Instability, Sensitivity, and Degeneracy of Discrete Exponential Families

Michael Schweinberger*

Abstract

In applications to dependent data, first and foremost relational data, a number of discrete exponential family models has turned out to be near-degenerate and problematic in terms of Markov chain Monte Carlo simulation and statistical inference. We introduce the notion of instability with an eye to characterize, detect, and penalize discrete exponential family models that are near-degenerate and problematic in terms of Markov chain Monte Carlo simulation and statistical inference. We show that unstable discrete exponential family models are characterized by excessive sensitivity and near-degeneracy. In special cases, the subset of the natural parameter space corresponding to non-degenerate distributions and mean-value parameters far from the boundary of the mean-value parameter space turns out to be a lower-dimensional subspace of the natural parameter space. These characteristics of unstable discrete exponential family models tend to obstruct Markov chain Monte Carlo simulation and statistical inference. In applications to relational data, we show that discrete exponential family models with Markov dependence tend to be unstable and that the parameter space of some curved exponential families contains unstable subsets.

Key words: social networks, statistical exponential families, curved exponential families, undirected graphical models, Markov chain Monte Carlo.

*Email: `michael.schweinberger@stat.psu.edu`. Support is acknowledged from the Netherlands Organisation for Scientific Research (NWO grant 446-06-029), the National Institute of Health (NIH grant 1R01HD052887-01A2), and the Office of Naval Research (ONR grant N00014-08-1-1015). The author is grateful to David Hunter and two anonymous referees for stimulating questions and suggestions.

1 Introduction

We consider discrete exponential families (Barndorff-Nielsen 1978) with emphasis on applications to relational data (Wasserman and Faust 1994). Examples of relational data are social networks, terrorist networks, the world wide web, intra- and inter-organizational networks, trade networks, and cooperation and conflict between nations. A common form of relational data is discrete-valued relationships Y_{ij} between pairs of nodes $i, j = 1, \ldots, n$. Let $\boldsymbol{Y}$ be the collection of relationships Y_{ij} given n nodes and $\mathcal{Y}$ be the sample space of $\boldsymbol{Y}$. Any distribution with support $\mathcal{Y}$ can be expressed in exponential family form (Besag 1974, Frank and Strauss 1986). Discrete exponential families of distributions with support $\mathcal{Y}$ were introduced by Frank and Strauss (1986), Wasserman and Pattison (1996), Snijders et al. (2006), Hunter and Handcock (2006), and others.

In terms of statistical computing, the most important obstacle is the fact that relational data tend to be dependent and discrete exponential families for dependent data come with intractable likelihood functions. Therefore, conventional maximum likelihood and Bayesian algorithms (e.g., Geyer and Thompson 1992, Snijders 2002, Handcock 2002a, Hunter and Handcock 2006, Møller et al. 2006, Koskinen et al. 2010) exploit draws from distributions with support $\mathcal{Y}$ to maximize the likelihood function and explore the posterior distribution, respectively. As Markov chain Monte Carlo (MCMC) is the foremost means to generate draws from distributions with support $\mathcal{Y}$, MCMC is key to both simulation and statistical inference.

In practice, MCMC simulation from discrete exponential family distributions with support $\mathcal{Y}$ has brought to light some serious issues: first, Markov chains may mix extremely slowly and hardly move for millions of iterations (Snijders 2002, Handcock 2003a); and second, the extremely slow mixing of Markov chains may be rooted in the stationary distribution: the stationary distribution may be near-degenerate in the sense of placing almost all probability mass on a small subset of the sample space $\mathcal{Y}$ (Strauss 1986, Jonasson 1999, Snijders 2002, Handcock 2003a, Hunter et al. 2008, Rinaldo et al. 2009). The most troublesome observation, though, is that the subset of the natural parameter space corresponding to non-degenerate distributions may be a negligible subset of the natural parameter space. These troublesome observations raise at least two questions. First, why is the effective natural parameter space of some discrete exponential families (e.g., Frank and Strauss 1986) negligible, while the effective natural parameter space of others (e.g., the Bernoulli model, under which the

Y_{ij} are i.i.d. Bernoulli random variables) is non-negligible? Second, which sufficient statistics can induce such problematic behavior?

Handcock (2002a, 2003a,b) adapted and extended results of Barndorff-Nielsen (1978, pp. 185–186) and pointed out that, as the natural parameters tend to the boundary of the natural parameter space, the probability mass is pushed to the boundary of the convex hull of the space of sufficient statistics (cf. Rinaldo et al. 2009, Geyer 2009, Koskinen et al. 2010). However, these results are applicable to both the Bernoulli model and Frank and Strauss (1986) and neither explain the striking contrast between them nor clarify which sufficient statistics can induce problematic behavior.

We introduce the notion of instability along the lines of statistical physics (Ruelle 1969) with an eye to characterize, detect, and penalize problematic discrete exponential families. Strauss (1986) was the first to observe that the problematic behavior of the discrete exponential families of Frank and Strauss (1986) is related to lack of stability of point processes in statistical physics (Ruelle 1969, p. 33). We adapt the notion of stability of point processes in the sense of Ruelle (1969, p. 33) to discrete exponential families and introduce the notions of unstable discrete exponential family distributions and unstable sufficient statistics. We show that unstable exponential family distributions are characterized by excessive sensitivity and near-degeneracy. In special cases, the subset of the natural parameter space corresponding to non-degenerate distributions and mean-value parameters far from the boundary of the mean-value parameter space turns out to be a lower-dimensional subspace of the natural parameter space. In applications to relational data, it turns out that the parameter space of exponential families with Markov dependence (Frank and Strauss 1986) tends to be unstable and that the parameter space of some curved exponential families (Snijders et al. 2006, Hunter and Handcock 2006) contains unstable subsets.

We introduce the notion of instability and its implications in Section 2, discuss its impact on MCMC simulation and statistical inference in Sections 3 and 4, respectively, and present applications to relational data and simulation results in Sections 5 and 6, respectively.

2 Instability, sensitivity, and degeneracy

Let $\boldsymbol{Y}_N$ be a discrete random variable with sample space $\mathcal{Y}_N = \mathcal{Z}^N$, where $\mathcal{Z}$ is a discrete set of M elements and N is the number of degrees of freedom. In applications to relational data (Wasserman and Faust 1994), $\boldsymbol{Y}_N$ may correspond to $N \leq n^2$ relationships among n nodes; in applications to spatial data (Besag 1974), N random variables located at N sites of a lattice; and in binomial sampling, N i.i.d. Bernoulli random variables.

We consider discrete exponential families of distributions $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ with probability mass functions of the form

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}_N) = \exp\left[\boldsymbol{\eta}_N^T(\boldsymbol{\theta})\, \boldsymbol{g}_N(\boldsymbol{y}_N) - \psi_N(\boldsymbol{\theta})\right],\ \boldsymbol{y}_N \in \mathcal{Y}_N, \tag{1}$$

where $\boldsymbol{\eta}_N : \Theta \mapsto \mathbb{R}^L$ is a vector of natural parameters and $\boldsymbol{g}_N : \mathcal{Y}_N \mapsto \mathbb{R}^L$ is a vector of sufficient statistics,

$$\psi_N(\boldsymbol{\theta}) = \log \sum_{\boldsymbol{x}_N \in \mathcal{Y}_N} \exp\left[\boldsymbol{\eta}_N^T(\boldsymbol{\theta})\, \boldsymbol{g}_N(\boldsymbol{x}_N)\right],\ \boldsymbol{\theta} \in \Theta \tag{2}$$

is the cumulant generating function, and $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^K : \psi_N(\boldsymbol{\theta}) < \infty\}$ is the parameter space. The vector of natural parameters $\boldsymbol{\eta}_N(\boldsymbol{\theta})$ may be a linear or non-linear function of parameter vector $\boldsymbol{\theta}$. If $\boldsymbol{\eta}_N(\boldsymbol{\theta}) = \mathbf{A}_N^T\, \boldsymbol{\theta}$ is a linear function of $\boldsymbol{\theta}$, where $\mathbf{A}_N$ is a $K \times L$ matrix, the non-uniqueness of the canonical form of exponential families can be exploited to absorb $\mathbf{A}_N$ into $\boldsymbol{g}_N(\boldsymbol{y}_N)$, so that $\boldsymbol{\eta}_N(\boldsymbol{\theta}) = \boldsymbol{\theta}$ can be assumed without loss of generality. If $\boldsymbol{\eta}_N(\boldsymbol{\theta})$ is a non-linear function of $\boldsymbol{\theta}$ and $K < L$, the exponential family is curved (Efron 1978).

Let $q_{\boldsymbol{\theta}}(\boldsymbol{y}_N) = \boldsymbol{\eta}_N^T(\boldsymbol{\theta})\, \boldsymbol{g}_N(\boldsymbol{y}_N)$, and $I_N(\boldsymbol{\theta}) = \min_{\boldsymbol{y}_N \in \mathcal{Y}_N}[q_{\boldsymbol{\theta}}(\boldsymbol{y}_N)]$ and $S_N(\boldsymbol{\theta}) = \max_{\boldsymbol{y}_N \in \mathcal{Y}_N}[q_{\boldsymbol{\theta}}(\boldsymbol{y}_N)]$ be the minimum and maximum of $q_{\boldsymbol{\theta}}(\boldsymbol{y}_N)$, respectively. Since $p_{\boldsymbol{\theta}}(\boldsymbol{y}_N)$ is invariant to translations of $q_{\boldsymbol{\theta}}(\boldsymbol{y}_N)$ by $-I_N(\boldsymbol{\theta})$, let $I_N(\boldsymbol{\theta}) = 0$ without loss of generality.

Definition: stable, unstable distributions. A discrete exponential family distribution $P_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta$, is stable if there exist constants $C > 0$ and $N_C > 0$ such that

$$S_N(\boldsymbol{\theta}) \leq CN \text{ for all } N > N_C, \tag{3}$$

and unstable if, for any $C > 0$, however large, there exists $N_C > 0$ such that

$$S_N(\boldsymbol{\theta}) > CN \text{ for all } N > N_C. \square \tag{4}$$

In general, instability may be induced by $\boldsymbol{\eta}_N(\boldsymbol{\theta})$ or $\boldsymbol{g}_N(\boldsymbol{y}_N)$. In the important special case where $\boldsymbol{\eta}_N(\boldsymbol{\theta})$ is a linear function of $\boldsymbol{\theta}$, in which case $\boldsymbol{\eta}_N(\boldsymbol{\theta}) = \boldsymbol{\theta}$ can be assumed without loss of generality, $\boldsymbol{g}_N(\boldsymbol{y}_N)$ is the exclusive source of instability. Let $\eta_{N,k}(\boldsymbol{\theta})$ and $g_{N,k}(\boldsymbol{y}_N)$ be the k-th coordinate of $\boldsymbol{\eta}_N(\boldsymbol{\theta})$ and $\boldsymbol{g}_N(\boldsymbol{y}_N)$, respectively, $L_{N,k} = \min_{\boldsymbol{y}_N \in \mathcal{Y}_N}[g_{N,k}(\boldsymbol{y}_N)]$ and $U_{N,k} = \max_{\boldsymbol{y}_N \in \mathcal{Y}_N}[g_{N,k}(\boldsymbol{y}_N)]$ be the minimum and maximum of $g_{N,k}(\boldsymbol{y}_N)$, respectively, and $L_{N,k} = 0$ without loss of generality, owing to the invariance of $p_{\boldsymbol{\theta}}(\boldsymbol{y}_N)$ to translations of $q_{\boldsymbol{\theta}}(\boldsymbol{y}_N)$ by $-\eta_{N,k}(\boldsymbol{\theta})\, L_{N,k}$ $(k = 1, \ldots, L)$.

Definition: stable, unstable sufficient statistics. A sufficient statistic $g_{N,k}(\boldsymbol{y}_N)$ is stable if there exists constants $C > 0$ and $N_C > 0$ such that

$$U_{N,k} \leq CN \text{ for all } N > N_C, \tag{5}$$

and unstable if, for any $C > 0$, however large, there exists $N_C > 0$ such that

$$U_{N,k} > CN \text{ for all } N > N_C.\square \tag{6}$$

While the notion of unstable discrete exponential families holds intuitive appeal, the parameter space Θ of most discrete exponential families of interest includes subsets indexing stable distributions. With a wide range of applications in mind, it is therefore preferable to study the characteristics of unstable sufficient statistics and unstable distributions and to detect in applications unstable sufficient statistics and subsets of Θ indexing unstable distributions. It is worthwhile to note that Handcock (2002a,b, 2003a) discussed an alternative, but unrelated notion of stability, calling discrete exponential families stable if small changes in natural parameters result in small changes of the probability mass function.

To demonstrate instability and its implications, we introduce two classic examples in Section 2.1. In Sections 2.2 and 2.3, we show that unstable exponential family distributions are characterized by excessive sensitivity and near-degeneracy.

2.1 Examples

A simple but common form of relational data is undirected graphs $\boldsymbol{y}_N$, where the relationships $y_{ij} \in \{0, 1\}$ satisfy the linear constraints $y_{ij} = y_{ji}$ (all $i < j$) and $y_{ii} = 0$ (all i), which reduces the number of degrees of freedom N from n^2 to $n(n-1)/2$. Two classic models of undirected graphs are the Bernoulli model with natural parameter θ and stable sufficient statistic $\sum_{i<j} y_{ij}$ and the 2-star model with natural parameter θ and unstable sufficient statistic $\sum_{i,\, j<k} y_{ij} y_{ik}$. The Bernoulli model arises from the

assumption that the random variables Y_{ij} are i.i.d. Bernoulli (all $i < j$), while the 2-star model can be motivated by Markov dependence (Frank and Strauss 1986). The Bernoulli model implies $S_N(\theta) = |\theta|\, N$ and is therefore stable for all θ, while the 2-star model implies $S_N(\theta) = |\theta|\,(n-2)\, N$ and is therefore unstable for all $\theta \neq 0$.

2.2 Instability and sensitivity

Unstable discrete exponential family distributions are characterized by excessive sensitivity.

Consider the smallest possible changes of $\boldsymbol{y}_N$, that is, changes of one element of $\boldsymbol{y}_N$, and let

$$\Lambda_N(\boldsymbol{x}_N, \boldsymbol{y}_N; \boldsymbol{\theta}) = \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{y}_N)}{p_{\boldsymbol{\theta}}(\boldsymbol{x}_N)},\ \boldsymbol{x}_N \sim \boldsymbol{y}_N,\ \boldsymbol{x}_N,\ \boldsymbol{y}_N \in \mathcal{Y}_N \tag{7}$$

be the log odds of $p_{\boldsymbol{\theta}}(\boldsymbol{y}_N)$ relative to $p_{\boldsymbol{\theta}}(\boldsymbol{x}_N)$, where $\boldsymbol{x}_N \sim \boldsymbol{y}_N$ means that $\boldsymbol{x}_N$ and $\boldsymbol{y}_N$ are nearest neighbors in the sense that $\boldsymbol{x}_N$ and $\boldsymbol{y}_N$ match in all but one element. The following theorem shows that, if an exponential family distribution is unstable, then the probability mass function is characterized by excessive sensitivity in the sense that the nearest neighbor log odds are unbounded and therefore even the smallest possible changes can result in extremely large log odds.

Theorem 1. If a discrete exponential family distribution $P_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta$, is unstable, then there exist no constants $C > 0$ and $N_C > 0$ such that

$$|\Lambda_N(\boldsymbol{x}_N, \boldsymbol{y}_N; \boldsymbol{\theta})| \leq C \text{ for all } \boldsymbol{x}_N \sim \boldsymbol{y}_N,\ \boldsymbol{x}_N,\ \boldsymbol{y}_N \in \mathcal{Y}_N \text{ for all } N > N_C. \square \tag{8}$$

Theorem 1 implies that some, but not necessarily all, nearest neighbor log odds are unbounded. It indicates that the probability mass function is excessively sensitive to small changes in subsets of $\mathcal{Y}_N$ and that some elements of $\mathcal{Y}_N$ dominate others in terms of probability mass. A walk through $\mathcal{Y}_N$ resembles a walk through a rugged, mountainous landscape: small steps in $\mathcal{Y}_N$ can result in dramatic increases or decreases in probability mass. An example is given by the 2-star model of Section 2.1: for all $\theta \neq 0$, the nearest neighbor log odds satisfy $|\Lambda_N(\boldsymbol{x}_N, \boldsymbol{y}_N; \theta)| \leq 2\,|\theta|\,(n-2)$ (all $\boldsymbol{x}_N \sim \boldsymbol{y}_N$) and are therefore $O(n)$. The excessive sensitivity of the 2-star model is well-known (Handcock 2003a), but Theorem 1 indicates that all unstable exponential family distributions suffer from excessive sensitivity.

Section 3 shows that the unbounded nearest neighbor log odds of unstable exponential family distributions have a direct impact on MCMC simulation.

2.3 Instability and degeneracy

Discrete exponential family distributions with support $\mathcal{Y}_N$ cannot be degenerate in the strict sense of the word. However, unstable discrete exponential family distributions turn out to be near-degenerate. Worse, in the important special case of discrete exponential families with unstable sufficient statistics, the subset of the natural parameter space corresponding to non-degenerate distributions turns out to be a lower-dimensional subspace of the natural parameter space.

Let $\mathcal{M}_N = \{\boldsymbol{y}_N \in \mathcal{Y}_N : q_{\boldsymbol{\theta}}(\boldsymbol{y}_N) = S_N(\boldsymbol{\theta})\}$ be the subset of modes and, for any $0 < \epsilon < 1$, let $\mathcal{M}_{\epsilon,N} = \{\boldsymbol{y}_N \in \mathcal{Y}_N : q_{\boldsymbol{\theta}}(\boldsymbol{y}_N) > (1-\epsilon)\, S_N(\boldsymbol{\theta})\}$ be the subset of ϵ-modes of the probability mass function $p_{\boldsymbol{\theta}}(\boldsymbol{y}_N)$. The following theorem shows that unstable exponential family distributions tend to concentrate almost all probability mass on the modes of the probability mass function.

Theorem 2. If a discrete exponential family distribution $P_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta$, is unstable, then it is degenerate in the sense that, for any $0 < \epsilon < 1$, however small,

$$P_{\boldsymbol{\theta}}(\boldsymbol{Y}_N \in \mathcal{M}_{\epsilon,N}) \longrightarrow 1 \text{ as } N \longrightarrow \infty. \square \tag{9}$$

A related result was reported by Strauss (1986) and Handcock (2003a). In general, the fact that almost all probability mass tends to be concentrated on the modes of the probability mass function is troublesome: first, because the effective support, the subset of the support $\mathcal{Y}_N$ with non-negligible probability mass, is reduced; and second, because in most applications the modes do not resemble observed data.

In the important special case of exponential families with unstable sufficient statistics, it is possible to gain more insight into near-degeneracy. Consider one-parameter exponential families $\{P_\theta, \theta \in \Theta\}$ with natural parameter $\eta_N(\theta) = \theta$ and sufficient statistic $g_N(\boldsymbol{y}_N)$. Let $L_N = 0$ (without loss of generality) and U_N be the minimum and maximum of $g_N(\boldsymbol{y}_N)$, respectively, and, for any $0 < \epsilon < 1$, let $\mathcal{L}_{\epsilon,N} = \{\boldsymbol{y}_N \in \mathcal{Y}_N : g_N(\boldsymbol{y}_N) < \epsilon\, U_N\}$ and $\mathcal{U}_{\epsilon,N} = \{\boldsymbol{y}_N \in \mathcal{Y}_N : g_N(\boldsymbol{y}_N) > (1-\epsilon)\, U_N\}$ be the subset of the sample space $\mathcal{Y}_N$ close to the minimum and maximum of $g_N(\boldsymbol{y}_N)$, respectively. The following result shows that one-parameter exponential families with unstable sufficient statistics $g_N(\boldsymbol{y}_N)$ tend to be degenerate with respect to $g_N(\boldsymbol{y}_N)$.

Theorem 3. A one-parameter exponential family $\{P_\theta, \theta \in \Theta\}$ with natural parameter θ and unstable sufficient statistic $g_N(\boldsymbol{y}_N)$ is degenerate with respect to $g_N(\boldsymbol{y}_N)$ in the sense that, for any $0 < \epsilon < 1$, however small, and for any $\theta < 0$,

$$P_\theta(\boldsymbol{Y}_N \in \mathcal{L}_{\epsilon,N}) \longrightarrow 1 \text{ as } N \longrightarrow \infty \tag{10}$$

and, for any $\theta > 0$,

$$P_\theta(\boldsymbol{Y}_N \in \mathcal{U}_{\epsilon,N}) \longrightarrow 1 \text{ as } N \longrightarrow \infty.\square \tag{11}$$

Thus, the probability mass is pushed to the minimum of $g_N(\boldsymbol{y}_N)$ for all $\theta < 0$ and the maximum of $g_N(\boldsymbol{y}_N)$ for all $\theta > 0$, and the subset of the natural parameter space Θ corresponding to non-degenerate distributions is a lower-dimensional subspace of Θ: the point $\theta = 0$. An example of a one-parameter exponential family with unstable sufficient statistic is given by the 2-star model of Section 2.1.

Consider K-parameter exponential families $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ with natural parameters $\eta_{N,1}(\boldsymbol{\theta}) = \theta_1, \ldots, \eta_{N,K}(\boldsymbol{\theta}) = \theta_K$ and $K-1$ stable sufficient statistics $g_{N,1}(\boldsymbol{y}_N), \ldots, g_{N,K-1}(\boldsymbol{y}_N)$ as well as one unstable sufficient statistic $g_{N,K}(\boldsymbol{y}_N)$. In accordance with the preceding paragraph, let $\mathcal{L}_{\epsilon,N,K}$ and $\mathcal{U}_{\epsilon,N,K}$ be the subset of the sample space $\mathcal{Y}_N$ close to the minimum and maximum of the unstable sufficient statistic $g_{N,K}(\boldsymbol{y}_N)$, respectively. The following result shows that K-parameter exponential families with $K-1$ stable and one unstable sufficient statistic tend to be degenerate with respect to the unstable sufficient statistic.

Theorem 4. A K-parameter exponential family $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ with natural parameters $\theta_1, \ldots, \theta_K$ and $K-1$ stable sufficient statistics $g_{N,1}(\boldsymbol{y}_N), \ldots, g_{N,K-1}(\boldsymbol{y}_N)$ as well as one unstable sufficient statistic $g_{N,K}(\boldsymbol{y}_N)$ is degenerate with respect to $g_{N,K}(\boldsymbol{y}_N)$ in the sense that, for any $0 < \epsilon < 1$, however small, and for any $\theta_K < 0$,

$$P_{\boldsymbol{\theta}}(\boldsymbol{Y}_N \in \mathcal{L}_{\epsilon,N,K}) \longrightarrow 1 \text{ as } N \longrightarrow \infty \tag{12}$$

and, for any $\theta_K > 0$,

$$P_{\boldsymbol{\theta}}(\boldsymbol{Y}_N \in \mathcal{U}_{\epsilon,N,K}) \longrightarrow 1 \text{ as } N \longrightarrow \infty.\square \tag{13}$$

In general, it is not straightforward to see where the probability mass of K-parameter exponential families with multiple unstable sufficient statistics ends up. In special cases, though, insight can be gained. Consider a K-parameter exponential family $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ with natural parameters $\theta_1, \ldots, \theta_K$ and sufficient statistics $g_{N,1}(\boldsymbol{y}_N), \ldots, g_{N,K}(\boldsymbol{y}_N)$, where $g_{N,1}(\boldsymbol{y}_N), \ldots, g_{N,K-1}(\boldsymbol{y}_N)$ may be unstable while $g_{N,K}(\boldsymbol{y}_N)$ is unstable and dominates $g_{N,1}(\boldsymbol{y}_N), \ldots, g_{N,K-1}(\boldsymbol{y}_N)$ in the sense that, for any $D > 0$, however large, there exists $N_D > 0$ such that

$$\frac{U_{N,K}}{U_{N,k}} > D \text{ for all } N > N_D,\ k = 1, \ldots, K-1. \tag{14}$$

A K-parameter exponential family with multiple unstable sufficient statistics, including an unstable, dominating sufficient statistic $g_{N,K}(\boldsymbol{y}_N)$, tends to be degenerate with respect to $g_{N,K}(\boldsymbol{y}_N)$.

Theorem 5. A K-parameter exponential family $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ with natural parameters $\theta_1, \ldots, \theta_K$ and sufficient statistics $g_{N,1}(\boldsymbol{y}_N)$, $\ldots$, $g_{N,K}(\boldsymbol{y}_N)$, where $g_{N,1}(\boldsymbol{y}_N)$, $\ldots$, $g_{N,K-1}(\boldsymbol{y}_N)$ may be unstable while $g_{N,K}(\boldsymbol{y}_N)$ is unstable and dominates $g_{N,1}(\boldsymbol{y}_N)$, $\ldots$, $g_{N,K-1}(\boldsymbol{y}_N)$, is degenerate with respect to $g_{N,K}(\boldsymbol{y}_N)$ in the sense that, for any $0 < \epsilon < 1$, however small, and for any $\theta_K < 0$,

$$P_{\boldsymbol{\theta}}(\boldsymbol{Y}_N \in \mathcal{L}_{\epsilon,N,K}) \longrightarrow 1 \text{ as } N \longrightarrow \infty \tag{15}$$

and, for any $\theta_K > 0$,

$$P_{\boldsymbol{\theta}}(\boldsymbol{Y}_N \in \mathcal{U}_{\epsilon,N,K}) \longrightarrow 1 \text{ as } N \longrightarrow \infty. \square \tag{16}$$

It is worthwhile to point out that whether most probability mass tends to be concentrated on one element of the sample space $\mathcal{Y}_N$ and the entropy of the distribution tends to 0 depends on the sufficient statistics. An exponential family that is degenerate with respect to sufficient statistics is as degenerate as it can be.

As we will see in Section 4, the degeneracy of exponential families with unstable sufficient statistics tends to push the mean-value parameters to the boundary of the mean-value parameter space, which tends to obstruct statistical inference.

3 Impact of instability on MCMC simulation

If a Markov chain with unstable stationary distribution is constructed by MCMC methods, the excessive sensitivity and near-degeneracy of the stationary distribution tend to have a direct impact on MCMC simulation.

The excessive sensitivity of unstable stationary distributions, excessive in the sense that the nearest neighbor log odds are unbounded, affects the probabilities of transition between nearest neighbors: e.g., in applications to undirected graphs (cf. Section 2.1), Gibbs samplers sample elements y_{ij} from full conditional distributions of the form

$$Y_{ij} \mid \boldsymbol{y}_{-ij} \sim \text{Bernoulli}(\pi_{ij}(\boldsymbol{y}_{-ij}; \boldsymbol{\theta})), \tag{17}$$

where $\boldsymbol{y}_{-ij}$ denotes the collection of elements $\boldsymbol{y}_N$ excluding y_{ij}, and the log odds of

$\pi_{ij}(\boldsymbol{y}_{-ij};\boldsymbol{\theta})$ is given by

$$\log \frac{\pi_{ij}(\boldsymbol{y}_{-ij};\boldsymbol{\theta})}{1-\pi_{ij}(\boldsymbol{y}_{-ij};\boldsymbol{\theta})} = \Lambda_N(\{\boldsymbol{y}_{-ij}, y_{ij}=0\}, \{\boldsymbol{y}_{-ij}, y_{ij}=1\};\boldsymbol{\theta}). \tag{18}$$

A Metropolis-Hastings algorithm moves from $\boldsymbol{x}_N$ to $\boldsymbol{y}_N$, generated from a probability mass function f with support $\{\boldsymbol{y}_N : \boldsymbol{y}_N \sim \boldsymbol{x}_N\}$, with probability

$$\alpha(\boldsymbol{x}_N, \boldsymbol{y}_N;\boldsymbol{\theta}) = \min\left\{1,\, \exp[\Lambda_N(\boldsymbol{x}_N, \boldsymbol{y}_N;\boldsymbol{\theta})]\, \frac{f(\boldsymbol{x}_N \mid \boldsymbol{y}_N)}{f(\boldsymbol{y}_N \mid \boldsymbol{x}_N)}\right\}. \tag{19}$$

Since the nearest neighbor log odds satisfy $\Lambda_N(\boldsymbol{x}_N, \boldsymbol{y}_N;\boldsymbol{\theta}) = -\Lambda_N(\boldsymbol{y}_N, \boldsymbol{x}_N;\boldsymbol{\theta})$ (all $\boldsymbol{x}_N \sim \boldsymbol{y}_N$) and are unbounded by Theorem 1, Markov chains with unstable stationary distributions can move extremely fast from some subsets of the sample space $\mathcal{Y}_N$ to other subsets and extremely slowly back. In addition, if the mode of the probability mass function is not unique, multiple Markov chains may be required, because Theorems 1 and 2 indicate that one Markov chain may be trapped at one of the modes. Worse, Theorems 3 and 4 suggest that MCMC simulation from exponential families with unstable sufficient statistics may be a waste of time and resources in the first place.

The most important conclusion, though, is that mixing problems of MCMC algorithms tend to be rooted in the unstable stationary distribution rather than the design of the MCMC algorithms, as is evident from the unbounded nearest neighbor log odds and the near-degeneracy of unstable stationary distributions. A related result and conclusion was reported by Handcock (2003a).

4 Impact of instability on statistical inference

The degeneracy of exponential families with unstable sufficient statistics tends to push the mean-value parameters to the boundary of the mean-value parameter space and therefore tends to obstruct maximum likelihood estimation.

Let $\mu_N : \Theta \mapsto \text{int}(\mathcal{C}_N)$ be the map from parameter space Θ to the mean-value parameter space $\text{int}(\mathcal{C}_N)$ (Barndorff-Nielsen 1978, p. 121) given by

$$\mu_N(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[g_N(\boldsymbol{Y}_N)] \in \text{int}(\mathcal{C}_N), \tag{20}$$

where $\text{int}(\mathcal{C}_N)$ denotes the interior of the convex hull $\mathcal{C}_N$ of $\{g_N(\boldsymbol{y}_N) : \boldsymbol{y}_N \in \mathcal{Y}_N\}$.

We start with one-parameter exponential families $\{P_\theta, \theta \in \Theta\}$ with natural parameter θ and unstable sufficient statistic $g_N(\boldsymbol{y}_N)$. Let $L_N = 0$ (without loss of generality) and U_N be the minimum and maximum of $g_N(\boldsymbol{y}_N)$, respectively, and

$$\frac{\mu_N(\theta)}{U_N} = \frac{E_\theta[g_N(\boldsymbol{Y}_N)]}{U_N} \in (0,1) \tag{21}$$

be the mean-value parameter, where re-scaling by $1/U_N$ ensures that the range of $\mu_N(\theta)/U_N$ is $(0,1)$. The following result shows that one-parameter exponential families with unstable sufficient statistics $g_N(\boldsymbol{y}_N)$ push the mean-value parameter $\mu_N(\theta)$ to its infinum for all $\theta < 0$ and its supremum for all $\theta > 0$.

Corollary 1. The mean-value parameter $\mu_N(\theta)$ of a one-parameter exponential family $\{P_\theta, \theta \in \Theta\}$ with natural parameter θ and unstable sufficient statistic $g_N(\boldsymbol{y}_N)$ tends to the boundary of the mean-value parameter space in the sense that, for any $\theta < 0$, however small,

$$\frac{\mu_N(\theta)}{U_N} \longrightarrow 0 \text{ as } N \longrightarrow \infty \tag{22}$$

and, for any $\theta > 0$, however small,

$$\frac{\mu_N(\theta)}{U_N} \longrightarrow 1 \text{ as } N \longrightarrow \infty. \square \tag{23}$$

By Corollary 1, the subset of the natural parameter space Θ corresponding to mean-value parameters far from the boundary of the mean-value parameter space tends to be a lower-dimensional subpace of Θ: the point $\theta = 0$. In addition, the mean-value parameter $\mu_N(\theta)$ can be expected to be extremely sensitive to changes of the natural parameter θ around 0.

The relationship between the natural parameter θ and the mean-value parameter $\mu_N(\theta)$ is problematic in terms of maximum likelihood estimation. If $g_N(\boldsymbol{y}_N) \in \text{int}(\mathcal{C}_N)$ denotes an observation in the interior of $\mathcal{C}_N$, the maximum likelihood estimate of θ exists and is unique (Barndorff-Nielsen 1978, p. 150) and is given by the root of the estimating function

$$\delta_N(\theta) = g_N(\boldsymbol{y}_N) - E_\theta[g_N(\boldsymbol{Y}_N)] = g_N(\boldsymbol{y}_N) - \mu_N(\theta). \tag{24}$$

The estimating function $\delta_N(\theta)$ depends on θ through $\mu_N(\theta)$, and since $\mu_N(\theta)$ tends to be extremely sensitive to changes of θ around 0, so does $\delta_N(\theta)$. If the observation $g_N(\boldsymbol{y}_N)$ is not close to the boundary of $\mathcal{C}_N$, the maximum likelihood estimate of θ tends to be close to 0, since only values of θ close to 0 map to values of $\mu_N(\theta)$ which

are not close to the boundary of $\mathcal{C}_N$. As a result, maximum likelihood algorithms tend to search for the maximum likelihood estimate of θ in a small neighborhood of 0, but are hampered by the extreme sensitivity of the estimating function $\delta_N(\theta)$ around $\theta = 0$ and tend to make small steps in the natural parameter space Θ around $\theta = 0$ and large steps in the mean-value parameter space $\text{int}(\mathcal{C}_N)$ and struggle to converge. A related result and conclusion was reported by Handcock (2003a).

The behavior of K-parameter exponential families $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ with natural parameters $\theta_1, \ldots, \theta_K$ and $K-1$ stable sufficient statistics $g_{N,1}(\boldsymbol{y}_N)$, $\ldots$, $g_{N,K-1}(\boldsymbol{y}_N)$ as well as one unstable sufficient statistic $g_{N,K}(\boldsymbol{y}_N)$ resembles the behavior of one-parameter exponential families with unstable sufficient statistic $g_{N,K}(\boldsymbol{y}_N)$. Let $L_{N,K} = 0$ (without loss of generality) and $U_{N,K}$ be the minimum and maximum of the unstable sufficient statistic $g_{N,K}(\boldsymbol{y}_N)$, respectively, and

$$\frac{\mu_{N,K}(\boldsymbol{\theta})}{U_{N,K}} = \frac{E_{\boldsymbol{\theta}}[g_{N,K}(\boldsymbol{Y}_N)]}{U_{N,K}} \in (0,1) \tag{25}$$

be the coordinate of the vector of mean-value parameters $\mu_N(\boldsymbol{\theta})$ corresponding to $g_{N,K}(\boldsymbol{y}_N)$.

Corollary 2. The vector of mean-value parameters $\mu_N(\boldsymbol{\theta})$ of a K-parameter exponential family $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ with natural parameters $\theta_1, \ldots, \theta_K$ and $K-1$ stable sufficient statistics $g_{N,1}(\boldsymbol{y}_N), \ldots, g_{N,K-1}(\boldsymbol{y}_N)$ as well as one unstable sufficient statistic $g_{N,K}(\boldsymbol{y}_N)$ tends to the boundary of the mean-value parameter space in the sense that, for any $\theta_K < 0$, however small,

$$\frac{\mu_{N,K}(\boldsymbol{\theta})}{U_{N,K}} \longrightarrow 0 \text{ as } N \longrightarrow \infty \tag{26}$$

and, for any $\theta_K > 0$, however small,

$$\frac{\mu_{N,K}(\boldsymbol{\theta})}{U_{N,K}} \longrightarrow 1 \text{ as } N \longrightarrow \infty. \square \tag{27}$$

To conclude, while some maximum likelihood algorithms may outperform others, Corollaries 1 and 2 indicate that all maximum likelihood algorithms can be expected to suffer from degeneracy with respect to sufficient statistics (cf. Handcock 2003a, Rinaldo et al. 2009).

5 Applications to relational data

The intention of the present section is to detect unstable subsets of the parameter space of discrete exponential families, because unstable discrete exponential family

distributions are characterized by excessive sensitivity and near-degeneracy (cf. Section 2), which tends to obstruct MCMC simulation (cf. Section 3) as well as statistical inference (cf. Section 4).

We focus on applications to relational data, but note that in applications to lattice systems (Besag 1974) and binomial sampling, exponential family models (with suitable neighborhood assumptions) tend to be stable (Ruelle 1969). We consider undirected graphs and the most widely used exponential family models of undirected graphs, so-called exponential family random graph models (ERGMs) with Markov dependence and curved exponential family random graph models (curved ERGMs). It is worthwhile to note that the number of degrees of freedom N is $O(n^2)$ and is therefore large even when the number of nodes n is small, suggesting that the large-N results of Sections 2–4 shed light on the behavior of ERGMs even when n is not large.

A simple and appealing class of ERGMs with Markov dependence (Frank and Strauss 1986) is given by

$$q_{\boldsymbol{\theta}}(\boldsymbol{y}_N) \;=\; \sum_{k=1}^{n-1} \eta_{N,k}(\boldsymbol{\theta})\, s_{N,k}(\boldsymbol{y}_N) + \eta_{N,n}(\boldsymbol{\theta}) \sum_{i<j<k} y_{ij} y_{jk} y_{ik}, \tag{28}$$

where $s_{N,k}(\boldsymbol{y}_N) = \sum_{i,\, j_1<\cdots<j_k} y_{ij_1} \cdots y_{ij_k}$ is the number of k-stars ($k = 1, \ldots, n-1$) and $\sum_{i<j<k} y_{ij} y_{jk} y_{ik}$ is the number of triangles. Since the number of natural parameters of (28) is n, it is common to impose linear or non-linear constraints on the natural parameters of (28) with an eye to reduce the number of parameters to be estimated. The following ERGMs are special cases of (28) obtained by imposing suitable linear constraints on the natural parameters of (28).

Result 1. ERGMs with 2-star terms of the form

$$q_{\boldsymbol{\theta}}(\boldsymbol{y}_N) \;=\; \theta_1 \sum_{i<j} y_{ij} + \theta_2 \sum_{i,\, j<k} y_{ij} y_{ik} \tag{29}$$

are unstable for all $\theta_2 \neq 0$.□

Result 2. ERGMs with triangle terms of the form

$$q_{\boldsymbol{\theta}}(\boldsymbol{y}_N) \;=\; \theta_1 \sum_{i<j} y_{ij} + \theta_2 \sum_{i<j<k} y_{ij} y_{jk} y_{ik} \tag{30}$$

are unstable for all $\theta_2 \neq 0$.□

Results 1 and 2 are in line with existing results: both ERGMs are known to be near-degenerate and problematic in terms of MCMC simulation and statistical inference (Strauss 1986, Jonasson 1999, Snijders 2002, Handcock 2003a, Rinaldo et al.

2009). The most striking conclusion is that in both cases the subset of the natural parameter space $\mathbb{R}^2$ corresponding to non-degenerate distributions is a lower-dimensional subspace of $\mathbb{R}^2$: the line $(\theta_1, 0)$. In terms of MCMC, the nearest neighborhood log odds $|\Lambda_N(\boldsymbol{x}_N, \boldsymbol{y}_N; \boldsymbol{\theta})|$ are $O(n)$, which suggests that MCMC algorithms tend to suffer from extremely slow mixing, as is well-known (Snijders 2002, Handcock 2002a, 2003a).

To reduce the problematic behavior of ERGMs of the form (29) and (30), it has sometimes been suggested to counterbalance positive instability-inducing terms by negative instability-inducing terms.

Result 3. ERGMs with 2-star and triangle terms of the form

$$q_{\boldsymbol{\theta}}(\boldsymbol{y}_N) \;=\; \theta_1 \sum_{i<j} y_{ij} + \theta_2 \sum_{i,\,j<k} y_{ij} y_{ik} + \theta_3 \sum_{i<j<k} y_{ij} y_{jk} y_{ik} \tag{31}$$

are unstable for all θ_2 and θ_3 excluding $\theta_2 = \theta_3 = 0$ and $\theta_2 = -\theta_3/3$.□

Result 3 demonstrates that counterbalancing instability-inducing terms does not, in general, work: the subset of $\mathbb{R}^3$ corresponding to non-degenerate distributions is severely constrained by the linear constraints $\theta_2 = \theta_3 = 0$ and $\theta_2 = -\theta_3/3$.

We turn to the curved ERGMs of Snijders et al. (2006) and Hunter and Handcock (2006), which were motivated by the problematic behavior of ERGMs with Markov dependence. Three of the best-known curved ERGM terms are geometrically weighted degree (GWD), geometrically weighted dyadwise shared partner (GWDSP), and geometrically weighted edgewise shared partner (GWESP) terms (cf. Hunter et al. 2008).

Result 4. Curved ERGMs with GWD terms of the form

$$q_{\boldsymbol{\theta}}(\boldsymbol{y}_N) = \theta_1 \sum_{i<j} y_{ij} + \theta_2 \exp[\theta_3] \sum_{k=1}^{n-1} \left[1 - (1 - \exp[-\theta_3])^k\right] D_{N,k}(\boldsymbol{y}_N), \tag{32}$$

where $D_{N,k}(\boldsymbol{y}_N)$ is the number of nodes i with degree $\sum_{j \neq i} y_{ij} = k$, are unstable for all $\theta_2 \neq 0$ and $\theta_3 < -\log 2$.□

Result 5. Curved ERGMs with GWDSP terms of the form

$$q_{\boldsymbol{\theta}}(\boldsymbol{y}_N) = \theta_1 \sum_{i<j} y_{ij} + \theta_2 \exp[\theta_3] \sum_{k=1}^{n-2} \left[1 - (1 - \exp[-\theta_3])^k\right] DSP_{N,k}(\boldsymbol{y}_N), \tag{33}$$

where $DSP_{N,k}(\boldsymbol{y}_N)$ is the number of pairs of nodes $\{i, j\}$ with $\sum_{h \neq i,j} y_{ih} y_{jh} = k$ dyadwise shared partners, are unstable for all $\theta_2 \neq 0$ and $\theta_3 < -\log 2$.□

Result 6. Curved ERGMs with GWESP terms of the form

$$q_{\boldsymbol{\theta}}(\boldsymbol{y}_N) = \theta_1 \sum_{i<j} y_{ij} + \theta_2 \exp[\theta_3] \sum_{k=1}^{n-2} \left[1 - (1 - \exp[-\theta_3])^k\right] ESP_{N,k}(\boldsymbol{y}_N), \tag{34}$$

where $ESP_{N,k}(\boldsymbol{y}_N)$ is the number of pairs of nodes $\{i, j\}$ with $y_{ij} \sum_{h \neq i,j} y_{ih} y_{jh} = k$ edgewise shared partners, are unstable for all $\theta_2 \neq 0$ and $\theta_3 < -\log 2$.□

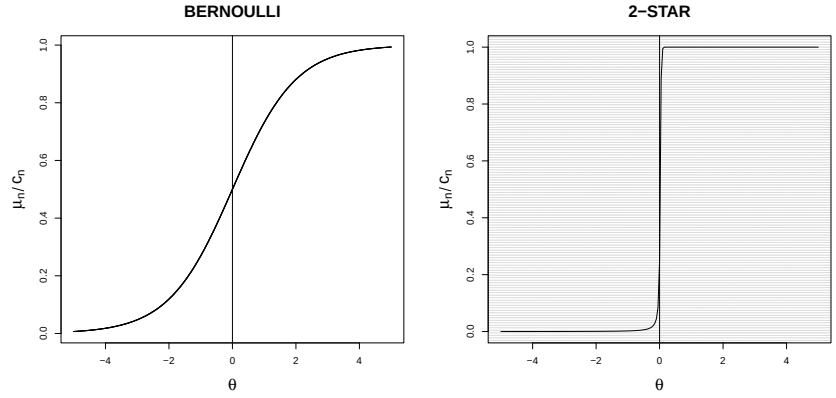
Thus, the parameter space of curved ERGMs with GWD, GWDSP, and GWESP terms contains unstable subsets. In terms of MCMC, in unstable subsets of the parameter space the curved ERGMs tend to be worse than the ERGMs with Markov dependence: if $\theta_2 \neq 0$ and $\theta_3 < -\log 2$, the nearest neighborhood log odds $|\Lambda_N(\boldsymbol{x}_N, \boldsymbol{y}_N; \boldsymbol{\theta})|$ are $O(\exp[n])$. On the other hand, the curved ERGMs with GWD, GWDSP, and GWESP terms are stable provided $\theta_2 \neq 0$ and $\theta_3 \geq -\log 2$, which is encouraging and indicates that the effective parameter space is non-negligible, in contrast to ERGMs with Markov dependence. The unstable subsets of the parameter space of curved ERGMs should be penalized by specifying suitable penalties in a maximum likelihood framework and suitable priors in a Bayesian framework.

6 Simulation results

To demonstrate that unstable discrete exponential family distributions are characterized by excessive sensitivity and near-degeneracy (cf. Section 2) and tend to obstruct MCMC simulation (cf. Section 3) and statistical inference (cf. Section 4), we resort to MCMC simulation of undirected graphs with $n = 32$ nodes and $N = 496$ degrees of freedom from the ERGMs of Results 1–6 (cf. Section 5). Since the computational cost of MCMC simulation is prohibitive, we exploit the fact that Results 1–6 hold regardless of the value of θ_1, the natural parameter corresponding to the sufficient statistic $\sum_{i<j} y_{ij}$, and fix the value of θ_1 at -1 and the value of θ_2 of the ERGMs of Results 3–6 at 1. For every ERGM and every non-fixed parameter, we consider 200 values in the interval $[-5, 5]$. At every such value, we generate an MCMC sample of size 2,000,000, discarding 1,000,000 draws as burn-in and recording every 1,000th post-burn-in draw. The MCMC samples were generated by a Metropolis-Hastings algorithm of the form (19) (Hunter et al. 2008).

We start with two classic examples: the Bernoulli model with stable sufficient statistic $g_N(\boldsymbol{y}_N) = \sum_{i<j} y_{ij}$ and the 2-star model with unstable sufficient statistic

Figure 1: MCMC sample estimate of mean-value parameter $\mu_N(\theta)$ plotted against natural parameter θ of Bernoulli model and 2-star model, where C_N ensures that the range of $\mu_N(\theta)/C_N$ is $(0, 1)$; shaded regions indicate unstable regions

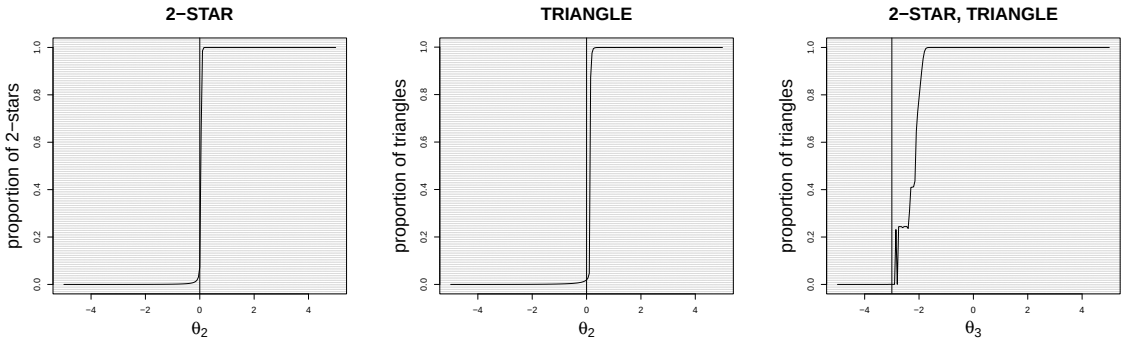


$g_N(\boldsymbol{y}_N) = \sum_{i,\, j<k} y_{ij} y_{ik}$ (cf. Section 2.1). Figure 1 plots the MCMC sample estimates of the mean-value parameters $\mu_N(\theta) = E_\theta[g_N(\boldsymbol{Y}_N)]$ of these models against the corresponding natural parameters θ. The MCMC sample estimate of the mean-value parameter $\mu_N(\theta)$ of the Bernoulli model is close to the exact value $\mu_N(\theta) = N/(1+\exp[-\theta])$ (within two standard deviations of the sample average based on random samples of size 1,000), demonstrating that MCMC simulation from the Bernoulli model is hardly problematic. The MCMC sample estimate of the mean-value parameter $\mu_N(\theta)$ of the 2-star model is, in line with Corollary 1, close to its infinum for all $\theta < 0$ and close to its supremum for all $\theta > 0$, and extremely sensitive to small changes of θ around 0.

The ERGMs with Markov dependence (Results 1–3) are expected to be degenerate with respect to the unstable sufficient statistics, the number of 2-stars (Result 1), the number of triangles (Result 2), and the number of triangles (Result 3 with 2-star parameter equal to 1), and the corresponding mean-value parameters are expected to be close to the boundary of the mean-value parameter space. Figure 2 plots the proportion of 2-stars (Result 1) and triangles (Results 2 and 3) against the corresponding natural parameter and confirms these considerations.

Concerning the curved ERGMs with GWD, GWDSP, and GWESP terms (Results 4–6), since the number of sufficient statistics is linear in n, we focus on the sufficient statistic $\sum_{i<j} y_{ij}$, one of the most fundamental functions of undirected graphs $\boldsymbol{y}_N$. We take the coefficient of variation CV_N, defined as the standard deviation of $\sum_{i<j} y_{ij}$

Figure 2: MCMC sample proportion of 2-stars (Result 1) and triangles (Results 2 and 3) plotted against corresponding natural parameter; shaded regions indicate unstable regions

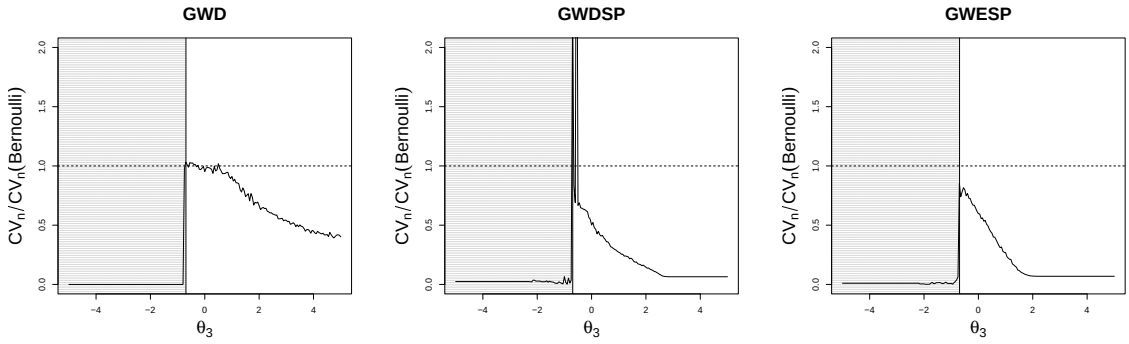


divided by the mean of $\sum_{i<j} y_{ij}$, as an indicator of mixing and near-degeneracy: low coefficients of variation indicate slow mixing and near-degeneracy. We divide the coefficients of variation CV_N by the coefficient of variation CV_N(Bernoulli) under the corresponding ERGM with $\theta_1 = -1$ and $\theta_2 = 0$, which corresponds to the Bernoulli model of Section 2.1 with $\theta = -1$. Figure 3 plots the MCMC sample coefficients of variation $\mathrm{CV}_N/\mathrm{CV}_N$(Bernoulli) against the critical parameter θ_3 of the ERGMs of Results 4–6. The simulation results indicate that in the unstable subset of the parameter space, corresponding to $\theta_3 < -\log 2$, the coefficients of variation are close to 0, as expected, and around $\theta_3 = -\log 2$, the coefficients of variation rise to a value comparable to the coefficient of variation CV_N(Bernoulli) under the corresponding Bernoulli model.

7 Discussion

Building on the work of Strauss (1986) and Handcock (2002a, 2003a,b), we have introduced the notion of instability and shown that unstable discrete exponential family distributions are characterized by excessive sensitivity and near-degeneracy. In the important special case of exponential families with unstable sufficient statistics, the subset of the natural parameter space corresponding to non-degenerate distributions and mean-value parameters far from the boundary of the mean-value parameter space turns out to be a lower-dimensional subspace of the natural parameter space. These

Figure 3: MCMC sample coefficient of variation CV_N of curved ERGM with GWD term (Result 4), GWDSP term (Result 5), and GWESP term (Result 6), re-scaled by $1/\mathrm{CV}_N(\text{Bernoulli})$; shaded regions indicate unstable regions



characteristics of instability tend to obstruct MCMC simulation and statistical inference. In applications to relational data, we find that exponential families with Markov dependence tend to be unstable and that the parameter space of some curved exponential families contains unstable subsets. We conclude that unstable subsets of the parameter space of curved exponential families should be penalized by specifying suitable penalties in a maximum likelihood framework and suitable priors in a Bayesian framework.

It is worthwhile to point out that, while instability implies undesirable behavior such as near-degeneracy, stability is not—and cannot be—an insurance against near-degeneracy. Indeed, every discrete exponential family, with or without unstable sufficient statistics, includes near-degenerate distributions provided the natural parameters are sufficiently large (cf. Barndorff-Nielsen 1978, pp. 185–186, Handcock 2002a, 2003a,b). In addition, while unstable sufficient statistics can be stabilized, there are good reasons to be sceptical of simple stabilization strategies. Consider one-parameter exponential families with natural parameter θ and unstable sufficient statistic $g_N(\boldsymbol{y}_N)$. The unstable sufficient statistic $g_N(\boldsymbol{y}_N)$ can be transformed into the stable sufficient statistic $g_N(\boldsymbol{y}_N)/U_N$ by dividing $g_N(\boldsymbol{y}_N)$ by its maximum U_N. Since the canonical form of exponential families is not unique (Brown 1986, pp. 7–8), mapping $g_N(\boldsymbol{y}_N)$ to $g_N(\boldsymbol{y}_N)/U_N$ is equivalent to mapping θ to θ/U_N and can therefore be regarded as a reparameterization of the exponential family with unstable sufficient statistic $g_N(\boldsymbol{y}_N)$. Let $\eta_N(\theta) = \theta/U_N$. By the parameterization invariance of maximum likelihood estimators, the maximum likelihood estimators $\hat{\theta}$ and $\hat{\eta}_N \stackrel{\text{def}}{=} \widehat{\eta_N(\theta)}$ of

θ respectively $\eta_N(\theta)$ satisfy $\hat{\theta} = \hat{\eta}_N U_N$. The probability of data under the maximum likelihood estimator is the same under both parameterizations. The simple stabilization strategy therefore fails to address the problem of lack of fit: even under the maximum likelihood estimator, the probability of data may be extremely low relative to other elements of the sample space and the fit of the model thus unacceptable (cf. Hunter et al. 2008). The argument extends to K-parameter exponential families and linear transformations of sufficient statistics (Brown 1986, pp. 7–8).

Last, while the conditions under which maximum likelihood estimators of discrete exponential families for dependent data exist and are unique are well-understood (cf. Barndorff-Nielsen 1978, p. 151, Handcock 2002a, 2003a,b, Rinaldo et al. 2009), it is an open question which conditions ensure consistency and asymptotic normality of maximum likelihood estimators (cf. Hunter and Handcock 2006, Rinaldo et al. 2009). An anonymous referee suggested semi-group structure (cf. Lauritzen 1988, pp. 140–146). Semi-group structure implies stability and holds promise.

A Appendix: proofs

Proof of Theorem 1. We prove Theorem 1 by contradiction. Given an unstable discrete exponential family distribution, suppose that there exist $C > 0$ and $N_C > 0$ such that

$$|\Lambda_N(\boldsymbol{x}_N, \boldsymbol{y}_N; \boldsymbol{\theta})| \leq C \text{ for all } \boldsymbol{x}_N \sim \boldsymbol{y}_N,\ \boldsymbol{x}_N,\ \boldsymbol{y}_N \in \mathcal{Y}_N \text{ for all } N > N_C. \tag{35}$$

Consider a given $N \geq 1$. Let $\boldsymbol{a}_N \in \{\boldsymbol{y}_N \in \mathcal{Y}_N : q_{\boldsymbol{\theta}}(\boldsymbol{y}_N) = I_N(\boldsymbol{\theta})\}$ and $\boldsymbol{b}_N \in \{\boldsymbol{y}_N \in \mathcal{Y}_N : q_{\boldsymbol{\theta}}(\boldsymbol{y}_N) = S_N(\boldsymbol{\theta})\}$, and let $K_N \leq N$ be the number of non-matching elements of $\boldsymbol{a}_N$ and $\boldsymbol{b}_N$. By changing the non-matching elements of $\boldsymbol{a}_N$ and $\boldsymbol{b}_N$ one by one, it is possible to go from $\boldsymbol{a}_N$ to $\boldsymbol{b}_N$ within $K_N \leq N$ steps. Let $\boldsymbol{y}_{N,0}, \boldsymbol{y}_{N,1}, \ldots, \boldsymbol{y}_{N,K_N-1}, \boldsymbol{y}_{N,K_N}$ be a path from $\boldsymbol{a}_N$ to $\boldsymbol{b}_N$ such that $\boldsymbol{y}_{N,0} = \boldsymbol{a}_N$ and $\boldsymbol{y}_{N,K_N} = \boldsymbol{b}_N$ and $\boldsymbol{y}_{N,k-1} \sim \boldsymbol{y}_{N,k}$ $(k = 1, \ldots, K_N)$. By Jensen's inequality and (35), there exist $C > 0$ and $N_C > 0$ such that, for any $N > N_C$,

$$\left|\sum_{k=1}^{K_N} \Lambda_N(\boldsymbol{y}_{N,k-1}, \boldsymbol{y}_{N,k}; \boldsymbol{\theta})\right| \ \leq\ \sum_{k=1}^{K_N} |\Lambda_N(\boldsymbol{y}_{N,k-1}, \boldsymbol{y}_{N,k}; \boldsymbol{\theta})| \ \leq\ C\,N. \tag{36}$$

The left-hand side of (36) is, by definition of $\boldsymbol{a}_N$ and $\boldsymbol{b}_N$, given by

$$\left|\sum_{k=1}^{K_N} \Lambda_N(\boldsymbol{y}_{N,k-1}, \boldsymbol{y}_{N,k}; \boldsymbol{\theta})\right| \ =\ |q_{\boldsymbol{\theta}}(\boldsymbol{b}_N) - q_{\boldsymbol{\theta}}(\boldsymbol{a}_N)| \ =\ S_N(\boldsymbol{\theta}). \tag{37}$$

Thus, (35) implies that there exist $C > 0$ and $N_C > 0$ such that

$$S_N(\boldsymbol{\theta}) \;\leq\; C\,N \text{ for all } N > N_C, \tag{38}$$

which contradicts the assumption of instability.□

Proof of Theorem 2. For any $0 < \delta < \epsilon < 1$, however small, and any $N \geq 1$,

$$P_{\boldsymbol{\theta}}(\boldsymbol{Y}_N \in \mathcal{M}_{\epsilon,N}) \geq P_{\boldsymbol{\theta}}(\boldsymbol{Y}_N \in \mathcal{M}_{\delta,N}) \geq \exp[(1-\delta)S_N(\boldsymbol{\theta}) - \psi_N(\boldsymbol{\theta})] \tag{39}$$

using the fact that $\mathcal{M}_{\delta,N}$ contains at least one element, and

$$1 - P_{\boldsymbol{\theta}}(\boldsymbol{Y}_N \in \mathcal{M}_{\epsilon,N}) < \exp[N\log M + (1-\epsilon)S_N(\boldsymbol{\theta}) - \psi_N(\boldsymbol{\theta})] \tag{40}$$

using the fact that $\mathcal{Y}_N \setminus \mathcal{M}_{\epsilon,N}$ contains at most $\exp[N\log M] - 1 < \exp[N\log M]$ elements. Thus, the log odds of $P_{\boldsymbol{\theta}}(\boldsymbol{Y}_N \in \mathcal{M}_{\epsilon,N})$ is given by

$$\omega_{\epsilon,N} = \log \frac{P_{\boldsymbol{\theta}}(\boldsymbol{Y}_N \in \mathcal{M}_{\epsilon,N})}{1 - P_{\boldsymbol{\theta}}(\boldsymbol{Y}_N \in \mathcal{M}_{\epsilon,N})} > (\epsilon - \delta)\, S_N(\boldsymbol{\theta}) - N\log M. \tag{41}$$

By instability, for any $C > 0$, however large, there exists $N_C > 0$ such that

$$\omega_{\epsilon,N} > (\epsilon-\delta)\, S_N(\boldsymbol{\theta}) - N\log M > [(\epsilon-\delta)\, C - \log M]\, N \text{ for all } N > N_C. \tag{42}$$

Since $\epsilon - \delta > 0$ and $C > 0$ can be as large as desired, $\omega_{\epsilon,N} \longrightarrow \infty$ as $N \longrightarrow \infty$ and (9) holds.□

Proof of Theorems 3 and 4. We prove Theorem 4, since Theorem 3 can be considered to be a special case of Theorem 4.

Case 1: $\theta_K < 0$. For any $0 < \delta < \epsilon < 1$, however small, and any $N \geq 1$,

$$\begin{aligned} & P_{\boldsymbol{\theta}}(\boldsymbol{Y}_N \in \mathcal{L}_{\epsilon,N,K}) \;\geq\; P_{\boldsymbol{\theta}}(\boldsymbol{Y}_N \in \mathcal{L}_{\delta,N,K}) \\ & \geq\; \exp\left\{ \sum_{k=1}^{K-1} \min_{\boldsymbol{y}_N \in \mathcal{L}_{\delta,N,K}} [\theta_k\, g_{N,k}(\boldsymbol{y}_N)] + \theta_K\, \delta\, U_{N,K} - \psi_N(\boldsymbol{\theta}) \right\} \end{aligned} \tag{43}$$

and

$$\begin{aligned} & 1 - P_{\boldsymbol{\theta}}(\boldsymbol{Y}_N \in \mathcal{L}_{\epsilon,N,K}) \\ & <\; \exp\left\{ N\log M + \sum_{k=1}^{K-1} \max_{\boldsymbol{y}_N \in \mathcal{Y}_N \setminus \mathcal{L}_{\epsilon,N,K}} [\theta_k\, g_{N,k}(\boldsymbol{y}_N)] + \theta_K\, \epsilon\, U_{N,K} - \psi_N(\boldsymbol{\theta}) \right\}. \end{aligned} \tag{44}$$

Thus, the log odds of $P_{\boldsymbol{\theta}}(\boldsymbol{Y}_N \in \mathcal{L}_{\epsilon,N,K})$ is given by

$$\begin{aligned} \omega_{\epsilon,N,K} &= \log \frac{P_{\boldsymbol{\theta}}(\boldsymbol{Y}_N \in \mathcal{L}_{\epsilon,N,K})}{1 - P_{\boldsymbol{\theta}}(\boldsymbol{Y}_N \in \mathcal{L}_{\epsilon,N,K})} \\ &> -\theta_K \, (\epsilon - \delta) \, U_{N,K} - N \log M - \sum_{k=1}^{K-1} |\theta_k| \, U_{N,k}. \end{aligned} \tag{45}$$

Since $-\theta_K > 0$, $\epsilon - \delta > 0$, and the sufficient statistic $g_{N,K}(\boldsymbol{y}_N)$ is unstable, the term $-\theta_K \, (\epsilon - \delta) \, U_{N,K}$ on the right-hand side of (45) is positive and not bounded by N, while the stability of the sufficient statistics $g_{N,1}(\boldsymbol{y}_N)$, $\ldots$, $g_{N,K-1}(\boldsymbol{y}_N)$ implies that the other terms on the right-hand side of (45) are bounded by N. Thus, for any $\theta_K < 0$, $\omega_{\epsilon,N,K} \longrightarrow \infty$ as $N \longrightarrow \infty$ and (12) holds.□

Case 2: $\theta_K > 0$. The case $\theta_K > 0$ proceeds along the same lines as the case $\theta_K < 0$, mutatis mutandis, to show that (13) holds.□

Proof of Theorem 5. A proof of Theorem 5 proceeds along the same lines as the proof of Theorem 4, with the exception that the sufficient statistics $g_{N,1}(\boldsymbol{y}_N)$, $\ldots$, $g_{N,K-1}(\boldsymbol{y}_N)$ may be unstable but are dominated by the unstable sufficient statistic $g_{N,K}(\boldsymbol{y}_N)$.□

Proof of Corollaries 1 and 2. We prove Corollary 2, since Corollary 1 can be considered to be a special case of Corollary 1.

Case 1: $\theta_K < 0$. For any $0 < \gamma < 1$, however small, and any $N \geq 1$, one can partition the sample space $\mathcal{Y}_N$ into the subsets $\mathcal{L}_{\gamma,N,K}$ and $\mathcal{Y}_N \setminus \mathcal{L}_{\gamma,N,K}$. Therefore,

$$\frac{\mu_{N,K}(\boldsymbol{\theta})}{U_{N,K}} \leq \gamma \, P_{\boldsymbol{\theta}}(\boldsymbol{Y} \in \mathcal{L}_{\gamma,N,K}) + P_{\boldsymbol{\theta}}(\boldsymbol{Y} \in \mathcal{Y}_N \setminus \mathcal{L}_{\gamma,N,K}). \tag{46}$$

By Theorem 4, for any $0 < \delta < 1$, however small, and any $\theta_K < 0$, there exists $N_\delta > 0$ such that

$$\gamma \, P_{\boldsymbol{\theta}}(\boldsymbol{Y} \in \mathcal{L}_{\gamma,N,K}) + P_{\boldsymbol{\theta}}(\boldsymbol{Y} \in \mathcal{Y}_N \setminus \mathcal{L}_{\gamma,N,K}) < \gamma + \delta = \epsilon \text{ for all } N > N_\delta. \tag{47}$$

Since γ and δ can be as small as desired, (26) holds.

Case 2: $\theta_K > 0$. The case $\theta_K > 0$ proceeds along the same lines as the case $\theta_K < 0$, mutatis mutandi, to show that (27) holds.□

Proof of Results 1–6. Let $\mathbf{0}_N$ be the empty graph ($0_{ij} = 0$, all $i < j$) and $\mathbf{1}_N$ be the complete graph ($1_{ij} = 1$, all $i < j$) given n nodes and $N = n(n-1)/2$ degrees of freedom. For every ERGM of Results 1–6, every $\boldsymbol{\theta} \in \Theta$, and every $n > 1$, $q_{\boldsymbol{\theta}}(\mathbf{0}_N) = 0$ and

$$S_N(\boldsymbol{\theta}) - I_N(\boldsymbol{\theta}) \geq |q_{\boldsymbol{\theta}}(\mathbf{1}_N) - q_{\boldsymbol{\theta}}(\mathbf{0}_N)| = |q_{\boldsymbol{\theta}}(\mathbf{1}_N)|. \tag{48}$$

Therefore, all $\boldsymbol{\theta} \in \Theta$ such that $|q_{\boldsymbol{\theta}}(\mathbf{1}_N)|$ is not bounded by N give rise to unstable distributions $P_{\boldsymbol{\theta}}$, proving Results 1–6.□

References

Barndorff-Nielsen, O. E. (1978), *Information and Exponential Families in Statistical Theory*, New York: Wiley.

Besag, J. (1974), "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society, Series B*, 36, 192–225.

Brown, L. (1986), *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*, Hayworth, CA, USA: Institute of Mathematical Statistics.

Efron, B. (1978), "The geometry of exponential families," *Annals of Statistics*, 6, 362–376.

Frank, O., and Strauss, D. (1986), "Markov graphs," *Journal of the American Statistical Association*, 81, 832–842.

Geyer, C. J. (2009), "Likelihood inference in exponential families and directions of recession," *Electronic Journal of Statistics*, 3, 259–289.

Geyer, C. J., and Thompson, E. A. (1992), "Constrained Monte Carlo maximum likelihood for dependent data," *Journal of the Royal Statistical Society, Series B*, 54, 657–699.

Handcock, M. (2002a), "Degeneracy and inference for social network models," Paper presented at the Sunbelt XXII International Social Network Conference in New Orleans, LA.

— (2002b), "Degeneracy and inference for social network models," Paper presented at the Joint Statistical Meetings in New York, NY.

Handcock, M. (2003a), "Assessing degeneracy in statistical models of social networks," Tech. rep., Center for Statistics and the Social Sciences, University of Washington, http://www.csss.washington.edu/Papers.

— (2003b), "Statistical Models for Social Networks: Inference and Degeneracy," in *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, eds. Breiger, R., Carley, K., and Pattison, P., Washington, D.C.: National Academies Press.

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

Hunter, D. R., and Handcock, M. S. (2006), "Inference in curved exponential family models for networks," *Journal of Computational and Graphical Statistics*, 15, 565–583.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., and Morris, M. (2008), "ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks," *Journal of Statistical Software*, 24, 1–29.

Jonasson, J. (1999), "The random triangle model," *Journal of Applied Probability*, 36, 852–876.

Koskinen, J. H., Robins, G. L., and Pattison, P. E. (2010), "Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation," *Statistical Methodology*, 7, 366–384.

Lauritzen, S. L. (1988), *Extremal Families and Systems of Sufficient Statistics*, Heidelberg: Springer.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006), "An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants," *Biometrika*, 93, 451–458.

Rinaldo, A., Fienberg, S. E., and Zhou, Y. (2009), "On the geometry of discrete exponential families with application to exponential random graph models," *Electronic Journal of Statistics*, 3, 446–484.

Ruelle, D. (1969), *Statistical mechanics. Rigorous results*, London and Singapore: Imperial College Press and World Scientific.

Snijders, T. A. B. (2002), "Markov chain Monte Carlo Estimation of exponential random graph models," *Journal of Social Structure*, 3, 1–40.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006), "New specifications for exponential random graph models," *Sociological Methodology*, 36, 99–153.

Strauss, D. (1986), "On a general class of models for interaction," *SIAM Review*, 28, 513–527.

Wasserman, S., and Faust, K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge: Cambridge University Press.

Wasserman, S., and Pattison, P. (1996), "Logit Models and Logistic Regression for Social Networks: I. An Introduction to Markov Graphs and p^*," *Psychometrika*, 61, 401–425.