

Curved Exponential Family Models for Social Networks*

David R. Hunter, Penn State University

March 2, 2006

Abstract

Curved exponential family models are a useful generalization of exponential random graph models (ERGMs). In particular, models involving the alternating k -star, alternating k -triangle, and alternating k -twopath statistics of Snijders et al (2006) may be viewed as curved exponential family models. This article unifies recent material in the literature regarding curved exponential family models for networks in general and models involving these alternating statistics in particular. It also discusses the intuition behind rewriting the three alternating statistics in terms of the degree distribution and the recently introduced shared partner distributions. This intuition suggests a redefinition of the alternating k -star statistic. Finally, this article demonstrates the use of the `statnet` package in R for fitting models of this sort, comparing new results on an oft-studied network dataset with results found in the literature.

Key Words: exponential random graph model, maximum likelihood estimation, p-star model

1 Introduction

For a fixed set of n actors, or nodes, and a network on those nodes, assume that \mathbf{Y} denotes the $n \times n$ adjacency matrix for the network; that is,

$$Y_{ij} = \begin{cases} 1 & \text{if an edge exists from } i \text{ to } j; \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

In some social networks applications, the goal is to produce a probabilistic model for \mathbf{Y} based on an observed network dataset. It is the goal of this article to explain a

*This research is supported by Grant DA012831 from NIDA and Grant HD041877 from NICHD.

particular class of models, called curved exponential family models, for achieving this end. We assume here that the reader is at least somewhat conversant in certain basic techniques of statistical modelling such as logistic regression, though not necessarily familiar with the intricacies of statistical modelling of networks.

Implicit in definition (1) is the fact that we assume no valued or multiple edges are allowed; furthermore, we disallow self-edges, so $Y_{ii} = 0$ for all i . Finally, we will treat only undirected networks in this article, which implies that $Y_{ij} = Y_{ji}$, though we make this choice only to simplify some of the arguments; there is little difficulty in extending all of the results here to the case of directed networks.

Beyond the network information contained in \mathbf{Y} , there are often additional data, such as a set of measured characteristics for each node in the network. For instance, when the nodes are people, we may know each person’s age and sex. Throughout this article, we let \mathbf{X} denote the additional data.

We assume throughout this article that the probability of observing a particular network is a function of statistics that may depend on the network itself as well as covariates measured on the nodes. For the particular class of models known as exponential random graph models (ERGMs), the relationship between a particular graph \mathbf{y} and its probability of occurrence conditional on the additional data \mathbf{X} is generally expressed as

$$P_{\boldsymbol{\eta}}(\mathbf{Y} = \mathbf{y}) = \frac{\exp\{\sum_{i=1}^p \eta_i g_i(\mathbf{y}, \mathbf{X})\}}{\kappa(\boldsymbol{\eta})} = \frac{\exp\{\boldsymbol{\eta}^t \mathbf{g}(\mathbf{y}, \mathbf{X})\}}{\kappa(\boldsymbol{\eta})}, \quad (2)$$

where $\mathbf{g}(\mathbf{y}, \mathbf{X})$ is a user-defined p -vector of statistics and $\boldsymbol{\eta} \in R^p$ denotes the statistical parameter governing the probabilistic formation of the network. The denominator, $\kappa(\boldsymbol{\eta})$, is a normalizing constant that ensures that the sum of (2) over all possible \mathbf{y} equals 1.

ERGMs are sometimes known in the social networks literature as p-star models (Wasserman and Pattison, 1996). We use “ERGM” instead of “p-star” here due to the vast statistical literature covering exponential family models (e.g., Barndorff-Nielsen, 1978; Brown, 1986). Nevertheless, we consider “p-star” to be synonymous with “ERGM” in this article, with one caveat: Wasserman and Pattison (1996) used a method of parameter estimation, maximum pseudo-likelihood estimation, that has

come to be closely associated with the p-star models themselves. However, in this article we separate the name of the models (ERGMs or p-star models) from the method of estimating their parameters. In particular, we do not discuss maximum pseudo-likelihood estimation here, focusing instead on the better-understood method of maximum likelihood estimation. ERGMs are discussed in detail by Robins et al (2006a).

The remainder of this article concerns generalizations of (2) known as curved exponential family models. Section 2 discusses these models in general terms, while Sections 3 and 4 focus on specific examples — namely, models involving the alternating k -star, alternating k -triangle, and alternating k -twopath statistics developed by Snijders et al (2006). These statistics have recently shown great promise in producing parsimonious models that fit certain social network datasets well; they are discussed further by Robins et al (2006b). Much of Sections 3 and 4 is devoted to a reformulation of these statistics in terms of the degree statistics and the recently-developed shared partner statistics, as well as an attempt to reveal how this reformulation aids in interpreting these statistics. Finally, Section 5 demonstrates the use of these models on real data. The analysis, which recreates and extends some earlier work using the same dataset, is carried out using the `statnet` package for R, available for use at csde.washington.edu/statnet. The computer code used in Section 5 may be found in the Appendix.

2 Curved exponential families

In model (2), the maximum likelihood estimator (MLE) of the parameter vector $\boldsymbol{\eta}$ is, by definition, the vector that maximizes $P_{\boldsymbol{\eta}}(\mathbf{Y} = \mathbf{y}_{\text{obs}})$ as a function of $\boldsymbol{\eta}$, where \mathbf{y}_{obs} is the observed network. In other words, if we let $\hat{\boldsymbol{\eta}}$ denote the MLE and we assume that $\boldsymbol{\eta}$ is a vector contained in p -dimensional space (denoted \mathbb{R}^p), then we may write

$$\hat{\boldsymbol{\eta}} = \arg \max_{\boldsymbol{\eta} \in \mathbb{R}^p} \frac{\exp\{\boldsymbol{\eta}^t \mathbf{g}(\mathbf{y}_{\text{obs}}, \mathbf{X})\}}{\kappa(\boldsymbol{\eta})}. \quad (3)$$

It is worth noting here that it is extremely difficult to find $\hat{\boldsymbol{\eta}}$ except in the case of a simplistic model (e.g., one in which all edges are assumed independent of one

another, so the model is simply logistic regression) or a very small network of no more than about ten nodes. This is part of the reason that the science of *fitting* models like (2) — that is, estimating the parameters — has lagged so far behind the *proposal* of models like (2) in the literature. Two similar yet distinct methods to approximate $\hat{\boldsymbol{\eta}}$ are outlined by Snijders (2002) and Hunter and Handcock (2006). Though these methods are beyond the scope of the current article, suffice it to say that there are many potential pitfalls; these are discussed by Snijders (2002), Handcock (2002; 2003), Hunter and Handcock (2006), and Robins et al (2006a).

To add yet another layer of complexity to the problem of finding $\hat{\boldsymbol{\eta}}$, suppose that $\hat{\boldsymbol{\eta}}$ is not allowed to be *any* element of \mathbb{R}^p , but rather that it must be contained in some subset $S \subset \mathbb{R}^p$:

$$\hat{\boldsymbol{\eta}}^* = \arg \max_{\boldsymbol{\eta} \in S} \frac{\exp\{\boldsymbol{\eta}^t \mathbf{g}(\mathbf{y}_{\text{obs}}, \mathbf{X})\}}{\kappa(\boldsymbol{\eta})}. \quad (4)$$

Note that if S is taken to be \mathbb{R}^p itself, then equations (4) and (3) are identical; thus, (4) generalizes (3). The reason for considering the constrained optimization problem of equation (4) may not be immediately clear, especially since constrained optimization is generally more difficult to accomplish than unconstrained optimization as in equation (3).

To see how a problem such as (4) might arise in theory, suppose that $\boldsymbol{\theta}$ is a vector in q -dimensional space \mathbb{R}^q , where $q < p$. That is, $\boldsymbol{\theta}$ has fewer components than $\boldsymbol{\eta}$. Now suppose that we regard $\boldsymbol{\eta}$ as a function of $\boldsymbol{\theta}$, so that model (2) becomes

$$P_{\boldsymbol{\eta}}(\mathbf{Y} = \mathbf{y}) = \frac{\exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^t \mathbf{g}(\mathbf{y}, \mathbf{X})\}}{\kappa[\boldsymbol{\eta}(\boldsymbol{\theta})]}. \quad (5)$$

Then the maximum likelihood estimator of $\boldsymbol{\theta}$ is the q -dimensional vector $\hat{\boldsymbol{\theta}}$ that maximizes (5) as a function of $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \mathbb{R}^q} \frac{\exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^t \mathbf{g}(\mathbf{y}, \mathbf{X})\}}{\kappa[\boldsymbol{\eta}(\boldsymbol{\theta})]}. \quad (6)$$

One may think of this as constrained optimization (4) in which S is the set of all $\boldsymbol{\eta}(\boldsymbol{\theta})$ as $\boldsymbol{\theta}$ ranges over all of \mathbb{R}^q , but it is probably easier to view the problem as unconstrained maximization as in (6).

Note that matters simplify considerably if $\boldsymbol{\eta}(\boldsymbol{\theta})$ is a linear function of $\boldsymbol{\theta}$, say $\boldsymbol{\eta}(\boldsymbol{\theta}) = \mathbf{M}\boldsymbol{\theta}$ for some constant $p \times q$ matrix \mathbf{M} . For in that case, model (5) becomes

$$P_{\boldsymbol{\eta}}(\mathbf{Y} = \mathbf{y}) = \frac{\exp\{\boldsymbol{\theta}^t \mathbf{M}^t \mathbf{g}(\mathbf{y}, \mathbf{X})\}}{\kappa(\mathbf{M}\boldsymbol{\theta})};$$

thus, by letting $\mathbf{g}^*(\mathbf{y}, \mathbf{X}) = \mathbf{M}^t \mathbf{g}(\mathbf{y}, \mathbf{X})$ and $\kappa^*(\boldsymbol{\theta}) = \kappa(\mathbf{M}\boldsymbol{\theta})$, we obtain the model

$$P_{\boldsymbol{\eta}}(\mathbf{Y} = \mathbf{y}) = \frac{\exp\{\boldsymbol{\theta}^t \mathbf{g}^*(\mathbf{y}, \mathbf{X})\}}{\kappa^*(\boldsymbol{\theta})},$$

which is an ERGM just like (2). In other words, replacing $\boldsymbol{\eta}$ by $\boldsymbol{\eta}(\boldsymbol{\theta})$ is only interesting if $\boldsymbol{\eta}(\boldsymbol{\theta})$ is *not* a linear function of $\boldsymbol{\theta}$, which is why we call model (5) a *curved* exponential family model.

Curved exponential families were first referred to as such by Efron (1975), who later wrote another article discussing the geometric intuition behind them (Efron, 1978). Finding a maximum likelihood estimator (6) in a curved exponential family is generally more difficult than in a regular (non-curved) exponential family. For instance, the Robbins-Monro algorithm (Robbins and Monro, 1951), exploited by Snijders (2002) to approximate an MLE for model (2), is based on an elegant equality that is true for regular exponential families but not for curved exponential families. Nonetheless, it is still possible to perform approximate maximum likelihood estimation for curved exponential families. Efron (1978) elegantly describes the geometry underlying such maximization, and Hunter and Handcock (2006) discuss the technical details of implementing it for the case of network models. However, we now lay general theory aside in favor of a particular application, namely, the alternating statistics of Snijders et al (2006).

3 Rewriting alternating k -stars

The first curved exponential family model we will consider involves the k -star statistics $S_1(\mathbf{y}), \dots, S_{n-1}(\mathbf{y})$, where $S_k(\mathbf{y})$ denotes the number of k -stars in the graph \mathbf{y} . A k -star is a set of k distinct edges that all share an endpoint. In particular, a 1-star is simply an edge. Note that the number of edges in the graph \mathbf{y} is sometimes denoted by $L(\mathbf{y})$, though we use $S_1(\mathbf{y})$ in this article.

When each k -star statistic has its own coefficient in an ERGM, the resulting model is

$$P(\mathbf{Y} = \mathbf{y}) = \frac{\exp\{\sum_{i=1}^{n-1} \eta_i S_i(\mathbf{y})\}}{\kappa(\boldsymbol{\eta})}. \quad (7)$$

Recent work by Snijders et al (2006) considers a more refined model based on the k -star statistics. If λ_s is a known positive constant (the subscripted s is for “star”), then we may define the scalar-valued network statistic

$$a^{(s)}(\mathbf{y}; \lambda_s) = S_2(\mathbf{y}) - \frac{S_3(\mathbf{y})}{\lambda_s} + \cdots + (-1)^{n-3} \frac{S_{n-1}(\mathbf{y})}{\lambda_s^{n-3}}. \quad (8)$$

Because of the alternating signs in front of the $S_k(\mathbf{y})$ terms, Snijders et al (2006) call $a^{(s)}(\mathbf{y}; \lambda_s)$ an *alternating k -star statistic*. Now suppose that we formulate a simple ERGM with only a single scalar parameter, say ξ , and the single statistic $a^{(s)}$. That is, suppose that

$$P(\mathbf{Y} = \mathbf{y}) = \frac{\exp\{\xi a^{(s)}(\mathbf{y}; \lambda_s)\}}{\kappa(\xi, \lambda_s)}. \quad (9)$$

As long as λ_s is fixed, then model (9) is a standard ERGM, by which we mean a model of the form (2). But how should λ_s be chosen? Certainly there is no hope of making an appropriate selection without having some sense of the intuitive meaning of λ_s . We will delve into this intuition later in this section. But even if a reasonable value may be decided on for a particular application, how can we know whether such a choice is the best choice?

If one wishes to automate the choice of λ_s , then λ_s becomes a parameter in the model just like ξ . However, if both ξ and λ_s are parameters to be estimated, then model (9) is no longer a standard ERGM. Instead, one may check that model (9) is just like model (7) as long as the constraints

$$\eta_1 = 0 \quad \text{and} \quad \eta_i \equiv \eta_i(\xi, \lambda_s) = \frac{(-1)^i \xi}{\lambda_s^{i-2}}, 2 \leq i \leq n-1$$

are satisfied. Therefore, if λ_s is unknown and to be estimated, model (9) is a curved exponential family model of the form (5), where $\boldsymbol{\theta}$ is the vector (ξ, λ_s) in \mathbb{R}^2 (we ignore the problems that arise when $\lambda_s \leq 0$ for now).

Now that we have established that model (9) is a curved exponential family model, it might seem that we need only apply the ideas of Hunter and Handcock (2006) to obtain approximate maximum likelihood estimates for the parameters ξ and λ_s . However, to do so blindly would be poor modelling practice; it is better to develop a sense of what the model terms mean. It is partly with this goal in mind that we now demonstrate one way to rewrite the alternating k -star statistic in terms of the degree statistics D_0, \dots, D_{n-1} , where $D_k(\mathbf{y})$ denotes the number of nodes in the network \mathbf{y} that have k neighbors in the network.

A node with i neighbors is the center of $\binom{i}{k}$ k -stars (but the “center” of a 1-star may be considered arbitrarily to be either of two nodes), which implies that the k -star statistics may be rewritten in terms of the degree statistics as follows:

$$S_k(\mathbf{y}) = \sum_{i=k}^{n-1} \binom{i}{k} D_i(\mathbf{y}), \quad 2 \leq k \leq n-1, \quad \text{and} \quad S_1(\mathbf{y}) = \frac{1}{2} \sum_{i=1}^{n-1} i D_i(\mathbf{y}). \quad (10)$$

Equations (10) are linear in the d_i and they may be inverted, though we do not give the inverse formulas here.

Using equations (10) and the binomial expansion

$$\left(1 - \frac{1}{\lambda_s}\right)^i = \sum_{j=0}^i \binom{i}{j} \left(-\frac{1}{\lambda_s}\right)^j, \quad (11)$$

the alternating k -star statistic (8) may be rewritten

$$a^{(s)}(\mathbf{y}; \lambda_s) = \lambda_s^2 \sum_{i=1}^{n-1} \left\{ \left(1 - \frac{1}{\lambda_s}\right)^i - 1 \right\} D_i(\mathbf{y}) + 2\lambda_s S_1(\mathbf{y}). \quad (12)$$

Note that the $S_1(\mathbf{y})$ term in (12) is the number of edges, or 1-stars, in \mathbf{y} . The derivation of expression (12) proceeds by first rewriting the alternating k -star statistic

(8) using summation notation:

$$\begin{aligned}
\sum_{k=2}^{n-1} \left(-\frac{1}{\lambda_s}\right)^{k-2} S_k(\mathbf{y}) &= \sum_{k=2}^{n-1} \sum_{i=k}^{n-1} \binom{i}{k} \left(-\frac{1}{\lambda_s}\right)^{k-2} D_i(\mathbf{y}) && \text{by (10)} \\
&= \lambda_s^2 \sum_{i=2}^{n-1} \sum_{k=2}^i \binom{i}{k} \left(-\frac{1}{\lambda_s}\right)^k D_i(\mathbf{y}) \\
&= \lambda_s^2 \sum_{i=2}^{n-1} \left\{ \left(1 - \frac{1}{\lambda_s}\right)^i + \frac{i}{\lambda_s} - 1 \right\} D_i(\mathbf{y}) && \text{by (11)} \\
&= \lambda_s^2 \sum_{i=1}^{n-1} \left\{ \left(1 - \frac{1}{\lambda_s}\right)^i - 1 \right\} D_i(\mathbf{y}) + 2\lambda_s S_1(\mathbf{y}) && \text{by (10)}.
\end{aligned}$$

Note that the second step above involves switching the order of summation.

We may rewrite equation (12) slightly as follows. First, to ensure that λ_s is positive, define a new parameter θ_s such that $\lambda_s = \exp\{\theta_s\}$. Then equation (12) becomes

$$a^{(s)}(\mathbf{y}; \lambda_s) = e^{\theta_s} [2S_1(\mathbf{y}) - u(\mathbf{y}; \theta_s)], \quad (13)$$

where $u(\mathbf{y}; \theta_s)$ is defined to be the *geometrically weighted degree* (GWD) statistic:

$$u(\mathbf{y}; \theta_s) = e^{\theta_s} \sum_{i=1}^{n-1} \left\{ 1 - (1 - e^{-\theta_s})^i \right\} D_i(\mathbf{y}). \quad (14)$$

The G in GWD reflects the fact that this statistic is based on the geometric sequence $(1 - e^{-\theta_s})^i$. Snijders et al (2006, equation 14) give an equation very similar to equation (13), yet their equation is slightly different because they use a different definition of the geometrically weighted degree statistic; see the remark at the end of this section for an explanation of this difference.

Equation (13) shows that the alternating k -star statistic $a^{(s)}$ is a linear combination of u and S_1 , and it is easy to invert this expression to give u as a linear combination of $a^{(s)}$ and S_1 . Therefore, an ERGM containing both $a^{(s)}(\mathbf{y}; \lambda_s)$ and $S_1(\mathbf{y})$ is mathematically equivalent to a model containing both $u(\mathbf{y}; \log \lambda_s)$ and $S_1(\mathbf{y})$. Because of this equivalence, and the fact that nearly every model of interest contains the $S_1(\mathbf{y})$ term (for much the same reason that nearly every linear regression model contains an intercept term), the alternating k -star statistic and the GWD statistic are essentially interchangeable from a modelling standpoint.

However, we argue that u is slightly preferable to $a^{(s)}$ for several reasons. First, the degree statistics are more ubiquitous in the networks literature than the k -star statistics. Second, as we see in the next section, u has an analogous form to the other two alternating statistics of Snijders et al (2006), the alternating k -triangle and alternating k -twopath statistics; it is only the alternating k -star statistic that has a qualitatively different form than its corresponding geometrically weighted statistic [to see why u is appealing in this light, compare equation (14) to equations (25) and (26)]. Finally, u makes the interpretation of its parameter values slightly easier than $a^{(s)}$ does, as we now demonstrate.

Instead of model (9), consider the related model

$$P(\mathbf{Y} = \mathbf{y}) = \frac{\exp\{\phi u(\mathbf{y}; \theta_s)\}}{\kappa(\phi, \theta_s)}. \quad (15)$$

To understand how θ_s and ϕ influence the probabilistic model for the formation of networks, consider the effect of adding a single edge to the graph. The result of this addition is that the degrees of each of the incident nodes increase by one. Let us focus on just one such increase, where a node of degree k becomes a node of degree $k+1$; i.e., (D_k, D_{k+1}) is replaced by $(D_k - 1, D_{k+1} + 1)$ for some k and no other change to the degree distribution is made. How would this change affect the probability of the graph? If p_{before} and p_{after} denote the probabilities before and after this change, substituting (14) into (15) reveals that

$$\frac{p_{\text{after}}}{p_{\text{before}}} = \frac{\exp\{\phi \lambda_s [(D_k - 1)(1 - \rho^k) + (D_{k+1} + 1)(1 - \rho^{k+1})]\}}{\exp\{\phi \lambda_s [D_k(1 - \rho^k) + D_{k+1}(1 - \rho^{k+1})]\}} = \exp\{\phi \rho^k\}, \quad (16)$$

where $\rho = 1 - e^{-\theta_s}$ to simplify notation. Using equation (16) to account for the increase in degree of *both* incident nodes, we see that adding an edge between two nodes of degrees k and ℓ results in an increase of $\phi(\rho^k + \rho^\ell)$ in the log-probability. However, we continue to focus on equation (16) for the sake of simplicity.

Whereas some network models entail preferential attachment, in which high-degree nodes are more likely than low-degree nodes to form edges (e.g., Albert and Barabási, 2002), we see from equation (16) that $u(\mathbf{y}; \theta_s)$ may be thought of as a sort of anti-preferential attachment model term: The benefit of adding an additional edge, expressed on the log-odds scale, decreases geometrically with the degree of the nodes

involved (after accounting for all other model terms, of course). The speed of the geometric decay is controlled by the θ_s parameter: As θ_s increases in the positive real numbers, so does the ratio ρ that determines the geometric rate of decay of the log-odds; the higher the θ_s , the slower the decay. The ϕ parameter is the multiplier for this effect. In particular, $\phi > 0$ implies a preference for adding edges, whereas $\phi < 0$ implies a preference for deleting them. When $\theta_s = 0$, the preference for adding new nodes disappears regardless of the value of ϕ , whereas for very large values of θ_s , this preference does not diminish much but remains roughly constant at ϕ . Negative values of θ_s are allowed, but they are more difficult to interpret.

If we use model (9), containing the alternating k -star statistic $a^{(s)}(\mathbf{y}; \lambda_s)$, instead of model (15), then the analogue of equation (16) is

$$\frac{p_{\text{after}}}{p_{\text{before}}} = \exp\{\xi \lambda_s (1 - \rho^k)\} = \exp\{\xi e^{\theta_s} (1 - \rho^k)\}.$$

Recall that $\rho = 1 - e^{-\theta_s}$, so that the multiplier ξe^{θ_s} is mixed up with the geometric ratio ρ . This makes interpretation of the ξ coefficient slightly more difficult than in equation (16). Furthermore, as k increases, the $(1 - \rho^k)$ term *increases* to 1 at a geometric rate. This is a bit like preferential attachment, as Snijders et al (2006) point out; however, the preferential effect is nearly nonexistent for ρ near zero and/or for nodes with large degree (large k), since $(1 - \rho^k) \approx 1$ in either case. This makes interpretation of the model parameters tricky: Using model (9) with $\xi > 0$ and $\lambda_s > 1$ (which gives $0 < \rho < 1$) results in a preferential attachment effect for small-degree nodes, but this preference tapers off and eventually disappears for larger-degree nodes.

We conclude that u and $a^{(s)}$ produce mathematically equivalent models, as long as they are used together with the S_1 statistic, yet the interpretation of the u parameters is slightly more straightforward than the interpretation of the $a^{(s)}$ parameters. For this reason, we recommend the use of the geometrically weighted degree statistic u instead of the alternating k -star statistic $a^{(s)}$. Fortunately, this is not an issue for either of the alternating statistics in the next section: Both the alternating k -triangle statistic and the alternating k -twopath statistic coincide exactly with corresponding geometrically weighted statistics defined analogously to equation (14), as we shall demonstrate in Section 4.

Interestingly, combining equations (13) and (8) shows that

$$u(\mathbf{y}; \log \lambda_s) = 2S_1(\mathbf{y}) - \frac{S_2(\mathbf{y})}{\lambda_s} + \cdots + (-1)^n \frac{S_{n-1}(\mathbf{y})}{\lambda_s^{n-2}}, \quad (17)$$

which has the same look as the alternating k -star statistic, except that it begins with the S_1 term instead of the S_2 term.

Remark: There is a geometrically weighted degree statistic in Snijders et al (2006), but this statistic is not the same as u in the current article. Their statistic, labeled $u_\alpha^{(d)}(\mathbf{y})$ and defined by

$$u_\alpha^{(d)}(\mathbf{y}) = \sum_{k=0}^{n-1} \exp\{-\alpha k\} D_k(\mathbf{y}), \quad (18)$$

is used in Snijders et al (2006) mostly to demonstrate the relationship between the alternating k -star statistic and the degree statistics. Although it is in a certain sense mathematically equivalent to the u statistic, $u_\alpha^{(d)}$ is not constructed to make its parameters interpretable in a meaningful, intuitive way. Thus, by “geometrically weighted degree statistic” we will always mean the u statistic of equation (14). The relationship among λ_s , θ_s , and the α parameter of equation (18) is summarized by $1/\lambda_s = e^{-\theta_s} = 1 - e^{-\alpha}$.

4 Shared partner statistics

Section 3 defines the alternating k -star statistic and shows that it may be rewritten as in equation (12) in terms of the degree statistics. In an analogous way, we now define the alternating k -triangle and alternating k -twopath statistics of Snijders et al (2006) and show that they may be rewritten in terms of the edgewise and dyadic shared partner statistics, respectively, which we also define.

The k -triangle and the k -twopath are concepts that generalize the ideas of triangle and 2-star, respectively. [Snijders et al (2006) actually coin the term “ k -independent 2-path,” but Hunter et al (2005) shorten this to the simpler “ k -twopath”.] A k -triangle is defined to be a set of k distinct triangles that share a common edge. A k -twopath is a set of k distinct paths of length two joining the same pair of nodes. In particular, a 1-triangle is the same thing as a triangle and a 1-twopath is the

same thing as a 2-star – however, note in the latter case that a 1-twopath is usually associated with its *endpoints* whereas a 2-star is usually associated with its *center*. We denote the number of k -triangles and k -twopaths for the graph \mathbf{y} by $T_k(\mathbf{y})$ and $P_k(\mathbf{y})$, respectively, where k may take any value from 1 to $n - 2$.

In analogy to the alternating k -star statistic (8), Snijders et al (2006) also define the alternating k -triangle and alternating k -twopath statistics as

$$3T_1(\mathbf{y}) - \frac{T_2(\mathbf{y})}{\lambda_t^1} + \dots + (-1)^{n-3} \frac{T_{n-2}(\mathbf{y})}{\lambda_t^{n-3}} \quad (19)$$

and

$$P_1(\mathbf{y}) - \frac{2P_2(\mathbf{y})}{\lambda_p^1} + \dots + (-1)^{n-3} \frac{P_{n-2}(\mathbf{y})}{\lambda_p^{n-3}}, \quad (20)$$

respectively. To complete the analogy, we will show how to rewrite (19) and (20) in a form similar to (12). For this purpose, we need the shared partner statistics.

We define two distinct sets of shared partner statistics, the *edgewise* shared partner statistics and the *dyadic* shared partner statistics. These statistics are closely related to the two-path statistics L_{2ij} defined by Snijders et al (2006). The edgewise shared partner statistics are denoted $EP_0(\mathbf{y}), \dots, EP_{n-2}(\mathbf{y})$, where $EP_k(\mathbf{y})$ is defined as the number of unordered pairs $\{i, j\}$ such that $y_{ij} = 1$ and i and j have exactly k common neighbors. The requirement that $y_{ij} = 1$ distinguishes the edgewise shared partner statistics from the dyadic shared partner statistics $DP_0(\mathbf{y}), \dots, DP_{n-2}(\mathbf{y})$: We define $DP_k(\mathbf{y})$ to be the number of pairs $\{i, j\}$ such that i and j have exactly k common neighbors, regardless of the value of y_{ij} . By these definitions, it is always true that $DP_k(\mathbf{y}) \geq EP_k(\mathbf{y})$, and in fact $DP_k(\mathbf{y}) - EP_k(\mathbf{y})$ equals the number of unordered pairs $\{i, j\}$ for which $y_{ij} = 0$ and i and j share exactly k common neighbors. The edgewise shared partner statistics first appear in Hunter and Handcock (2006), who refer to them merely as shared partner statistics. Hunter et al (2005) introduce the dyadic shared partner statistics.

Just as the degree statistics $D_i(\mathbf{y})$ are related to the k -star statistics $S_k(\mathbf{y})$ by (10), the edgewise and dyadic shared partner statistics are related to the k -triangle and k -twopath statistics, respectively, by the equations

$$T_k(\mathbf{y}) = \sum_{i=k}^{n-2} \binom{i}{k} EP_i(\mathbf{y}), \quad 2 \leq k \leq n - 2 \quad (21)$$

and

$$P_k(\mathbf{y}) = \sum_{i=k}^{n-2} \binom{i}{k} DP_i(\mathbf{y}), \quad 1 \leq k \leq n-2, k \neq 2. \quad (22)$$

The cases not covered in (21) and (22) are those of $T_1(\mathbf{y})$, the number of triangles, and $P_2(\mathbf{y})$, the number of 4-cycles. In the former case, there is an extra factor of $1/3$ because any of the three edges in a triangle may be considered the “common” edge of a 1-triangle; in the latter case, there is an extra factor of $1/2$ because any 4-cycle can be considered a 2-path between two distinct pairs of nodes. Thus, we obtain

$$T_1(\mathbf{y}) = \frac{1}{3} \sum_{i=0}^{n-2} i EP_i(\mathbf{y}); \quad (23)$$

$$P_2(\mathbf{y}) = \frac{1}{2} \sum_{i=2}^{n-2} \binom{i}{2} DP_i(\mathbf{y}). \quad (24)$$

A more detailed explanation of equations (21) through (24) is given by Hunter et al (2005).

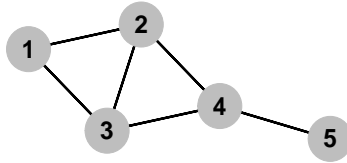


Figure 1: For this simple five-node network, the edgewise and dyadic shared partner distributions are $(EP_0, \dots, EP_3) = (1, 4, 1, 0)$ and $(DP_0, \dots, DP_3) = (2, 6, 2, 0)$, respectively; the k -triangle and k -twopath distributions are $(T_1, T_2, T_3) = (2, 1, 0)$ and $(P_1, P_2, P_3) = (10, 1, 0)$, respectively.

As a concrete example, the simple network of Figure 1 has two 1-triangles, one 2-triangle, ten 1-twopaths, and one 2-twopath. Note that the 2-twopath joining nodes 1 and 4 is the same as the 2-twopath joining nodes 2 and 3, so it is counted only once. One may check that equations (21) through (24) hold for this network.

To rewrite the alternating k -triangle and alternating k -twopath statistics in terms of the EP_k and DP_k statistics, begin by substituting equations (21) through (24) into (19) and (20). Next, simplify as in the derivation of equation (12). Finally, introduce

the parameters $\theta_t = \log \lambda_t$ and $\theta_p = \log \lambda_p$ to ensure λ_t and λ_p are positive. The result is that the alternating k -triangle statistic equals

$$v(\mathbf{y}; \theta_t) = e^{\theta_t} \sum_{i=1}^{n-2} \left\{ 1 - (1 - e^{-\theta_t})^i \right\} \text{EP}_i(\mathbf{y}) \quad (25)$$

and the alternating k -twopath statistic equals

$$w(\mathbf{y}; \theta_p) = e^{\theta_p} \sum_{i=1}^{n-2} \left\{ 1 - (1 - e^{-\theta_p})^i \right\} \text{DP}_i(\mathbf{y}). \quad (26)$$

Because of the geometric series $(1 - e^{-\theta_t})^i$ and $(1 - e^{-\theta_p})^i$, we call $v(\mathbf{y}; \theta_t)$ and $w(\mathbf{y}; \theta_p)$ the “geometrically weighted edgewise shared partner” (GWESP) and “geometrically weighted dyadic shared partner” (GWDSP) statistics, respectively.

Because the functional forms of $v(\mathbf{y}; \theta_t)$ and $w(\mathbf{y}; \theta_p)$ are the same as the form of $u(\mathbf{y}; \theta_u)$ — as one may verify by comparing equations (25) and (26) to equation (14) — their interpretations are also the same, except of course that the interpretations involve the shared partner statistics instead of the degree statistics. We will apply these interpretations to real data in Section 5.

5 Lazega’s Lawyer dataset

Lazega (Lazega and Pattison, 1999; Lazega, 2001) collected and analyzed data on working relations among 36 partners in a New England law firm. Here, we deal with only a subset of the data: An (undirected) edge will be said to exist between two partners if and only if each indicates a collaboration with the other. These data are analyzed by both Snijders et al (2006) and Hunter and Handcock (2006). Here, we employ models similar to those used in both of these articles in order to compare and contrast the results. We refer to these two articles so frequently in this section that we abbreviate them as SPRH (for the four authors’ names) and HH, respectively.

The nodal covariates measured on each partner are sex, office location (there are three offices in different cities), type of practice (there are two possible types of practice, litigation=0 and corporate law=1), and seniority (rank order of entry into the firm, so that smaller numbers correspond to more senior partners). These

covariates comprise \mathbf{X} , which in this case is a 36×4 matrix because there are 36 nodes and 4 covariates measured on each node.

As in both SPRH and HH, the goal here is to model the observed network of collaborations as a function of both network statistics and nodal covariates. To do so requires the selection of the $\mathbf{g}(\mathbf{y}, \mathbf{X})$ statistics that will go into model (5). Following SPRH, we allow $\mathbf{g}(\mathbf{y}, \mathbf{X})$ to include several statistics that depend on both \mathbf{X} and \mathbf{y} . These include so-called main effects of seniority and practice along with homophily effects of practice, sex, and office location. The main effect of, say, seniority is defined to be the sum of the seniority value of the endpoints of each of the edges of \mathbf{y} . This may be expressed as

$$\sum_{1 \leq i < j \leq n} y_{ij} (\text{seniority}_i + \text{seniority}_j).$$

In other words, whenever an edge is added to the network, the seniority main effect is increased by the sum of the seniority attributes of its two endpoints. The homophily effect of, say, practice is defined to be the number of edges in \mathbf{y} for which both endpoints have the same value of practice. Thus, if an edge is added to the network, it increases the homophily effect of practice by one if and only if the endpoints have the same value of practice. The homophily effects in SPRH are defined to be twice as large, with the result that the corresponding parameter values are half as large; this may be seen by comparing Table 1, Model 1 in HH with Table 1, Model 2 in SPRH.

The models are fit using the R package `statnet`, available for general use on a secure server at csde.washington.edu/statnet. The first model, whose parameter estimates are shown in Table 1, includes the edge count $s_1(\mathbf{y})$ along with the geometrically weighted degree (GWD), geometrically weighted edgewise shared partner (GWESP), and geometrically weighted dyadic shared partner (GWDSP) statistics and their respective θ parameters. In SPRH, these θ parameters are considered fixed (which makes the model a standard ERGM), but here we estimate them as curved exponential family model parameters. Actually, even in SPRH, the θ parameters are not quite fixed; those authors use several different values for each θ and settle on values that give high estimates of the likelihood function. The curved exponential family techniques used here and in HH automate this procedure.

| Parameter | estimate (s.e.) |
|----------------------------------|-------------------------------|
| Edges | -6.714 (0.740) ^{***} |
| Main Seniority | 0.023 (0.007) ^{**} |
| Main Practice | 0.404 (0.124) ^{**} |
| Homophily Practice | 0.786 (0.204) ^{***} |
| Homophily Sex | 0.671 (0.264) [*] |
| Homophily Office | 1.167 (0.212) ^{***} |
| GWD | 0.294 (0.385) |
| θ_s (= $\log \lambda_s$) | -0.741 (0.150) ^{***} |
| GWESP | 0.920 (0.265) ^{**} |
| θ_t (= $\log \lambda_t$) | 0.808 (0.109) ^{***} |
| GWDSP | 0.035 (0.064) |
| θ_p (= $\log \lambda_p$) | -0.006 (3.162) |
| *0.005 \leq p-value < 0.05 | |
| **0.0005 \leq p-value < 0.005 | |
| ***p-value < 0.0005 | |

Table 1: Approximate maximum likelihood results for a curved exponential family model containing the GWD, GWESP, and GWDSP terms in which the corresponding parameters (θ_s , θ_t , and θ_p) are estimated.

| Parameter | estimate (s.e.) | HH |
|----------------------------------|-------------------------------|------------------------------|
| Edges | -6.457 (0.649) ^{***} | N/A |
| Main Seniority | 0.023 (0.007) ^{**} | 0.023 (0.006) ^{***} |
| Main Practice | 0.407 (0.121) ^{**} | 0.390 (0.117) ^{***} |
| Homophily Practice | 0.761 (0.192) ^{***} | 0.757 (0.194) ^{***} |
| Homophily Sex | 0.697 (0.250) [*] | 0.688 (0.248) ^{***} |
| Homophily Office | 1.131 (0.202) ^{***} | 1.123 (0.194) ^{***} |
| GWESP | 0.898 (0.216) ^{***} | 0.878 (0.279) ^{***} |
| θ_t (= $\log \lambda_t$) | 0.779 (0.098) ^{***} | 0.814 (0.196) ^{***} |
| *0.005 \leq p-value < 0.05 | | |
| **0.0005 \leq p-value < 0.005 | | |
| ***p-value < 0.0005 | | |

Table 2: Approximate maximum likelihood results for a reduced model that includes only the significant terms. The HH column lists corresponding parameter estimates for the model reported in Hunter and Handcock (2006) for a model fit conditional on the number of edges being fixed.

The significance tests performed on each of the parameters, the results of which are indicated by asterisks in Table 1, involve the usual null hypotheses that the true parameter values equal zero. Note that in the case of the three parameters θ_s , θ_t , and θ_p , this test does not have the usual interpretation due to the fact that the value zero does not indicate the absence of the corresponding effect. Indeed, in the case of GWESP and GWDSP, the zero value was treated specially by SPRH: For GWESP, a value of $\theta_t = 0$ means that this statistic measures the number of pairs of nodes that are connected both by a direct edge and by a two-path through another node; for GWDSP, $\theta_p = 0$ leads to a count of all pairs of nodes that are connected by a two-path. SPRH refers to these statistics as “number of pairs directly & indirectly connected” and “number of pairs indirectly connected”, respectively. Model 1 of Table 1 in SPRH even includes these two terms in the ERGM, though it is pointed out that these terms are highly collinear with their GWESP and GWDSP counterparts that use the non-zero values of θ_t and θ_p .

The estimated parameter values $\hat{\theta}_s = -0.741$, $\hat{\theta}_t = 0.808$, and $\hat{\theta}_p = -0.0061$ given in Table 1 may be exponentiated to give their corresponding λ values 0.477, 2.243, and 0.994. SPRH fixes $\lambda = 3$ for each of these values, which is clearly quite different from the estimates obtained for the GWD and GWDSP statistics; however, neither of these statistics has a statistically significant leading coefficient, which means we should not overinterpret the estimates of λ_s or λ_p . Only the GWESP statistic, with its estimate of $\hat{\lambda}_t = \exp\{.808\} = 2.243$, is significant, and its positive coefficient of 0.898 indicates transitivity in this network.

The transitivity indicated by a significant and positive GWESP coefficient is beyond the transitivity that may be explained solely by nodal characteristics such as sex, office, practice, or seniority, since terms for each of these characteristics are included in the model (recall that every term in the model should be interpreted as though all other model effects have been accounted for). Specifically, we may interpret the GWESP estimates of 0.898 and 0.779 by considering how the probability of the network changes due to the GWESP term when a single pair of connected nodes increases its number of shared partners by one, assuming nothing else changes and all other model effects have been accounted for. Thus, for some k , (EP_k, EP_{k+1}) is

replaced by $(EP_k - 1, EP_{k+1} + 1)$ but all other EP_i statistics remain unchanged, and we let p_{before} and p_{after} denote the probabilities before and after this change. Even though such a specific change to the EP_i statistics is not generally possible by a simple alteration such as adding an edge, consideration of this case helps aid intuition about which types of networks this model will favor. After a bit of algebra we obtain

$$\log\left(\frac{p_{\text{after}}}{p_{\text{before}}}\right) = 0.898 \times (1 - \exp\{-0.779\})^k = .898 \times 0.54^k.$$

In other words, it is easiest to complete a triangle when none exists already ($k = 0$); such a move results in an increase of 0.898 on the log-probability scale beyond the effects predicted by the model’s other terms. However, this additional increase is roughly halved for each unit increase in k . Thus, completing a 2-triangle when a triangle already exists only results in an additional increase of 0.898×0.54 ; completing a 3-triangle when a 2-triangle already exists only gives 0.898×0.54^2 ; and so on. Thus, as two neighbors share more and more partners, the impetus to find additional shared partners decreases.

In comparing parameter estimates of this article with parameter estimates in both SPRH and HH, it is important to keep in mind a major difference among these models: The models here include an edges term, whereas both SPRH and HH carry out model-fitting under the assumption of a reduced sample space, including only those graphs that have exactly the same number of edges — in this case, 115 — as the observed graph. Both SPRH and HH cite numerical expediency as a reason for this choice (it is easier to carry out the very difficult computations necessary for estimation using the reduced sample space), though HH is also motivated by the desire to compare results directly with those of SPRH. We can see the effect of this major difference in Table 2, where we compare two models side-by-side: In the right column are the results from Model 2 in Table 1 of HH, and in the left column are the same results for a model that includes an edge term. These models have removed the insignificant GWD and GWDSP terms from the earlier model; all remaining terms are statistically significant. The results with the edge term are remarkably similar to the results without the edge term, underscoring a statement made in HH that “the density of collaboration is approximately ancillary to the other statistics.” This

statement, which admittedly uses the term “ancillary” rather loosely, expresses the idea that the inclusion of the density (number of edges) statistic in the model does not appear to influence the parameter estimates for the other terms in this model.

In summary, in this section we have considered the GWDSP term, unlike HH; we have allowed the θ_s , θ_t , and θ_p parameters to be estimated, unlike SPRH; and we have dropped the restriction in both SPRH and HH that considers only networks with 115 nodes. Yet the overall findings, both qualitative and quantitative, are remarkably similar among the three analyses.

6 Discussion

The class of curved exponential family models is a major generalization of the ERGM model class. Though we have focused here only on very particular models arising from the work of Snijders et al (2006), we have shown how curved exponential family models can achieve parsimonious descriptions of data by reducing a large number of parameters to only a few (e.g., in the case of the GWD term, reducing from $n - 1$ parameters to only two). These reductions also allow much better behavior of maximum likelihood estimation algorithms, which until recently have been plagued by issues of model degeneracy. These models are beginning to show great promise in their ability to fit social network data well. For example, both Hunter et al (2005) and Goodreau (2006) apply these models to networks with more than a thousand nodes.

Finally, we should stress that this is only the beginning. The models discussed here, since they are applied only to cross-sectional data (networks observed only at a single point in time), are static by nature. They assume that the parameter values are unchanging over time, and even the set of nodes is static. Currently, much work is being done on dynamic network models that will describe the evolution of a network over time. Yet quality longitudinal network data sets are scarce, and allowing models to take changes over time into account can lead to an explosion in their complexity. Therefore, curved exponential family models, which can achieve parsimonious yet high-quality fit to data, may have an important role to play in the development of

dynamic network models.

7 Appendix: R code

Here, we give the computer code that produced the two models summarized in Tables 1 and 2. These models are fit using a random algorithm. As such, the results are slightly different each time the algorithm is run. However, repeated runs always produce nearly the same results for all of the models in this article. The essential idea of the algorithm is that we run a Markov chain that produces a random sample from a particular probability distribution on the sample space of all networks. We then use this sample to approximate the true likelihood function in much the same way that a sample mean is often used to approximate a population mean. Finally, the approximated likelihood function is maximized, which yields the parameter estimates. This entire procedure is described in detail in Hunter and Handcock (2006).

The code below is written for the R environment and requires the loading of a package called `ergm` that is part of `statnet`. See csde.washington.edu/statnet for information on `statnet`.

After invoking R and ensuring that the `statnet` package is properly installed, type `library(statnet)`. Then, the following two commands will create R objects called `model1` and `model2` as long as the `lazega` network is correctly in place (naturally, `lazega` may be replaced by another network dataset). To learn more about the `ergm` function used below, type `help(ergm)`. Note that the `ergm` function by default uses a burnin of 1000 Markov chain steps and only samples once every 100 steps; thus, the declaration `MCMCsamplesize=10000` actually results in running a Markov chain for over 1,000,000 steps.

Model from Table 1:

```
formula1 <- lazega ~ edges + nodecov("seniority") +
  nodecov("practice") + match("practice") +
  match("gender") + match("office") +
  gwd(1, fixed=FALSE) + gwesp(1, fixed=FALSE) +
  gwdsp(0, fixed=FALSE)
model1 <- ergm(formula1, MCMCsamplesize=10000)
summary(model1)
```

Model from Table 2:

```
formula2 <- lazega ~ edges + nodecov("seniority") +
  nodecov("practice") + match("practice") +
  match("gender") + match("office") +
  gwesp(1, fixed=FALSE)
model2 <- ergm(formula2, MCMCsamplesize=10000)
summary(model2)
```

References

- Albert, R. and Barabási, A.-L. (2002), Statistical mechanics of complex networks, *Reviews of Modern Physics*, **74**, 47–97.
- Barndorff-Nielsen, O. E. (1978), *Information and Exponential Families in Statistical Theory*, New York: Wiley.
- Brown, L. D. (1986) *Fundamentals of Statistical Exponential Families*, IMS Lecture notes – Monograph series 9.
- Efron, B. (1975), Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion), *Annals of Statistics*, **3**: 1189–1242.
- Efron, B. (1978), The geometry of exponential families, *Annals of Statistics*, **6**: 362–376.
- Goodreau, S. M. (2006), Advances in Exponential Random Graph (p*) Models Applied to a Large Social Network, *Social Networks*, this issue.
- Handcock, M. S. (2002) Statistical Models for Social Networks: Inference and Degeneracy, pp. 229 – 240 in *Dynamic Social Network Modelling and Analysis: Workshop Summary and Papers*, edited by Ronald Breiger, Kathleen Carley, and Philippa E. Pattison. Washington, DC: National Academy Press.
- Handcock, M. S. (2003), Assessing degeneracy in statistical models of social networks, Working Paper no. 39, Center for Statistics and the Social Sciences, University of Washington. Available from <http://www.csss.washington.edu/Papers/>
- Hunter, D. R., S. M. Goodreau, and M. S. Handcock (2005), Goodness of fit of social network models, Penn State Department of Statistics technical report number 05-02. Available from <http://www.stat.psu.edu/reports/2005/>
- Hunter, D. R. and M. S. Handcock (2006), Inference in curved exponential family models for networks, *Journal of Computational and Graphical Statistics*, in press.
- Lazega, E. (2001), *The Collegial Phenomenon : The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership*, Oxford: Oxford University Press.
- Lazega, E. and P. E. Pattison (1999), Multiplexity, generalized exchange and cooperation in organizations: a case study. *Social Networks*, **21**: 67–90.

- Robbins, H. and S. Monro (1951), A stochastic approximation method, *Annals of Mathematical Statistics*, **22**: 400–407.
- Robins, G. L., P. E. Pattison, Y. Kalish, and D. Lusher (2006a), An introduction to exponential random graph (p^*) models for social networks, *Social Networks*, this issue.
- Robins, G. L., T. A. B. Snijders, P. Weng, M. S. Handcock, and P. E. Pattison (2006b), Recent developments in exponential random graph (p^*) models for social networks, *Social Networks*, this issue.
- Snijders, T. A. B. (2002), Markov Chain Monte Carlo estimation of exponential random graph models, *Journal of Social Structure*, **3**. Available at www.cmu.edu/joss/content/articles/volume3/Snijders.pdf
- Snijders, T. A. B., P. E. Pattison, G. L. Robins, and M. S. Handcock (2006), New specifications for exponential random graph models, *Sociological Methodology*, in press.
- Wasserman, S. and P. E. Pattison (1996), Logit models and logistic regression for social networks: I. An introduction to Markov graphs and p^* , *Psychometrika*, **61**: 401–425.