On the Question of Effective Sample Size in Network Modeling

Eric D. Kolaczyk and Pavel N. Krivitsky*

October 23, 2012

Abstract

The modeling and analysis of networks and network data has seen an explosion of interest in recent years and represents an exciting direction for potential growth in statistics. Despite the already substantial amount of work done in this area to date by researchers from various disciplines, however, there remain many questions of a decidedly foundational nature — natural analogues of standard questions already posed and addressed in more classical areas of statistics — that have yet to even be posed, much less addressed. Here we raise and consider one such question in connection with network modeling. Specifically, we ask, “Given an observed network, what is the sample size?” Using simple, illustrative examples from the class of exponential random graph models, we show that the answer to this question can very much depend on basic properties of the networks expected under the model, as the number of vertices $N_v$ in the network grows. In particular, we show that whether the networks are sparse or not under our model (i.e., having relatively few or many edges between vertices, respectively) is sufficient to change the asymptotic rates for maximum likelihood parameter estimation by an order of magnitude, from $N_v^{1/2}$ to $N_v$. We then explore

*Eric D. Kolaczyk is Professor, Department of Mathematics and Statistics, Boston University, Boston, MA 02215 and Pavel N. Krivitsky is Research Associate, Department of Statistics, Pennsylvania State University, University Park, PA 16802. (Email for correspondence: kolaczyk@bu.edu.) This work was begun during the 2010–2011 Program on Complex Networks at SAMSI and was partially supported by ONR award N000140910654 (EDK), ONR award N000140811015 (PNK), NIH award 1R01HD068395-01 (PNK), and Portuguese Foundation for Science and Technology Ciência 2009 Program (PNK). Krivitsky also wishes to thank David Nolin for the Lalamera food-sharing network dataset.
some practical implications of this result, using both simulation and data on food-sharing from Lamalera, Indonesia.

**Keywords:** Asymptotic normality; Consistency; Exponential random graph model; Maximum likelihood.

## 1 Introduction

Since roughly the mid-1990s, the study of networks has increased dramatically. Researchers from across the sciences — including biology, bioinformatics, computer science, economics, engineering, mathematics, physics, sociology, and statistics — are more and more involved with the collection and statistical analysis of data associated with networks. As a result, statistical methods and models are being developed in this area at a furious pace, with contributions coming from a wide spectrum of disciplines. See, for example, Jackson (2008), Kolaczyk (2009), and Newman (2010) for recent overviews from the perspective of economics, statistics, and statistical physics, respectively.

A network is typically represented mathematically by a graph, say, $G = (V, E)$, where $V$ is a set of $N_v$ vertices (commonly written $V = \{1, \ldots, N_v\}$) and $E$ is a set of $N_e$ edges (represented as vertex pairs $(u, v) \in E$). Edges can be either directed (wherein $(u, v)$ is distinct from $(v, u)$) or undirected. Prominent examples of networks represented in this fashion include the World Wide Web graph (with vertices representing web-pages and directed edges representing hyper-links pointing from one page to another), protein-protein interaction networks in biology (with vertices representing proteins and undirected edges representing an affinity for two proteins to bind physically), and friendship networks (with vertices representing people and edges representing friendship nominations in a social survey).

A great deal of attention in the literature has been focused on the natural problem of modeling networks. There are by now a wide variety of network models that have been proposed, ranging from models of largely mathematical interest to models designed to be fit statistically to data. See, for example, the sources cited above or, for a shorter treatment, the review paper by Airoldi et al. (2009). The derivation and study of network models is a unique endeavor, due to a number of factors. First, the defining aspect of networks is their relational nature, and hence the task is effectively one of modeling complex dependencies among the vertices. Second, quite often there is no convenient space associated with the network, and so the type of distance and geometry that can be exploited in modeling...
other dependent phenomena, like time series and spatial processes, generally are not available when modeling networks. Finally, network problems frequently are quite large, involving hundreds if not thousands or hundreds of thousands of vertices and their edges. Since a network of \( N_v \) vertices can in principle have on the order of \( O(N_v^2) \) edges, in network modeling and analysis — particularly statistical analysis of network data — the sheer magnitude of the network can be a critical factor in this area.

Suppose that we observe a network, in the form of a directed graph \( G = (V,E) \), where \( V \) is a set of \( N_v = |V| \) vertices and \( E \) is a set of ordered vertex pairs, indicating edges. Alternatively, we may think of \( G \) in terms of its \( N_v \times N_v \) adjacency matrix \( Y \), where \( Y_{ij} = 1 \), if \((i,j) \in E\), and 0, otherwise. What is our sample size in this setting? At the August, 2010 opening workshop of the recent Program on Complex Networks, held at the Statistical and Applied Mathematical Sciences Institute (SAMSI), in North Carolina, USA, this question in fact evoked three different responses:

1. it is the number of unique entries in \( Y \), i.e., \( N_v(N_v - 1) \);
2. it is the number of vertices, i.e., \( N_v \); or
3. it is the number of networks, i.e., one.

Which answer is correct?

Despite the already vast literature on network modeling, to the best of our knowledge this question has yet to be formally posed much less answered. That this should be so is particularly curious given that the analogous questions have been asked and answered in other areas involving dependent data, most notably in time series analysis and the analysis of spatial data. Specifically, in both of the latter contexts, it is often possible to show that, whereas nominally the asymptotic variance of maximum likelihood estimates for parameters scales inversely with the sample size \( n \), under dependency a different scaling obtains, reflecting a combination of (a) the nominal sample size \( n \), and (b) the dependency structure in the data. Thus, it seems not unreasonable to hope that similar results might be produced in the context of networks, with the asymptotics shown to be a function of the number vertices \( N_v \), modified by characteristics of the network structure itself.

Following similar practice in these other fields, therefore, we will interpret the scaling of the asymptotic variances of maximum likelihood estimates in a network model as an effective sample size. In this paper we provide some initial insight
into the question of what is the effective sample size in network modeling, focusing on the impact of what is arguably the most fundamental of network characteristics — sparsity. A now commonly acknowledged characteristic of real-world networks is that the actual number of edges tends to scale much more like the number of vertices (i.e., $O(N_v)$) than the nominal number of edges (i.e., $O(N_v^2)$). Here we demonstrate that two very different regimes of asymptotics, corresponding to responses 1 and 2 above, obtain for maximum likelihood estimates in the context of a simple case of the popular exponential random graph models, under non-sparse and sparse variants of the models.

Notably, the specification of our models and the derivations of our results all utilize concepts and tools accessible to a first-year graduate student in statistics. Accordingly, our results serve to highlight in a straightforward and illustrative manner how the question of effective sample size in network settings can in fact be expected to be non-trivial and that the answer in general is likely to be subtle, depending substantially on basic model assumptions.

The rest of this paper is organized as follows. Some background and definitions are provided in Section 2. Our main results are presented in Section 3, first for the case where edges arise as independent coin flips and, second, for the case in which flips corresponding to edges to and from a given pair of vertices are dependent. We then illustrate some practical implications of our results, through a simulation study in Section 4, exploring coverage of confidence intervals associated with our asymptotic arguments, and through application to food-sharing networks in Section 5, where we examine the extent to which real-world data can be found to support non-sparse versus sparse variants of our models. Finally, some additional discussion may be found in Section 6.

2 Background

There are many models for networks. See Kolaczyk (2009), Chapter 6, or the review paper by Airoldi et al. (2009). The class of exponential random graph models has a history going back roughly 30 years and is particularly popular with practitioners in social network analysis. This class of models specifies that the distribution of the adjacency matrix $Y$ follow an exponential family form, i.e., $p_\theta(Y = y) \propto \exp(\theta^\top g(y))$, for vectors $\theta$ of parameters and $g(\cdot)$ of sufficient statistics. However, despite this seemingly appealing feature, work in the last five years has shown that exponential random graph models must be handled with some care, as both their theoretical properties and computational tractability can
be rather sensitive to model specification. See Robins et al. (2007), for example, and Chatterjee and Diaconis (2011), for a more theoretical treatment.

Here we concern ourselves only with certain examples of the simplest type of exponential random graph models, wherein the dyads \((Y_{ij}, Y_{ji})\) and \((Y_{k\ell}, Y_{\ell k})\) are assumed independent, for \((i, j) \neq (k, \ell)\), and identically distributed. These independent dyad models arguably have the smallest amount of dependency to still be interesting as network models. A variant of the models introduced by Holland and Leinhardt (1981), they are in fact too simple to be appropriate for modeling in most situations of practical interest. However, they are ideal for our purposes, as they allow us to quickly obtain non-trivial insight into the question of effective sampling size in network modeling, using relatively standard tools and arguments.

The models we consider are all variations of the form

\[
p_{\alpha, \beta}(Y = y) = \prod_{i<j} \exp \left\{ \alpha (y_{ij} + y_{ji}) + \beta y_{ij}y_{ji} \right\}
\left\{ 1 + 2e^{\alpha} + e^{2\alpha+\beta} \right\}^{\binom{N_v}{2}} \times \exp \left\{ \alpha s(y) + \beta m(y) \right\},
\]

(1)

with sufficient statistics

\[
s(y) \equiv \sum_{i<j} (y_{ij} + y_{ji}) \quad \text{and} \quad m(y) \equiv \sum_{i<j} y_{ij}y_{ji},
\]

(2)

a so-called Bernoulli model with reciprocity. The parameter \(\alpha\) governs the propensity of vertices \(i\) and \(j\) to form an edge \((i, j)\), and the parameter \(\beta\) governs the tendency towards reciprocity, forming an edge \((j, i)\) that reciprocates \((i, j)\). Of interest will be both this general model and the restricted model \(p_{\alpha} \equiv p_{\alpha, 0}\), wherein \(\beta = 0\) and there is no reciprocity. We will refer to this latter model simply as the Bernoulli model. Realizations of networks from this model without and with reciprocity (holding expected edge count \(s(y)\) fixed) are given in Figure [1(a)] and [1(b)], respectively.

Importantly, in both the Bernoulli model and the Bernoulli model with reciprocity, we will examine the question of effective sample size under both the original model parameterization and a reparameterisation in which parameter(s) are shifted by a value \(\log N_v\). Krivitsky et al. (2011) introduced such shifts as a way of adjusting models like (1) for network size such that realizations with fixed \(\alpha\) and \(\beta\) would produce network distributions with asymptotically constant expected mean degree, \(E_{\alpha, \beta}[s(Y)/N_v]\), for varying \(N_v\). That is, a configuration \((\alpha, \beta)\) that
Figure 1: Sampled networks drawn from four configurations of (1). (a) shows a realization from a model with expected mean degree 1 on 100 vertices, and no reciprocity effect. (b) shows a realization from model with the same network size and mean degree as (a) but with reciprocity parameter $\beta$ set such that the expected number of mutual ties is 25. (c) is a realization of the model from (a) scaled to 200 vertices, preserving density; while (d) preserves mean degree.
would produce a typical $N_v = 100$ realization like that in Figure 1a. Would produce an $N_v = 200$ realization like that in Figure 1d. The model’s baseline asymptotic behavior is to have a constant expected density, $E_{\alpha,\beta}[s(Y)/\{N_v(N_v - 1)\}]$, such that a parameter configuration that would produce a network like 1a for $N_v = 100$ would produce a network like 1c for $N_v = 200$. Motivated by similar concerns, we use the presence or absence of such shifts to produce two different types of asymptotic behavior in our network model classes, corresponding to sparse (asymptotically finite mean degree) and non-sparse (asymptotically infinite mean degree) networks, respectively. Because it is widely recognized that most large real-world networks are sparse networks, this distinction is critical, and, as we show below, it has fundamental implications on effective sample size.

3 Main Results

3.1 Bernoulli Model

We first present our results for the Bernoulli model. Let $p_\alpha$ denote the model $p_{\alpha,0}$, as defined above, and let $p_\alpha^\dagger$ denote the same model, but under the mapping \( \alpha \mapsto \alpha - \log N_v \) of the density parameter. Then, it is easy to show that under $p_\alpha$ the mean vertex in- and out-degree tends to infinity and the network density stays at \( \logit^{-1}(\alpha) \) as $N_v \to \infty$, while under $p_\alpha^\dagger$, the mean degree tend to $e^\alpha$ while the density tends to zero. In fact, the limiting in- and out-degree distributions tend to a Poisson law with the stated mean.

From the perspective of traditional random graph theory, the offset model of Krivitsky et al. (2011) is asymptotically equivalent to the standard formulation of an Erdős-Rényi random graph, in which the probability of an edge scales like $e^{\alpha}/N_v$. Alternatively, from the perspective of social network theory, it is useful to examine the log-odds that $Y_{ij} = 1$, conditional on the status $Y_{[-ij]} = y_{[-ij]}$ of all other edges log-odds of an edge, i.e.,

$$
\log \frac{p(Y_{ij} = 1 \mid Y_{[-ij]} = y_{[-ij]})}{p(Y_{ij} = 0 \mid Y_{[-ij]} = y_{[-ij]})}.
$$

This quantity goes from being a constant value $\alpha$ under $p = p_\alpha$ to a value $\alpha - \log N_v$ under $p_\alpha^\dagger$. This reflects the intuition that as long as there is a cost associated with forming and maintaining a network tie, an individual will be able to maintain ties with a shrinking fraction of the network as the network grows, with the
average number of maintained ties being unaffected by the growth of the network beyond a certain point (Krivitsky et al., 2011).

Given the observation of a network $Y$ randomly generated with respect to either of these models, initial insight into the effective sample size can be obtained by studying the asymptotic behavior of the Fisher information, which we denote $\mathcal{I}(\alpha)$ and $\mathcal{I}^\dagger(\alpha)$ under $p_\alpha$ and $p_\alpha^\dagger$, respectively. Straightforward calculation shows that while

$$\mathcal{I}(\alpha) = \left(\frac{N_v}{2}\right) \frac{2e^\alpha}{(1 + e^\alpha)^2},$$

in contrast,

$$\mathcal{I}^\dagger(\alpha) = \left(\frac{N_v}{2}\right) \frac{2e^\alpha/N_v}{(1 + e^\alpha/N_v)^2} \approx N_v e^\alpha.$$

So $\mathcal{I}(\alpha) = O(N_v^2)$, while $\mathcal{I}(\alpha)^\dagger = O(N_v)$, a difference by an order of magnitude.

The implications of this difference are immediately apparent when we consider the asymptotic behavior of the maximum likelihood estimates of $\alpha$ under the two models.

**Theorem 1.** Let $\hat{\alpha}$ and $\hat{\alpha}^\dagger$ denote the maximum likelihood estimates of the parameter $\alpha_0$ under the models $p_\alpha$ and $p_\alpha^\dagger$, respectively, where $\alpha_0 \in [\alpha_{\min}, \alpha_{\max}]$, for finite $\alpha_{\min}, \alpha_{\max}$. Then under the model $p_\alpha$, the estimator $\hat{\alpha}$ is $\left(\frac{N_v}{2}\right)^{1/2}$-consistent for $\alpha_0$, and

$$\left(\frac{N_v}{2}\right)^{1/2} (\hat{\alpha} - \alpha_0) \to N \left(0, \frac{2e^{\alpha_0}}{(1 + e^{\alpha_0})^2} \right)^{-1},$$

while under the model $p_\alpha^\dagger$, the estimator $\hat{\alpha}^\dagger$ is $N_v^{1/2}$-consistent for $\alpha_0$, and

$$\sqrt{N_v} (\hat{\alpha}^\dagger - \alpha_0) \to N \left(0, e^{-\alpha_0} \right).$$

The proof of these results uses largely standard techniques for asymptotics of estimating equations, but with a few interesting twists. Note that, for fixed $N_v$, the dyads $(Y_{ij}, Y_{ji})$ constitute $N_v(N_v - 1)/2$ independent and identically distributed bivariate random variables under both $p_\alpha$ and $p_\alpha^\dagger$. Consistency of the estimators in both cases can be argued by verifying, for example, the conditions of Theorem 5.9 of Van der Vaart (2000) for consistency of estimating equations. Similarly, the
proof of asymptotic normality of the estimators can be based on the usual technique of a Taylor series expansion of the log-likelihood and, due to the fact that we have assumed an exponential family distribution, the asymptotic normality of the sufficient statistic $s(y)$ in (2). However, in the case of the sparse model $p^\dagger$, the dyads $\{(Y_{ij}, Y_{ji})\}_{i<j}$ follow a different distribution for each $N_v$, and therefore an array-based central limit theorem is required to show the asymptotic normality of $s(y)$. But since increasing the number of vertices from, say, $N_v - 1$ to $N_v$, as $N_v \to \infty$, increases the number of dyads in our model by $N_v - 1$, a standard triangular array central limit theorem is not appropriate here. Rather, a double array central limit theorem is needed, such as Theorem 7.1.2 of Chung (2001). A full derivation is provided in Appendix A.2.

3.2 Bernoulli Model with Reciprocity

From Theorem 1 we see that the effective sample size in this context can be either $N_v$ or $N_v^2$, depending on the scaling of the assumed model, i.e., on whether the model is sparse or not. From a non-network perspective, these results can be largely anticipated by the rescaling involved, in that the transformation $\alpha \mapsto \alpha - \log N_v$ induces a rescaling of the expected number of edges by $N_v^{-1}$. Now, however, consider the full Bernoulli model with reciprocity, $p_{\alpha,\beta}$, defined in (1). Even with just two parameters the situation becomes notably more subtle.

Let $I(\alpha, \beta)$ be the $2 \times 2$ Fisher information matrix under this model. Then calculations analogous to those required for our previous results show that $I(\alpha, \beta) = O(N_v^2)$ and, similarly, asymptotic properties of the maximum likelihood estimate of $(\alpha, \beta)$ analogous to those for $p_{\alpha}$ hold.

Let us focus then on sparse versions of $p_{\alpha,\beta}$. The offset used previously, i.e., mapping $\alpha$ to $\alpha - \log N_v$, is not by itself satisfactory. Call the resulting model $p^\dagger_{\alpha,\beta}$. Standard arguments show that the limiting in- and out-degree distributions under this model will be Poisson with mean parameter $e^\alpha$. On the other hand, the expected number of reciprocated out-ties a vertex has, $E^\dagger_{\alpha,\beta}[2m(Y)/N_v]$, behaves like $e^{2\alpha+\beta}/N_v$, and therefore tends to zero as $N_v \to \infty$. Thus, $\beta$ plays no role in the limiting behavior of the model, and, indeed, reciprocity vanishes. This fact can also be understood through examination of the Fisher information matrix, say $I^\dagger(\alpha, \beta)$, in that direct calculation shows

$$I^\dagger(\alpha, \beta) = \begin{bmatrix} O(N_v) & O(1) \\ O(1) & O(1) \end{bmatrix}.$$
That is, only the information on $\alpha$ grows with the network. Under $p^{\dagger}_{\alpha,\beta}$, only the affinity parameter $\alpha$ can be inferred in a reliable manner.

However, the same intuition that suggests that as the network becomes larger, a given actor $i$ will have an opportunity for contact with a smaller and smaller fraction of it also suggests that if there is a preexisting relationship in the form of a tie from $j$ to $i$, such an opportunity likely exists regardless of how large the network may be. This, as well as direct examination of the exact expression for the information matrix $\mathcal{I}^{\dagger}(\alpha, \beta)$, suggests that the $-\log N_v$ penalty on tie log-probability should not apply to reciprocating ties, which may be implemented by mapping $\beta \mapsto \beta + \log N_v$. Call this model, in which $p^{\dagger}_{\alpha,\beta}$ is augmented with this additional offset for $\beta$, the model $p^{\ddagger}_{\alpha,\beta}$. The corresponding conditional log-odds of an edge now have the form

$$
\log \frac{p^{\ddagger}_{\alpha,\beta}(Y_{ij} = 1 | Y_{[-ij]} = y_{[-ij]})}{p^{\ddagger}_{\alpha,\beta}(Y_{ij} = 0 | Y_{[-ij]} = y_{[-ij]})} = \begin{cases} 
\alpha - \log N_v, & \text{if } y_{ji} = 0, \\
\alpha + \beta, & \text{if } y_{ji} = 1,
\end{cases}
$$

which exactly captures the intuition described.

It can be shown that under $p^{\ddagger}_{\alpha,\beta}$ we have $\mathcal{I}^{\ddagger}(\alpha, \beta) = O(N_v)$, indicating that information on both parameters grows at the same rate in $N_v$. It can also be shown that the limiting in- and out-degree distribution is now Poisson with mean parameter $e^{\alpha} + e^{2\alpha+\beta}$, and that $E^{\ddagger}_{\alpha,\beta}[2m(Y)/N_v]$ tends to $e^{2\alpha+\beta}$. So, both parameters play a role in the limiting behavior of the model and the additional offset induces an asymptotically constant expected per-vertex reciprocity in addition to asymptotically constant expected mean degree.

Finally, we have the following analogue of Theorem 1.

**Theorem 2.** Let $(\hat{\alpha}^{\ddagger}, \hat{\beta}^{\ddagger})$ denote the maximum likelihood estimate of the parameter $(\alpha_0, \beta_0)$ under the model $p^{\ddagger}_{\alpha_0,\beta_0}$, where $(\alpha_0, \beta_0) \in [\alpha_{\min}, \alpha_{\max}] \times [\beta_{\min}, \beta_{\max}]$, for finite $\alpha_{\min}, \alpha_{\max}, \beta_{\min}, \beta_{\max}$. Then $(\hat{\alpha}^{\ddagger}, \hat{\beta}^{\ddagger})$ is $N_v^{1/2}$-consistent for $(\alpha_0, \beta_0)$, and

$$
\sqrt{N_v} \begin{pmatrix} \hat{\alpha}^{\ddagger} - \alpha_0 \\ \hat{\beta}^{\ddagger} - \beta_0 \end{pmatrix} \to N \begin{pmatrix} 0, & e^{-\alpha_0} \begin{bmatrix} 1 & -2 \\ -2 & 4 + 2 e^{-\alpha_0-\beta_0} \end{bmatrix} \end{pmatrix}.
$$

Proof of this theorem, using arguments directly analogous to those of Theorem 1 may be found in the Appendix A.3. From the theorem we see that under the sparse model $p^{\ddagger}_{\alpha,\beta}$, as under $p^{\dagger}_{\alpha}$, the effective sample size is $N_v$. 

10
Coverage of Wald Confidence Intervals

Our asymptotic arguments in Section 3 were developed primarily for the purpose of establishing the scaling associated with the asymptotic variance, so as to provide insight into the question of effective sample size — our main focus here. However, the asymptotically normal distributions we have derived are of no little independent interest themselves, as they serve as a foundation for doing formal inference on the model parameters in practice. By way of illustration, here we explore their use for constructing confidence intervals, particularly those based on Theorem 2: under a model \( p^\dagger_{\alpha,\beta} \), the Wald confidence intervals using plug-in estimators for the standard errors are

\[
\hat{\alpha}^\dagger \pm z^*_{(1-CL)/2} \sqrt{e^{-\hat{\alpha}^\dagger}/N_v} \text{ for } \alpha
\]

and

\[
\hat{\beta}^\dagger \pm z^*_{(1-CL)/2} \sqrt{e^{-\hat{\alpha}^\dagger}(4 + 2e^{-\hat{\alpha}^\dagger - \hat{\beta}^\dagger})/N_v} \text{ for } \beta.
\]

Because our asymptotics are in \( N_v = |V| \), we examine a variety of network sizes. The desired asymptotic properties of the network are expressed in terms of the per-capita mean value parameters — \( E_{\alpha,\beta}[s(Y)/N_v] \) and \( E_{\alpha,\beta}[2m(Y)/N_v] \). We study two configurations:

1. \( (E_{\alpha,\beta}[s(Y)/N_v], E_{\alpha,\beta}[2m(Y)/N_v]) = (1, 0.25) \)
2. \( (E_{\alpha,\beta}[s(Y)/N_v], E_{\alpha,\beta}[2m(Y)/N_v]) = (1, 0.40) \).

In other words, the expected mean outdegree is set to 1, and expected numbers of out-ties that are reciprocated are \( 0.25 \times 2 = 0.5 \) and \( 0.40 \times 2 = 0.8 \) per actor, respectively. These represent two levels of mutuality, though note that even (1) represents substantial mutuality, especially for larger networks.

For each \( N_v = 10, 15, 20, \ldots, 200 \), we estimate the natural parameters of the model \( p^\dagger_{\alpha,\beta} \) corresponding to the desired mean value parameters, and then simulate 100,000 networks from each configuration, evaluating the MLE and constructing a Wald confidence interval at each of level of the customary 80%, 90%, 95%, and 99%, for \( \alpha \) and for \( \beta \) (individually), checking the coverage.

For some the smaller sample sizes, the simulated network statistics were not in the interior of their convex hull \( (0 < s(y) < N_v(N_v-1) \text{ and } 0 < m(y) < s(y)/2) \), so the MLE did not exist. (For (1), the fraction was 8.2% for \( N_v = 10 \) and none of the 100,000 realizations had no MLE for \( N_v \geq 55 \). For (2), the it was 14.2% for \( N_v = 10 \) and none of the realizations had no MLE for \( N_v \geq 65 \).)

Our results are conditional on the MLE existing. From the frequentist perspective, one might argue that if the MLE did not exist for a real dataset, we would not
Table 1: Simulated Theorem 2 confidence interval coverage levels for selected network sizes and two levels of reciprocity: lower (1) and higher (2).

| Coverage  | 80.0%  | 90.0%  | 95.0%  | 99.0%  |
|-----------|--------|--------|--------|--------|
| \( N_v \) | \( \alpha \) | \( \beta \) | \( \alpha \) | \( \beta \) | \( \alpha \) | \( \beta \) | \( \alpha \) | \( \beta \) |
| (1) 10    | 72.4%  | 77.3%  | 85.3%  | 93.2%  | 96.4%  | 99.4%  | 99.4%  | 99.4%  |
| 20       | 74.5%  | 77.3%  | 86.0%  | 92.9%  | 98.3%  | 99.5%  | 99.5%  | 99.5%  |
| 50       | 80.9%  | 78.8%  | 87.6%  | 94.7%  | 98.9%  | 99.2%  | 99.2%  | 99.2%  |
| 100      | 77.4%  | 79.6%  | 90.0%  | 94.6%  | 98.9%  | 99.1%  | 99.1%  | 99.1%  |
| 200      | 79.0%  | 79.5%  | 90.1%  | 94.9%  | 98.9%  | 99.0%  | 99.0%  | 99.0%  |
| (2) 10    | 84.0%  | 84.2%  | 86.6%  | 93.6%  | 96.3%  | 98.2%  | 98.2%  | 98.2%  |
| 20       | 81.8%  | 80.3%  | 92.8%  | 95.1%  | 98.1%  | 98.8%  | 98.8%  | 98.8%  |
| 50       | 75.3%  | 79.5%  | 91.7%  | 95.6%  | 98.8%  | 99.0%  | 99.0%  | 99.0%  |
| 100      | 78.5%  | 79.7%  | 91.0%  | 94.5%  | 99.0%  | 99.1%  | 99.1%  | 99.1%  |
| 200      | 82.2%  | 79.9%  | 90.5%  | 95.3%  | 99.2%  | 99.1%  | 99.1%  | 99.1%  |

have reported that type of confidence interval, so it should be excluded from the simulation as well.

We report coverages for selected network sizes in Table 1 and provide a visualization in Figure 2. Overall, the 80% coverage appears to be varied — and not very conservative — while higher levels of confidence appear to be more consistently conservative, particularly for estimates of \( \beta \). Coverage for \( \alpha \) appears to oscillate as a function of network size. This is particularly noticeable for the lower confidence levels and stronger mutuality (2). Tendency of a confidence interval for a binomial proportion to oscillate around the nominal level is a known phenomenon (Brown et al., 2001, 2002, and others), though it is interesting to note that it appears to be more prominent for the density, rather than mutuality, parameter and that it appears to be stronger for stronger mutuality.

5 Example: Food-Sharing Networks in Lamalera

While the results of Section 3 are important in establishing how closely the question of effective sample size in network modeling is tied to the structural property of (non)sparseness expected of the networks modeled, there remains the important
Figure 2: Differences between simulated coverage and nominal coverage for the two configuration studied, as a function of network size $N_v$. Note that the differences are differences in percentage points (simulated $\% -$ nominal $\%$), not percent differences ($\frac{\text{simulated} \% - \text{nominal} \%}{\text{nominal} \%} \times 100\%$).
practical question of establishing in applications just which model (i.e., sparse or non-sparse) is most appropriate. While a full and detailed study of this question is beyond the scope of this work, we present here an initial exploration.

Note that, in exploring this question, we face a problem similar to that pointed out by Krivitsky et al. (2011): it requires a collection of closed networks of a variety of sizes yet substantively similar social structure. Furthermore, our results are limited to modeling density and reciprocity, so the networks should be well-approximated by this model. Here, we use data collected by Nolin (2010), in which each of 317 households in Lamalera, Indonesia was asked to list the households to whom they have given and households from whom they have received food in the preceding season. Lamalera is split, administratively, into two villages, which are further subdivided into wards, and then into neighborhoods. Nolin (2010) fit several ERGMs to the network, finding that distance between households had a significant effect on the propensity to share, as did kinship between members of the households involved. Nolin also found a significant positive mutuality effect.

In our study, we make use of the geographic effect by constructing a series of 24 overlapping subnetworks, consisting of Lamalera itself, its 2 constituent villages, 6 wards, and 15 neighborhoods, with network sizes ranging from 12 to 317. We then fit the baseline model $p_{\alpha,\beta}$ to each network. If $p_{\alpha,\beta}$ is the most realistic asymptotic regime for these data, we would expect estimates $\hat{\alpha}$ and $\hat{\beta}$ to have no relationship to $\log N_v$ for the corresponding network. If $p_{\alpha,\beta}^\dagger$ is the most realistic, we would expect no relationship between $\log N_v$ and $\hat{\beta}$, but an approximately linear relationship with $\hat{\alpha}$, with slope around $-1$. Lastly, if $p_{\alpha,\beta}^{\ddagger}$ is the most realistic, we would expect the slope of the relationship between $\log N_v$ and $\hat{\alpha}$ to be around $-1$ and between $\log N_v$ and $\hat{\beta}$ to be around $+1$.

The estimated coefficients and the slopes are given in Figure 3. The results are suggestive. The relationship between $\hat{\alpha}$ and $\log N_v$ is clearly negative, while the relationship between $\hat{\beta}$ and $\log N_v$ is clearly positive, and the magnitudes of both slopes are closer to 1 than to 0 (although both are far from equaling 1). Overlap between the subnetworks induces dependence among the coefficients, so it is not possible to formally test or estimate how significant this difference is. Nevertheless, the preponderance of evidence is that $p_{\alpha,\beta}^{\ddagger}$ is the best of the three considered. That is, a sparse model that does not enforce sparsity on reciprocating ties appears to be preferable here.

A possible explanation for why the magnitudes of the slopes are substantially less than 1 is that both the argument of Krivitsky et al. (2011) and our argument in
Figure 3: Maximum likelihood estimates from fitting $p_{\alpha,\beta}$ to each subdivision of the Lamalera food-sharing network. Colors indicate subdivision type. The least-squares coefficients from regressing $\hat{\alpha}$ and $\hat{\beta}$ on $\log N_v$ are $-0.72$ and $+0.60$, respectively.
Section 3.2 rely on the assumption that the network is closed, or, at least, that the stable mean degree and per-capita reciprocity are for the ties within the network observed. However, while there is likely to be very little food sharing out of or into Lamalera, and relatively little between the two villages it comprises (7% of all food-sharing ties in the network are between villages), there is more sharing between the wards (28% are between wards), and even more between neighborhoods (44%). Thus, the closed-network assumption is violated. (The respective between-subdivision percentages for reciprocated ties are 6%, 22%, and 39%.) When each of the subdivisions of the network is considered in isolation, these ties are lost, so the smaller subdivisions appear, to the model, to have smaller mean degree and per-capita mutuality. (See Figure 4.) This, in turn, means that smaller subdivisions have a decreased $\hat{\alpha}$ (increasing the slope for it in Fig. 3) and, because mutual ties suffer less of this “attrition” than ties do overall, the $\hat{\beta}$, after adjusting for the decreased $\hat{\alpha}$, is increased for smaller networks, thus reducing the slope for $\hat{\beta}$ in Fig. 3. It is not unlikely that this pattern will hold in any network with an unobserved spatial structure, whose subnetworks of interest are contiguous regions in this space.

### 6 Discussion

Unlike conventional data, network data typically do not have an unambiguous notion of sample size. The examples we have presented show that the effective sample size associated with a network depends strongly on the model assumed for how the network scales. In particular, in the case of reciprocity, whether or not the model for scaling takes into account the notion of preexisting relationship affects whether reciprocity is even meaningful for large networks.

Our model is, intentionally, a very simple one. However, with reciprocity, it includes an important aspect that already allows us a glimpse beyond the more sophisticated treatments of, say, Chatterjee et al. (2011) and Rinaldo et al. (2011), for so-called beta models, where the dependency induced here by reciprocity is absent. In addition, the results for reciprocity suggest that the effective modeling of triadic (e.g., friend of a friend of a friend) effects — arguably the most natural type of dependency to add next to the current model — in a manner indexed to network size is likely to require a more complex treatment yet, which, in turn, may further complicate the notion of effective sample size. While it is likely that insight into how to proceed from here can be gleaned from experience in other parts of the literature for dependent data, such as for time series and spatial data,
Figure 4: Per-capita network statistics as a function of $N_v$. Colors indicate subdivision type. Note that the larger subdivisions have more within-subdivision ties.
it is unlikely that the tools developed there can be applied directly. Rather, new tools and techniques almost certainly are required here.

We note that asymptotic theory supporting methods for the construction of confidence intervals for network parameters is only beginning to emerge. The most traction appears to have been gained in the context of stochastic block models (e.g., Bickel and Chen (2009); Choi et al. (2010); Celisse et al. (2011); Rohe et al. (2011)), although progress is beginning to be had with exponential random graph models as well (e.g., Chatterjee et al. (2011); Chatterjee and Diaconis (2011); Rinaldo et al. (2011)). Most of these works present consistency results for maximum likelihood and related estimators, with the exception of Bickel and Chen (2009), which also includes results on asymptotic normality of estimators. Our work contributes to this important but nascent area.

The lack of an established understanding of the distributional properties of parameter estimates in commonly used network models is particularly unfortunate given that a number of software packages now allow for the easy computation of such estimates. For example, packages for computing estimates of parameters in fairly general formulations of exponential random graph models routinely report both estimates and, ostensibly, standard errors, where the latter are based on standard arguments for exponential families. Unfortunately, practitioners do not always seem to be aware that the use of these standard errors for constructing normal-theory confidence intervals and tests is lacking in any formal justification. From that perspective, our work appears to be one of the first to begin laying the necessary theoretical foundation to justify practical confidence interval procedures in exponential random graph models. See Haberman (1981) for another contribution in this direction, proposed as part of the discussion of the original paper Holland and Leinhardt (1981).

References

E.M. Airoldi, D.M. Blei, S.E. Fienberg, and E.P. Xing. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2:129 – 233, 2009.

P.J. Bickel and A. Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068, 2009.

Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. Interval estimation for
a binomial proportion. *Statistical Science*, 16(2):101–117, May 2001. ISSN 0883-4237.

Lawrence D. Brown, T. Tony Cai, and Anirban Dasgupta. Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics*, 30(1):160–201, 2002.

A. Celisse, J.J. Daudin, and L. Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Arxiv preprint arXiv:1105.3288*, 2011.

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

S. Chatterjee, P. Diaconis, and A. Sly. Random graphs with a given degree sequence. *The Annals of Applied Probability*, 21(4):1400–1435, 2011.

D.S. Choi, P.J. Wolfe, and E.M. Airoldi. Stochastic blockmodels with growing number of classes. *Arxiv preprint arXiv:1011.4644*, 2010.

K.L. Chung. *A course in probability theory*. Academic Pr, 2001.

S.J. Haberman. An exponential family of probability distributions for directed graphs: Comment. *Journal of the American Statistical Association*, 76(373):60–61, 1981.

P.W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, pages 33–50, 1981.

M.O. Jackson. *Social and economic networks*. Princeton Univ Pr, 2008.

E.D. Kolaczyk. *Statistical analysis of network data: methods and models*. Springer Verlag, 2009.

P.N. Krivitsky, M.S. Handcock, and M. Morris. Adjusting for network size and composition effects in exponential-family random graph models. *Statistical Methodology*, 2011.

Peter McCullagh and John A. Nelder. *Generalized Linear Models*, volume 37 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, second edition, August 1989. ISBN 0-412-31760-5.
A. Derivations of Key Results

These supplementary materials contain the derivations of key results and expressions provided in the main paper. We begin by establishing preliminary notation and expressions in Section A.1. In Sections A.2 and A.3, we offer proofs of Theorems 1 and 2, respectively.

A.1 Preliminaries

Recall that the models we consider are all variations of the form (1) with sufficient statistics (2), and \((\alpha, \beta) \in [\alpha_{\text{min}}, \alpha_{\text{max}}] \times [\beta_{\text{min}}, \beta_{\text{max}}]\), with \(\alpha_{\text{min}}, \alpha_{\text{max}}, \beta_{\text{min}}, \) and \(\beta_{\text{max}}\) all finite. In particular, all models are of exponential family form — either with or without an offset term(s), in the terminology of McCullagh and Nelder (1989). Hence, for all models the probability mass function can be written as \(p_\theta(y) = (1/\kappa(\theta)) \exp(\theta^T g(y))\), and the log-likelihood, as \(\ell(\theta) = \theta^T g(y) - \psi(\theta)\), where \(\kappa(\theta)\) is the normalization term and \(\psi(\theta) = \log \kappa(\theta)\). Furthermore, the Fisher information matrix in each case is given by the formula \(I(\theta) = \partial^2 \psi(\theta) / \partial \theta \partial \theta^T\).
Let $\ell_{N_v}(\alpha)$ denote the log-likelihood under the Bernoulli model $p_\alpha$; let $\ell^\dagger_{N_v}(\alpha)$ denote the loglikelihood under the Bernoulli model $p^\dagger_\alpha$, with offset $\alpha \mapsto \alpha - \log N_v$; let $\ell_{N_v}(\alpha, \beta)$ denote the loglikelihood under the Bernoulli model with reciprocity $p_{\alpha,\beta}$; and let $\ell^\ddagger_{N_v}(\alpha, \beta)$ denote the loglikelihood under the Bernoulli model with reciprocity $p^{\ddagger}_{\alpha,\beta}$, with offsets $\alpha \mapsto \alpha - \log N_v$ and $\beta \mapsto \beta + \log N_v$. All expressions provided in the main paper for the orders of magnitude of the elements of the corresponding information matrices, under these various models, may be obtained directly by twice differentiating

$$
\psi(\alpha, \beta) = \left(\frac{N_v}{2}\right) \log \left(1 + 2e^\alpha + e^{2\alpha + \beta}\right),
$$

appropriately parameterized.

Thus, for example, it is straightforward to show that

$$
\mathcal{I}_{N_v}(\alpha) = N_v(N_v - 1) \left[ \frac{e^\alpha}{1 + e^\alpha} \left( 1 - \frac{e^\alpha}{1 + e^\alpha} \right) \right] = O(N_v^2),
$$

while

$$
\mathcal{I}^\dagger_{N_v}(\alpha) = N_v(N_v - 1) \left[ \frac{e^\alpha/N_v}{1 + e^\alpha/N_v} \left( 1 - \frac{e^\alpha/N_v}{1 + e^\alpha/N_v} \right) \right] = O(N_v).
$$

Similarly, for the two-parameter models, defining $A = e^\alpha$ and $B = e^{2\alpha + \beta}$, we find that

$$
\mathcal{I}_{N_v}(\alpha, \beta) = \frac{\binom{N_v}{2}}{(1 + 2A + B)^2} \begin{bmatrix}
2A + 4B + 2AB & 2(B + AB) \\
2(B + AB) & 2AB + B
\end{bmatrix},
$$

and so $\mathcal{I}_{N_v}(\alpha, \beta) = O(N_v^2)$. On the other hand, substituting $A/N_v$ and $B/N_v$ for $A$ and $B$, respectively, which captures the effects of the two offsets in the $p^{\ddagger}_{\alpha,\beta}$ model, and simplifying, we find that $\mathcal{I}^\ddagger_{N_v}(\alpha, \beta)$ behaves asymptotically like

$$
N_v \begin{bmatrix}
A + 2B & B \\
B & B/2
\end{bmatrix},
$$

and hence is $O(N_v)$. 

21
A.2 Proof of Theorem 1

Theorem 1 establishes consistency and asymptotic normality for \( \hat{\alpha} \) and \( \hat{\alpha}^\dagger \), the maximum likelihood estimates of \( \alpha_0 \) under the models \( p_{\alpha_0} \) and \( p_{\alpha_0}^\dagger \), respectively. Here we sketch the case of \( \hat{\alpha}^\dagger \), as the more interesting of the two. The case of \( \hat{\alpha} \) follows using conventional arguments.

There are various ways one might argue consistency of \( \hat{\alpha}^\dagger \). One approach would be to use techniques for M-estimators, the main requirement for which is that the log-likelihood converge uniformly in probability to a function with a well-defined maximum at \( \alpha_0 \). However, note that

\[
\frac{1}{N_v} \ell_{N_v}^\dagger (\alpha) = (N_v - 1) \left[ (\alpha - \log N_v) \bar{s}(y) - \log \left( 1 + \frac{e^\alpha}{N_v} \right) \right],
\]

where \( \bar{s}(y) = s(y) / \left[ N_v (N_v - 1) \right] \). Since

\[
(N_v - 1) \bar{s}(y) \to e^{\alpha_0}
\]

in probability, as \( N_v \to \infty \), and \( \log (1 + e^\alpha / N_v) \) behaves like \( e^\alpha / N_v \) for large \( N_v \), it follows that the log-likelihood \( \ell_{N_v}^\dagger (\alpha) / N_v \) behaves like \( (\alpha - \log N_v) e^{\alpha_0} - e^\alpha \) for large \( N_v \), and hence tends to \( -\infty \) for all \( \alpha \). As a result, while the maximization of the log-likelihood is well-defined for each finite \( N_v \), this method of proof is not amenable to demonstrating consistency here.

Instead, therefore, we study the behavior of the derivative of the log-likelihood with respect to \( \alpha \), i.e.,

\[
\Psi_{N_v}^\dagger (\alpha) = \frac{1}{N_v} \frac{d}{d\alpha} \ell_{N_v}^\dagger (\alpha)
= (N_v - 1) \left[ \bar{s}(y) - \frac{e^\alpha}{N_v + e^\alpha} \right],
\]

(4)

since \( \Psi_{N_v}^\dagger (\alpha) \to \Psi^\dagger (\alpha) \) in probability for each \( \alpha \in \Theta \), where \( \Psi^\dagger (\alpha) \equiv e^{\alpha_0} - e^\alpha \). By Theorem 5.9 of [Van der Vaart (2000)], since \( \Psi^\dagger (\alpha) \) has a unique zero at \( \alpha = \alpha_0 \), in order for us to demonstrate consistency it remains to show that \( \Psi_{N_v}^\dagger (\alpha) \) converges uniformly to \( \Psi^\dagger (\alpha) \) on \( \Theta \), i.e., that

\[
\sup_{\alpha \in \Theta} \left| \Psi_{N_v}^\dagger (\alpha) - \Psi^\dagger (\alpha) \right| \to 0
\]

in probability as \( N_v \to \infty \).
We have assumed that $\alpha_0 \in [\alpha_{\text{min}}, \alpha_{\text{max}}]$, a compact and connected set, and we have observed pointwise convergence of $\Psi_{N_v}^\dagger$ to $\Psi^\dagger$. Uniform convergence therefore follows if we can demonstrate stochastic equicontinuity of

$$Q_{N_v}(\alpha) = \Psi_{N_v}^\dagger(\alpha) - \Psi^\dagger(\alpha)$$

$$= (N_v - 1) \left[ \bar{s}(y) - \frac{e^\alpha}{N_v + e^\alpha} \right] - (e^{\alpha_0} - e^\alpha).$$

But a sufficient condition for stochastic equicontinuity is that $Q_{N_v}$ be Lipschitz. See Corollary 2.2 to Theorem 2.1 of Newey (1991), for example. That this last condition is true, however, can be argued easily enough, as $Q_{N_v}$ is a continuous function in $\alpha$ on a compact, connected domain, which by the mean value theorem allows us to write

$$|Q_{N_v}(\alpha) - Q_{N_v}(\alpha')| \leq K |\alpha - \alpha'|,$$

where $K = \sup_{\alpha \in \Theta} |Q_{N_v}(\alpha)|$ and $Q_{N_v}$. Hence $Q_{N_v}$ is Lipschitz and $\hat{\alpha}^\dagger$ has been shown to be consistent.

To establish asymptotic normality, we use a standard argument based on Taylor series expansions. Begin by writing

$$0 = \Psi_{N_v}^\dagger(\hat{\alpha}^\dagger) = \Psi_{N_v}^\dagger(\alpha_0) + (\hat{\alpha}^\dagger - \alpha_0)\dot{\Psi}_{N_v}^\dagger(\alpha_0) + \frac{1}{2}(\hat{\alpha}^\dagger - \alpha_0)^2 \ddot{\Psi}_{N_v}^\dagger(\tilde{\alpha}),$$

where $\tilde{\alpha}$ is a value between $\hat{\alpha}^\dagger$ and $\alpha_0$. (Here again we employ the dot notation for differentiation with respect to $\alpha$.) It follows that

$$\sqrt{N_v}(\hat{\alpha}^\dagger - \alpha_0) = \frac{-\sqrt{N_v}\Psi_{N_v}^\dagger(\alpha_0)}{\dot{\Psi}_{N_v}^\dagger(\alpha_0) + \frac{1}{2}(\hat{\alpha}^\dagger - \alpha_0)^2 \ddot{\Psi}_{N_v}^\dagger(\tilde{\alpha})}. \quad (5)$$

Consider the numerator in (5). Recalling the form of (4), we see that $\Psi_{N_v}^\dagger(\alpha_0)$ is simply proportional to an average of the $N_v(N_v - 1)$ independent and identically distributed link variables $y_{ij}$ (i.e., $\bar{s}(y)$) centered by its mean. However, two points are worth noting. First, this mean is changing as a function of $N_v$. Second, as $N_v$ is allowed to tend towards infinity, for an increase of $N_v$ to $N_v + 1$ there are a total of $N_v$ more observations $y_{ij}$ that define the average $\bar{s}(y)$. Therefore, we employ a double array central limit theorem in arguing for the asymptotic normality of a suitably normalized version of $\Psi_{N_v}^\dagger$.
Specifically, write

\[ \mu_{N_v}(\alpha_0) = E[\bar{s}(Y)] = \frac{e^{\alpha_0}}{N_v + e^{\alpha_0}} \]

and

\[ v_{N_v}(\alpha_0) = \text{Var}(\bar{s}(Y)) = \frac{1}{N_v - 1} \frac{e^{\alpha_0}}{(N_v + e^{\alpha_0})^2}, \]

and consider the standardized variables

\[ X_{ij} = \frac{Y_{ij} - \mu_{N_v}(\alpha_0)}{\nu_{N_v}^{1/2}(\alpha_0)}. \]

It is easy to see that, for each fixed \( N_v \), the \( N_v(N_v - 1) \) random variables \( X_{ij} \) are all bounded by a constant, say \( M_{N_v} \), and that \( \lim_{N_v \to \infty} M_{N_v} = 0 \). (Indeed, \( M_{N_v} = O(N_v^{-1/2}) \).) Hence the condition of the corollary immediately following Theorem 7.1.2 of [Chung (2001)] holds, and therefore the sum of the \( X_{ij} \)'s, which is the standardized mean

\[ \bar{s}(Y) - \mu_{N_v}(\alpha_0) \nu_{N_v}^{1/2}(\alpha_0), \]

tends in distribution to a standard normal. Noting then that this sum behaves asymptotically as \( N_v^{1/2} \hat{\Psi}_{N_v}^\dagger(\alpha_0) e^{-\alpha_0/2} \), we conclude that the numerator in (5) tends to a zero-mean normal random variable with variance \( e^{\alpha_0} \).

Now consider the denominator in (5). Direct calculation yields that

\[ \hat{\Psi}_{N_v}^\dagger(\alpha_0) = -(N_v - 1) \frac{e^{\alpha_0}/N_v}{(1 + e^{\alpha_0}/N_v)^2} \to -e^{\alpha_0}. \]

In addition, we know that \( \hat{\alpha} - \alpha_0 = o_P(1) \), while an exercise in calculus shows that \( \hat{\Psi}_{N_v}^\dagger(\hat{\alpha}) = O_P(1) \). As a result, the denominator in (5) tends to \( -e^{\alpha_0} \) in probability.

Combining these results, we conclude that \( N_v^{1/2} (\hat{\alpha} - \alpha_0) \) tends in distribution to a mean-zero normal random variable with variance \( e^{-\alpha_0} \), as was to be shown.

### A.3 Proof of Theorem 2

Theorem 2 establishes consistency and asymptotic normality for \( (\hat{\alpha}^\dagger, \hat{\beta}^\dagger) \), the maximum likelihood estimate of \( (\alpha_0, \beta_0) \) under the model \( p^\dagger_{\alpha_0, \beta_0} \). Our proof uses arguments analogous to those of Theorem [1].
We again use Theorem 5.9 of Van der Vaart (2000) to show consistency of our estimator. Analogous to (4), the vector of partial derivatives of the log-likelihood function \( \ell^\dagger_N(\alpha, \beta)/N_v \), with respect to \( \alpha \) and \( \beta \), has the form

\[
\Psi^\dagger_N(\alpha, \beta) = \frac{N_v - 1}{2} \left[ \bar{s}(y) \frac{1}{\bar{m}(y)} - 2 \left( e^{\alpha_0 + e^{2\alpha_0 + \beta_0}} \right) \right],
\]

where \( \bar{s}(y) = s(y)/(nverts) \) and \( \bar{m}(y) = m(y)/(N_v^2) \). Note that

\[
(N_v - 1) \left[ \begin{array}{c} \bar{s}(y) \\ \bar{m}(y) \end{array} \right] \rightarrow \left[ \begin{array}{c} 2 \left( e^{\alpha_0 + e^{2\alpha_0 + \beta_0}} \right) \\ e^{2\alpha_0 + \beta_0} - e^{2\alpha_0 + \beta_0} \end{array} \right]
\]

in probability as \( N_v \rightarrow \infty \). Therefore, \( \Psi^\dagger_N(\alpha, \beta) \rightarrow \Psi^\dagger(\alpha, \beta) \) in probability pointwise, where

\[
\Psi^\dagger(\alpha, \beta) = \frac{1}{2} \left[ 2 \left( e^{\alpha_0 + e^{2\alpha_0 + \beta_0}} \right) - 2 \left( e^{\alpha_0 + e^{2\alpha_0 + \beta_0}} \right) \right].
\]

It is straightforward to show that \( \Psi^\dagger(\alpha, \beta) = 0 \) has the unique solution \((\alpha, \beta) = (\alpha_0, \beta_0)\). So consistency follows from arguing that the convergence of \( \Psi^\dagger_N \) to \( \Psi^\dagger \) is uniform on the set \([\alpha_{\min}, \alpha_{\max}] \times [\beta_{\min}, \beta_{\max}]\). Following the same line of reasoning used above in the proof of Theorem 1, it suffices to show that \( Q_N = \Psi^\dagger_N - \Psi^\dagger \) is Lipschitz on this set. But this follows immediately from the facts that the gradient of \( Q_N \) is continuous on this set and the set is compact and connected, followed by an appeal to the multivariate version of the mean-value theorem. As a result, consistency follows.

To argue for asymptotic normality, we begin again with a Taylor series expansion, which allows us to write

\[
-N_v^{1/2} \Psi^\dagger_N(\alpha_0, \beta_0) = N_v^{1/2} \Psi^\dagger_N(\alpha_0, \beta_0) \left( \begin{array}{c} \hat{\alpha}^\dagger - \alpha_0 \\ \hat{\beta}^\dagger - \beta_0 \end{array} \right) + \frac{N_v^{1/2}}{2} \left( \begin{array}{c} \hat{\alpha}^\dagger - \alpha_0 \\ \hat{\beta}^\dagger - \beta_0 \end{array} \right)^\top \Psi^\dagger_{N_v}(\hat{\alpha}, \hat{\beta}) \left( \begin{array}{c} \hat{\alpha}^\dagger - \alpha_0 \\ \hat{\beta}^\dagger - \beta_0 \end{array} \right).
\]

We then argue that (a) on the left-hand side of this expression the term converges asymptotically to a multivariate normal distribution, while (b) on the right-hand side, the multiplier in the first-order term converges to a constant, and the second-order term is asymptotically negligible.
In order to demonstrate asymptotic normality on the left-hand side of (7), we employ the Cramér-Wold device. From (6) we see that the behavior of $Ψ_N$ is driven by that of the mean vector $(\bar{s}(y), \bar{m}(y))^\top$. Letting $A_0 = e^{\alpha_0}$ and $B_0 = e^{2\alpha_0+\beta_0}$, write
\[
\mu_N(\alpha_0, \beta_0) = E\left[\begin{pmatrix} \bar{s}(Y) \\ \bar{m}(Y) \end{pmatrix}\right] = (N_v + 2A_0 + B_0)^{-1} \begin{pmatrix} 2(A_0 + B_0) \\ B_0 \end{pmatrix}
\]
and
\[
V_N(\alpha_0, \beta_0) = \text{Cov}\left(\begin{pmatrix} \bar{s}(Y) \\ \bar{m}(Y) \end{pmatrix}\right) = \left[\begin{pmatrix} N_v \\ 2 \end{pmatrix} N_v (1 + 2A_0/N_v + B_0/N_v^2)^2\right]^{-1} \times \begin{pmatrix} 2A_0 + 4B_0 + 2A_0B_0/N_v \\ 2B_0 + 2A_0B_0/N_v \end{pmatrix}.
\]

Fix $t = (t_1, t_2)^\top$ and consider the standardized variables
\[
X_{ij}^{(t)} = \frac{t_1(Y_{ij} + Y_{ji}) + t_2Y_{ij}Y_{ji} - t^\top \mu_N(\alpha_0, \beta_0)}{(N_v/2)^{1/2} \sqrt{t^\top V_N(\alpha_0, \beta_0)t}}.
\]
Again using Theorem 7.1.2 of Chung (2001), we have that the sum of the $X_{ij}^{(t)}$'s, and hence
\[
t_1\bar{s}(Y) + t_2\bar{m}(Y) - t^\top \mu_N(\alpha_0, \beta_0)
\]
tends in distribution to a standard normal. But this holds true for all $t$, and therefore
\[
V_N^{-1/2}(\alpha_0, \beta_0) \begin{pmatrix} \bar{s}(Y) \\ \bar{m}(Y) \end{pmatrix} - \mu_N(\alpha_0, \beta_0)
\]
tends asymptotically to a bivariate normal with mean zero and covariance the identity.

Some algebra shows that the statistic in (8) behaves asymptotically like
\[
N_v^{1/2} V^{-1/2}(\alpha_0, \beta_0) \Psi_N^{\#}(\alpha_0, \beta_0),
\]
26
where
\[ V(\alpha_0, \beta_0) = \begin{bmatrix} A_0 + 2B_0 & B_0 \\ B_0 & B_0/2 \end{bmatrix}. \] (9)

Therefore, \( N_{v}^{1/2} \hat{\Psi}_{N_{v}}(\alpha_0, \beta_0) \) tends asymptotically to a bivariate normal distribution, with zero mean and covariance \( V(\alpha_0, \beta_0) \). This establishes the behavior of the left-hand side of (7).

Considering the right-hand side of (7), direct calculation shows that
\[
\dot{\hat{\Psi}}_{N_{v}}(\alpha_0, \beta_0) = -\frac{N_{v} - 1}{2} \left( \begin{array}{c} N_{v} \\ 2 \end{array} \right) V_{N_{v}}(\alpha_0, \beta_0) \to -V(\alpha_0, \beta_0).
\]

In addition, we know that \( \| (\hat{\alpha}^\dagger, \hat{\beta}^\dagger)^T - (\alpha_0, \beta_0)^T \| = o_P(1) \) and it is straightforward to show that \( \| \hat{\Psi}^\dagger \| = O_P(1) \).

Combining these results, we conclude that
\[
\sqrt{N_{v}} \left( \begin{array}{c} \hat{\alpha}^\dagger - \alpha \\ \hat{\beta}^\dagger - \beta \end{array} \right) \to N \left( 0, V^{-1}(\alpha_0, \beta_0) \right).
\]

Noting that
\[
V^{-1}(\alpha_0, \beta_0) = e^{-\alpha} \begin{bmatrix} 1 & -2 \\ -2 & 4 + 2e^{-\alpha-\beta} \end{bmatrix},
\]
completes the proof of the theorem.