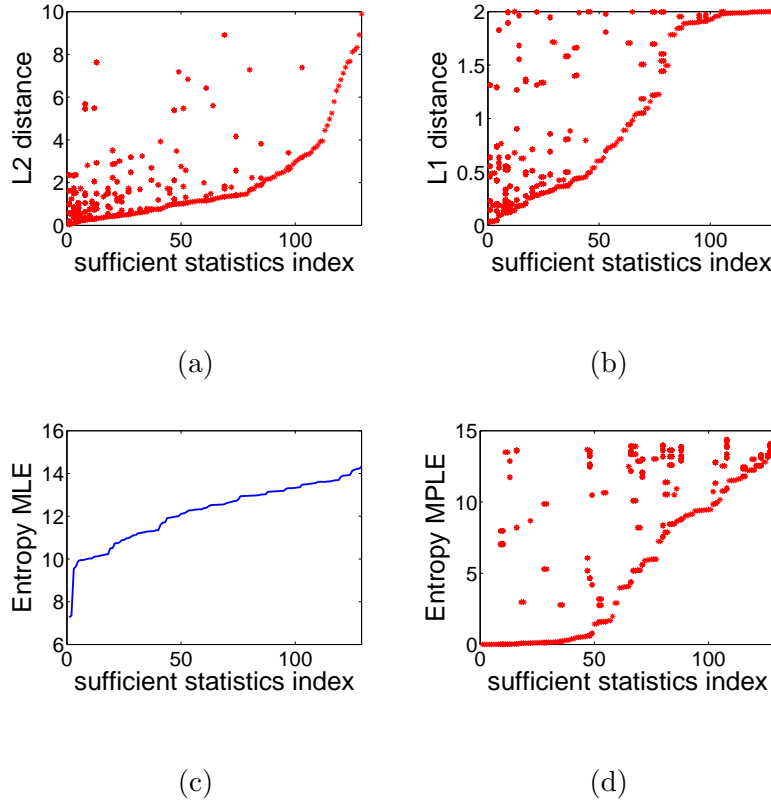


Figure 8: 7-node graphs: when both MLE and MPLE exists, (a) L_2 distance between $\hat{\theta}$ and $\tilde{\theta}$ (b) L_1 distance between $p_{\hat{\theta}}$ and $p_{\tilde{\theta}}$ (c) Entropy of $p_{\hat{\theta}}$ (d) Entropy of $p_{\tilde{\theta}}$

a model far from degenerate, we sample 100 graphs with specific number of nodes from this model. Fit MLEs and MPLEs using each group of graphs to get $\hat{\theta}$'s and $\tilde{\theta}$'s. Figure 9 shows the histograms of the edge parameter of $\hat{\theta}$'s and $\tilde{\theta}$'s when the number of nodes $n = 100, 200, 400$ and 600 . Table 5 shows the mean and standard error of the estimated edge parameter of $\hat{\theta}$'s and $\tilde{\theta}$'s. In this specific case, the MLE and MPLE do not differ very much, even asymptotically. Table 6 and 7 show the summary statistics for triangle and 2-star parameters. They also indicate that both MLE and MPLE are asymptotically unbiased and MPLE is a good approximation of MLE in this case.

Our experiments, however, show that most parameters in the parameter space correspond to degenerate models as the number of nodes increases. Now we pick a parameter value $\theta = (-3.1043, -1.8940, 1.0219)$ corresponding to a near degenerate model based on our previous experiment. We conduct similar study as on the non-degenerate case. Figure 10 and Table 8 show that the sufficient statistics of graphs sampled from $p_{\hat{\theta}}$ and $p_{\tilde{\theta}}$ no longer center around the true value which indicates that the model doesn't fit the data well. We show the results when $n = 100$ and $n = 200$. We further find that degeneracy happens when $n = 400$.

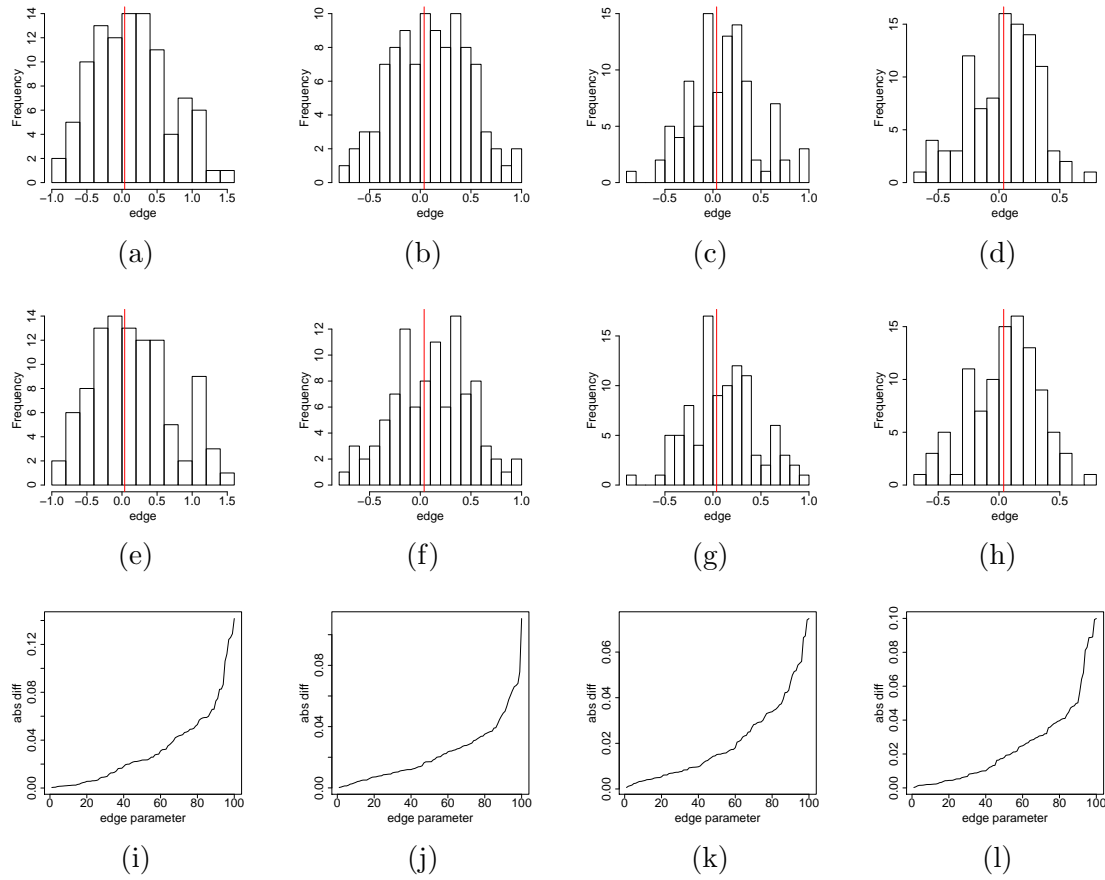


Figure 9: The histogram of edge parameters in $\hat{\theta}$ for (a) 100-node, (b) 200-node, (c) 400-node, and (d) 600-node graphs, edge parameters in $\hat{\theta}$ for (e) 100-node, (f) 200-node, (g) 400-node, and (h) 600-node graphs and the absolute value difference between $\hat{\theta}$ and $\bar{\theta}$ (i) 100-node, (j) 200-node, (k) 400-node and (l) 600-node graphs with true model $\theta = (0.0384, 0.2378, -0.0853)$

Table 5: The mean and standard error of simulated edge parameters with true value 0.0384

Edge parameter	MLE		MPLE	
	mean	se	mean	se
$n = 100$	0.139	0.541	0.161	0.555
$n = 200$	0.096	0.376	0.107	0.373
$n = 400$	0.097	0.331	0.102	0.327
$n = 600$	0.037	0.275	0.047	0.285

4.3. Example: Zachary's Karate Club

[37] collected network information on the relationships among the 34 members of a university karate club. Figure 11 shows the observed network structure. Our purpose here

Table 6: The mean and standard error of simulated triangle parameters with true value 0.2378

Triangle parameter	MLE		MPLE	
	mean	se	mean	se
$n = 100$	0.228	0.064	0.229	0.065
$n = 200$	0.235	0.036	0.235	0.036
$n = 400$	0.242	0.026	0.242	0.026
$n = 600$	0.239	0.027	0.239	0.027

Table 7: The mean and standard error of simulated 2-star parameters with true value -0.0853

2-star parameter	MLE		MPLE	
	mean	sd	mean	sd
$n = 100$	-0.088	0.021	-0.089	0.022
$n = 200$	-0.088	0.014	-0.088	0.014
$n = 400$	-0.087	0.008	-0.087	0.009
$n = 600$	-0.085	0.007	-0.086	0.007

Table 8: The mean and standard error of edge parameters in $\hat{\theta}$ and $\tilde{\theta}$ with true value -3.1043

Edge parameter	MLE		MPLE	
	mean	se	mean	se
$n = 100$	0.329	0.831	0.361	0.847
$n = 200$	0.162	0.442	0.192	0.458

is simply to use this network to illustrate differences between MLEs and MPLEs for dyadic-dependent ERGMs, and several of the fitted models actually provide a poor description of the data which, as other authors demonstrate and as Figure 11 shows, consist of two or three somewhat connected blocks.

We show the sufficient statistics for various ERGMs of increasing complexity associated with this network in Table 9.

We calculated the MLEs for these ERGMs using the MCMC sampler in the R package “ergm”. Table 10 shows the fitted parameters along with their estimated standard errors for the MLE and MPLE for a sequence of 8 increasingly complex ERGMs.

The MPLE exists for all of these models, which implies the MLE exists as well. Model 1 is the Erdős-Rényi model and for it everything is nice. The other models belong to the Markov random graph model family. From the table we see that the MLEs and MPLEs differ for these dyadic-dependent ERGMs; occasionally they even have different signs. Moreover, near-degeneracy of the MLE occurs when we fit these dependent models. For example, for model 2 the standard errors are very large suggesting that we are approaching the boundary of the parameter space. For model 3, 4 and 5 the MCMC sampler for computing the MLE gets stuck at one graph; thus the standard error estimate of 0.

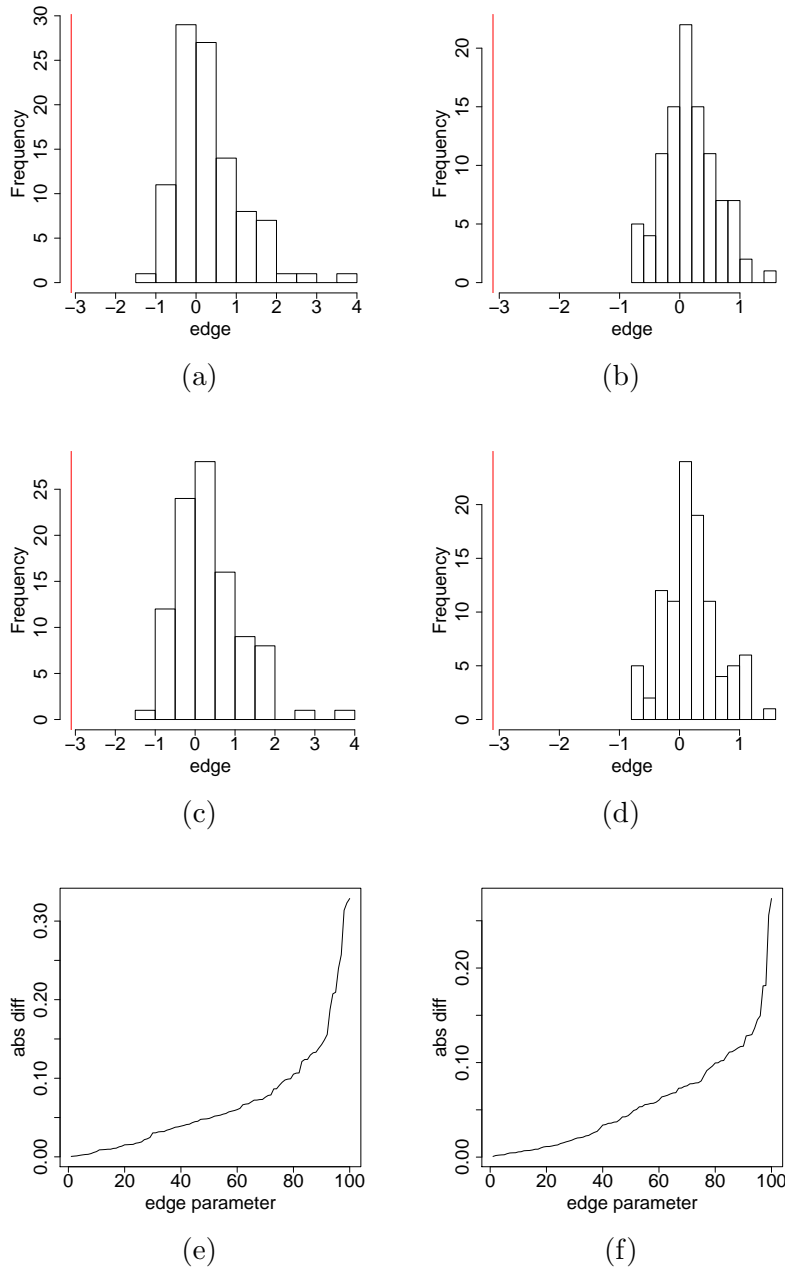


Figure 10: The histogram of edge parameters in $\hat{\theta}$ for (a) 100-node, (b) 200-node graphs, edge parameters in $\tilde{\theta}$ for (c) 100-node, (d) 200-node graphs and the absolute value difference between $\hat{\theta}$ and $\tilde{\theta}$ (e) 100-node, (f) 200-node graphs with true model $\theta = (-3.1043, -1.8940, 1.0219)$

The “ergm” package does not give valid estimates for model 7 and 8 and the MCMC sampler exits with errors. This behavior is due to the poor convergence of the MCMC

for these models. Finally for the intermediate model 6, the MLE results show estimated standard errors that are large relative to the estimated values, again suggesting some form of problematic behavior.

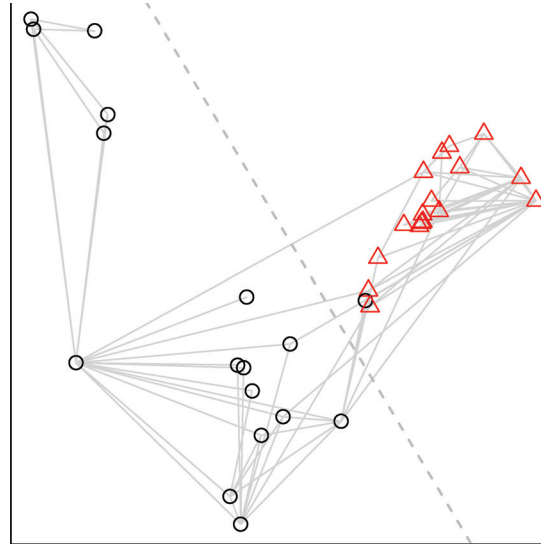


Figure 11: The network structure of the Zachary karate club data among 34 individuals

Table 9: Sufficient statistics of the Zachary karate club network data

edges	triangle	kstar2	kstar3	kstar4	kstar5	kstar6	kstar7	kstar8
78	45	528	1764	5082	11741	21604	31836	37729
kstar9	kstar10	kstar11	kstar12	kstar13	kstar14	kstar15	kstar16	kstar17
35981	27523	16756	8009	2940	800	152	18	1

5. Conclusion

In this paper, we have described the relationships among different subclasses of Exponential Random Graph Models (ERGMs) from the perspective of dependence among edges. The introduction of edge dependence, e.g., in the form of Markov dependence, yields a probability function for the graph that no longer factorizes in a nice way. We described the implications of the complexities introduced by edge dependence for estimation and we introduced Maximum Pseudo-Likelihood Estimation (MPLE) as an alternative to the more standard Maximum Likelihood Estimation (MLE), an approach which does not involve computing a complex normalizing constant and can be fitted using logistic regression. We then studied the relationship between MLE and MPLE for ERGMs, ex-

Table 10: MLEs and MPLEs for the parameters of 8 ERGMs with increasing parametric complexity with standard errors shown in the parenthesis

model	edges	kstar2	triangle	kstar3	kstar4	kstar5	kstar6
MLE							
1	-1.82(0.12)						
2	-3.67(1300.5)	0.18(20.3)					
3	-3.941(0)	0.11(0)	0.44(0)				
4	-3.41(0)	0.11(0)		0.04(0)			
5	-2.95(0)	-0.05(0)	0.58(0)	0.02(0)			
6	-0.46(3.86)	0.27(0.19)	-0.54(0.69)	0.19(0.17)	0.03(0.02)		
7	Degenerate model and no reasonable model is fitted using the “ergm” package.						
8							
MPLE							
1	-1.823(0.12)						
2	-3.68(0.3)	0.18(0.02)					
3	-3.95(0.33)	0.15(0.03)	0.463(0.13)				
4	-3.41(0.56)	0.13(0.09)		0.006(0.01)			
5	-2.96(0.57)	-0.05(0.1)	0.58(0.14)	0.02(0.01)			
6	1.24(0.9)	-1.21(0.23)	0.67(0.16)	0.38(0.06)	-0.05(0.01)		
7	0.75(1.39)	0.67(0.16)	-1.04(0.44)	0.29(0.2)	-0.02(0.06)	-0.004(0.01)	
8	9.99(2.19)	0.71(0.17)	-4.85(0.83)	2.83(0.51)	-1.24(0.23)	0.38(0.07)	-0.06(0.01)

amining theoretical properties, exact calculations for small graphs and a simulation. We also illustrated the connections using the Zachary karate club data.

The two forms of estimation we explore here, MLE and MPLE, differ in several aspects. The existence of MPLE for an ERGM implies the existence of MLE. When both of them exist, the estimation $\hat{\theta}$ and $\tilde{\theta}$ and the resulted model $p_{\hat{\theta}}$ and $p_{\tilde{\theta}}$ are different sometimes according to our simulation study on small graphs. The difference is quite large in the near-degenerate settings. We also evaluated their difference by sampling graphs with increasing number of nodes from the true model and fit parameters using the two methods when the number of nodes is more than 10. The MLE and MPLE differ little when the true parameter lies within the convex hull and the model has a large entropy. The experimental results empirically show that the variance of both the MLE and MPLE decreases as the number of nodes increases. This is a sign of consistency of MLE and MPLE for ERGMs. Our analysis of the karate club data also shows the MLEs and MPLEs can however be different, sometimes in substantial ways.

[29] described a form of model consistency as the size of the network n grows, which fails to hold for edge-dependent ERGMs. Yet the dual graph or dependence graph interpretation associated with Markov random graph models remains alluring. Thus, despite the estimation problems we have highlighted in this paper, we believe that the complexities of edge-dependent ERGMs bear further investigation.

Acknowledgements

This research was supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office and by contracts FA9550-12-1-0392 and AFOSR/DARPA grant FA9550-14-1-0141 from the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA).

References

- [1] Albert, A. and Anderson, J. A. “On the existence of maximum likelihood estimates in logistic regression models.” *Biometrika*, 71(1):1–10 (1984).
- [2] Arnold, B. C. and Strauss, D. “Pseudolikelihood estimation: Some examples.” *The Indian Journal of Statistics, Series B*, 53(2):233–243 (1991).
- [3] Barndorff-Nielsen, O. *Information and Exponential Families in Statistical Theory*. New York: John Wiley (1978).
- [4] Besag, J. “Spatial interaction and the statistical analysis of lattice systems.” *Journal of the Royal Statistical Society, Series B*, 36(2):192–236 (1974).
- [5] Besag, J. “Efficiency of pseudo-likelihood estimation for simple Gaussian fields.” *Biometrika*, 64(3):616–618 (1977).
- [6] Besag, J. “Markov chain Monte Carlo for statistical inference.” Technical report, University of Washington, Center for Statistics and the Social Sciences (2000).
- [7] Chatterjee, S., Diaconis, P., and Sly, A. “Random graphs with a given degree sequence.” *The Annals of Applied Probability*, 21(4):1400–1435 (2011).
- [8] Cox, D. R. and Reid, N. “A note on pseudo-likelihood constructed from marginal densities.” *Biometrika*, 91:729–737 (2004).
- [9] Demaris, B. and Cranmer, S. J. “Consistent confidence intervals for maximum pseudo-likelihood estimators.” In *NIPS workshop: Computational Social Science and the Wisdom of Crowds* (2010).
- [10] Fienberg, S. E. “A brief history of statistical models for network analysis and open challenges.” *Journal of Computational and Graphical Statistics*, 21(4):825–839 (2012).
- [11] Fienberg, S. E., Petrović, S., and Rinaldo, A. “Algebraic Statistics for p_1 Random Graph Models: Markov Bases and Their Uses.” In Dorans, N. J. and Sinharay, S. (eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, volume 202 of *Lecture Notes in Statistics*, 21–38. New York: Springer (2011).

- [12] Fienberg, S. E. and Wasserman, S. “An exponential family of probability distributions for directed graphs: Comment.” *Journal of the American Statistical Association*, 76(373):54–57 (1981).
- [13] Frank, O. and Strauss, D. “Markov graphs.” *Journal of the American Statistical Association*, 81(395):832–842 (1986).
- [14] Friedman, J., Hastie, T., and Tibshirani, R. “Sparse inverse covariance estimation with the graphical lasso.” *Biostatistics*, 9(3):432–41 (2008).
- [15] Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. “A survey of statistical network models.” *Foundations and Trends in Machine Learning*, 2(2):129–233 (2010).
- [16] Handcock, M. S., Robins, G., Snijders, T., Moody, J., and Besag, J. “Assessing degeneracy in statistical models of social networks.” *Center for Statistics and the Social Sciences, University of Washington, Working Paper No. 39* (2003).
- [17] Holland, P. W. and Leinhardt, S. “An exponential family of probability distributions for directed graphs.” *Journal of the American Statistical Association*, 76(373):33–50 (1981).
- [18] Kolaczyk, E. D. *Statistical Analysis of Network Data: Methods and Models*. New York: Springer (2009).
- [19] Meinshausen, N. and Bühlmann, P. “High dimensional graphs and variable selection with the Lasso.” *Annals of Statistics*, 34(3):1436–1462 (2006).
- [20] Pattison, P. and Wasserman, S. “Logit models and logistic regressions for social networks: II. Multivariate relations.” *British Journal of Mathematical and Statistical Psychology*, 52:169–193 (1999).
- [21] Petrović, S., Rinaldo, A., and Fienberg, S. E. “Algebraic statistics for a directed random graph model with reciprocation.” In Viana, M. and Wynn, H. (eds.), *Algebraic Methods in Statistics and Probability II*, volume 516 of *Contemporary Mathematics*, 261–283. American Mathematical Society (2010).
- [22] Rinaldo, A., Fienberg, S. E., and Zhou, Y. “On the geometry of discrete exponential families with application to exponential random graph models.” *Electronic Journal of Statistics*, 3:446–484 (2009).
- [23] Rinaldo, A., Petrovic, S., and Fienberg, S. E. “Maximum likelihood estimation in network models.” *arXiv:1105.6145v2* (2012).
- [24] Rinaldo, A., Petrović, S., and Fienberg, S. E. “Maximum likelihood estimation in the β -model.” *Annals of Statistics*, 41:1085–1110 (2013).

- [25] Robins, G. “Neighborhood-based models for social networks.” *Sociological Methodology*, 32:301–337 (2002).
- [26] Robins, G., Pattison, P., Kalish, Y., and Lusher, D. “An introduction to exponential random graph (p^*) models for social networks.” *Social Networks*, 29(2):173–191 (2007).
- [27] Santner, J. T. and Duffy, E. D. “A note on A. Albert and J. A. Anderson’s conditions for the existence of maximum likelihood estimates in logistic regression models.” *Biometrika*, 73(3):755–758 (1986).
- [28] Schweinberger, M. “Instability, sensitivity, and degeneracy of discrete exponential families.” *Journal of the American Statistical Association*, 106:1361–1370 (2011).
- [29] Shalizi, C. R. and Rinaldo, A. “Consistency under sampling of exponential random graph models.” *Annals of Statistics*, 41(3):508–535 (2013).
- [30] Snijders, T. A. B. “Markov chain Monte Carlo estimation of exponential random graph models.” *Journal of Social Structure*, 3(2) (2002).
- [31] Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. “New specifications for exponential random graph models.” *Sociological Methodology*, 36(1):99–153 (2006).
- [32] Strauss, D. “On a general class of models for interaction.” *SIAM Review*, 28(4):513–527 (1986).
- [33] Strauss, D. and Ikeda, M. “Pseudo-likelihood estimation for social networks.” *Journal of the American Statistical Association*, 85(409):204–212 (1990).
- [34] van Duijn, M., Gile, K., and Handcock, M. “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).
- [35] Wasserman, S. and Pattison, P. “Logit models and logistic regressions for social networks: an introduction to Markov graphs and p^* .” *Psychometrika*, 61(3):401–425–425 (1996).
- [36] Wasserman, S. and Robins, G. L. “An introduction to random graphs, dependence graphs, and p^* .” In Carrington, P. J., Scott, J., and Wasserman, S. (eds.), *Models and Methods in Social Network Analysis*, 148–161. Cambridge University Press (2005).
- [37] Zachary, W. W. “An information flow model for conflict and fission in small groups.” *Journal of Anthropological Research*, 33:452–473 (1977).