2017

Using contrastive divergence to seed Monte Carlo MLE for exponential-family random graph models

Pavel N. Krivitsky
University of Wollongong, pavel@uow.edu.au

Publication Details

Krivitsky, P. N. (2017). Using contrastive divergence to seed Monte Carlo MLE for exponential-family random graph models. Computational Statistics and Data Analysis, 107 149-161.

Using contrastive divergence to seed Monte Carlo MLE for exponential-family random graph models

Abstract

Exponential-family models for dependent data have applications in a wide variety of areas, but the dependence often results in an intractable likelihood, requiring either analytic approximation or MCMC-based techniques to fit, the latter requiring an initial parameter configuration to seed their simulations. A poor initial configuration can lead to slow convergence or outright failure. The approximate techniques that could be used to find them tend not to be as general as the simulation-based and require implementation separate from that of the MLE-finding algorithm. Contrastive divergence is a more recent simulation-based approximation technique that uses a series of abridged MCMC runs instead of running them to stationarity. Combining it with the importance sampling Monte Carlo MLE yields a method for obtaining adequate initial values that is applicable to a wide variety of modeling scenarios. Practical issues such as stopping criteria and selection of tuning parameters are also addressed. A simple generalization of the Monte Carlo MLE partial stepping algorithm to curved exponential families (applicable to MLE-finding as well) is also proposed. The proposed approach reuses the aspects of an MLE implementation that are model-specific, so little to no additional implementer effort is required to obtain adequate initial parameters. This is demonstrated on a series of network datasets and models drawn from exponential-family random graph model computation literature, also exploring the limitations of the techniques considered.

Keywords

graph, random, exponential-family, mle, carlo, monte, seed, models, divergence, contrastive

Disciplines

Engineering | Science and Technology Studies

Publication Details

Krivitsky, P. N. (2017). Using contrastive divergence to seed Monte Carlo MLE for exponential-family random graph models. Computational Statistics and Data Analysis, 107 149-161.

Using Contrastive Divergence to Seed Monte Carlo MLE for Exponential-Family Random Graph Models

Pavel N. Krivitsky[a]

[a] *School of Mathematics and Applied Statistics and National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong, New South Wales, Australia*

Abstract

Exponential-family models for dependent data have applications in a wide variety of areas, but the dependence often results in an intractable likelihood, requiring either analytic approximation or MCMC-based techniques to fit, the latter requiring an initial parameter configuration to seed their simulations. A poor initial configuration can lead to slow convergence or outright failure. The approximate techniques that could be used to find them tend not to be as general as the simulation-based and require implementation separate from that of the MLE-finding algorithm.

Contrastive divergence is a more recent simulation-based approximation technique that uses a series of abridged MCMC runs instead of running them to stationarity. Combining it with the importance sampling Monte Carlo MLE yields a method for obtaining adequate initial values that is applicable to a wide variety of modeling scenarios. Practical issues such as stopping criteria and selection of tuning parameters are also addressed. A simple generalization of the Monte Carlo MLE partial stepping algorithm to curved exponential families (applicable to MLE-finding as well) is also proposed.

The proposed approach reuses the aspects of an MLE implementation that are model-specific, so little to no additional implementer effort is required to obtain adequate initial parameters. This is demonstrated on a series of network datasets and models drawn from exponential-family ran-

Email address: `pavel@uow.edu.au` (Pavel N. Krivitsky)

[1] Datasets, R packages, and scripts to reproduce the simulation studies reported are included in a supplementary file. However, the R packages in particular are under continuing development, so a more recent version published to CRAN may perform better, even if it does not reproduce the simulation exactly.

dom graph model computation literature, also exploring the limitations of the techniques considered.

Keywords: curved exponential family, ERGM, network data, partial stepping

1. Introduction

Exponential family models for dependent data have found applications in point processes, social networks, statistical physics, and image analysis alike, but this dependence often produces likelihoods with intractable normalizing constants. A variety of techniques—frequentist and Bayesian—have been proposed for their estimation. Although some approximations are available, the exact techniques invariably require a starting parameter configuration $\boldsymbol{\theta}^0$, their performance and even feasibility depending on this value.

In this work, we focus on the problem of a general way of obtaining a good $\boldsymbol{\theta}^0$ with minimal additional implementer effort, particularly for the application of these models to modeling of social networks—the exponential-family random graph models (ERGMs) (Wasserman and Pattison, 1996), as extended to curved families by Snijders et al. (2006) and Hunter and Handcock (2006) and to networks with valued ties by Robins et al. (1999) and Krivitsky (2012). We consider the broad class of models defined as follows. Given a set $N = \{1, 2, \ldots, n\}$ of actors of interest, let $\mathbb{Y} \subseteq N \times N$ be the set of potential relationships among them (usually a proper subset, excluding self-loops or if only ties among specific subsets of actors are of interest). Then, with $\mathbb{S}$ being the set of possible relationship values (which could be simply $\{0, 1\}$ for binary networks), we define the sample space of mappings $\mathcal{Y} \subseteq \mathbb{S}^{\mathbb{Y}}$ (again, sometimes a proper subset if, say, we wish to constrain the network to have a specific number of ties or a specific degree distribution). In the interests of accessibility, we will focus on finite or countable $\mathbb{S}$ and $\mathcal{Y}$, using notation of probabilities and summations, rather than the more general case with Radon–Nikodym derivatives and Lebesgue integrals; but the ERGM formulation for uncountable $\mathbb{S}$ is analogous (Krivitsky, 2012, p. 1121), and given the mechanics of exponential families—centered on sufficient statistics—all of the review and developments in this manuscript should be applicable to that case as well.

We write $\boldsymbol{Y} \sim \mathrm{ERGM}_{\mathcal{Y},h,\boldsymbol{\eta},\boldsymbol{g}}(\boldsymbol{\theta})$ if

$$\mathrm{Pr}_{\mathcal{Y},h,\boldsymbol{\eta},\boldsymbol{g}}(\boldsymbol{Y}=\boldsymbol{y};\boldsymbol{\theta}) \overset{\text{def}}{=} \frac{h(\boldsymbol{y})\exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top \boldsymbol{g}(\boldsymbol{y})\}}{\kappa_{\mathcal{Y},h,\boldsymbol{\eta},\boldsymbol{g}}(\boldsymbol{\theta})},\ \boldsymbol{y}\in\mathcal{Y}:$$

an exponential family over a sample space $\mathcal{Y}$ of networks (potentially with valued ties), parametrized by a q-vector $\boldsymbol{\theta}$, and specified by a reference measure $h(\boldsymbol{y})$ (with $h(\boldsymbol{y}) \propto 1$ being typical for binary ERGMs), a mapping $\boldsymbol{\eta}$ from $\boldsymbol{\theta}$ to the p-vector of canonical parameters (and in non-curved ERGMs, $\boldsymbol{\eta}(\boldsymbol{\theta}) \equiv \boldsymbol{\theta}$ with $p \equiv q$), and a sufficient statistic p-vector $\boldsymbol{g}(\cdot)$. The normalizing constant $\kappa_{\mathcal{Y},h,\boldsymbol{\eta},\boldsymbol{g}}(\boldsymbol{\theta}) \overset{\text{def}}{=} \sum_{\boldsymbol{y}'\in\mathcal{Y}} h(\boldsymbol{y}')\exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top \boldsymbol{g}(\boldsymbol{y}')\}$, is often intractable for models that seek to reproduce more complex social effects, such as triadic closure. It also identifies the *natural parameter space* of the model, $\boldsymbol{\Theta}_{\mathrm{N}} \overset{\text{def}}{=} \{\boldsymbol{\theta} : \kappa_{\mathcal{Y},h,\boldsymbol{\eta},\boldsymbol{g}}(\boldsymbol{\theta}) < \infty\}$, which equals $\mathbb{R}^q$ for binary ERGMs, but which may be far more complex for valued ERGMs, such as if geometric or Conway–Maxwell–Poisson (CMP) distribution (Shmueli et al., 2005) is used for social interaction counts (Krivitsky, 2012). Unless it is relevant to the discussion, we will, generally, omit "$\mathcal{Y}, h, \boldsymbol{\eta}, \boldsymbol{g}$" from the subscript.

Given an observed network, $\boldsymbol{y}^{\text{obs}}$, it is desired to find the MLE, $\hat{\boldsymbol{\theta}} \overset{\text{def}}{=} \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \log \mathrm{Pr}(\boldsymbol{Y} = \boldsymbol{y}^{\text{obs}};\boldsymbol{\theta})$, or, equivalently (assuming a unique maximum, which holds for non-curved families but is not guaranteed for curved), to solve the score estimating equation,

$$\boldsymbol{U}(\hat{\boldsymbol{\theta}}) \overset{\text{def}}{=} \boldsymbol{\nabla}_{\boldsymbol{\theta}}\ell(\hat{\boldsymbol{\theta}}) = \boldsymbol{\eta}'(\hat{\boldsymbol{\theta}})^\top[\boldsymbol{g}(\boldsymbol{y}^{\text{obs}}) - \mathrm{E}\{\boldsymbol{g}(\boldsymbol{Y});\hat{\boldsymbol{\theta}}\}] = -\boldsymbol{\eta}'(\hat{\boldsymbol{\theta}})^\top\,\mathrm{E}\{\boldsymbol{z}(\boldsymbol{Y});\hat{\boldsymbol{\theta}}\} = \boldsymbol{0}, \tag{1}$$

(Hunter and Handcock, 2006, eq. 3.1), where $\boldsymbol{\eta}'(\cdot) \overset{\text{def}}{=} \boldsymbol{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\eta}(\cdot)$, $\mathrm{E}(\cdot;\cdot)$ denotes the expectation under the model and parameter configuration in question, and $\boldsymbol{z}(\boldsymbol{y}) \overset{\text{def}}{=} \boldsymbol{g}(\boldsymbol{y}) - \boldsymbol{g}(\boldsymbol{y}^{\text{obs}})$.

We use $\vec{\boldsymbol{y}}$ as shorthand for a sample or series of networks $\boldsymbol{y}^1,\ldots,\boldsymbol{y}^S$, with $\vec{\boldsymbol{y}}^{\boldsymbol{\theta}}$ in particular being a sample from $\mathrm{ERGM}(\boldsymbol{\theta})$, and we use $\boldsymbol{g}(\vec{\boldsymbol{y}})$ for a $p \times S$ matrix with sth column containing $\boldsymbol{g}(\boldsymbol{y}^s)$, with $\bar{\boldsymbol{g}}(\vec{\boldsymbol{y}}) \overset{\text{def}}{=} \boldsymbol{g}(\vec{\boldsymbol{y}})\,\boldsymbol{1}_S\,/S$, and, analogously $\boldsymbol{z}(\vec{\boldsymbol{y}})$ and $\bar{\boldsymbol{z}}(\vec{\boldsymbol{y}})$; and we define $\boldsymbol{U}_{\vec{\boldsymbol{y}}^{\boldsymbol{\theta}}}(\boldsymbol{\theta}) \overset{\text{def}}{=} -\boldsymbol{\eta}'(\boldsymbol{\theta})^\top \boldsymbol{z}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}})$, a $q \times S$ matrix whose sth column is the contribution to (1) from $\boldsymbol{y}^s$, so that $\bar{\boldsymbol{U}}_{\vec{\boldsymbol{y}}^{\boldsymbol{\theta}}}(\boldsymbol{\theta})$ is the sample estimate of $\boldsymbol{\nabla}_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})$. We also use the sample variance of a statistic $\boldsymbol{t}(\vec{\boldsymbol{y}})$,

$$\widetilde{\mathrm{Var}}\{\boldsymbol{t}(\vec{\boldsymbol{y}})\} \overset{\text{def}}{=} \frac{1}{S-1}\sum_{s=1}^{S}\{\boldsymbol{t}(\boldsymbol{y}^s) - \bar{\boldsymbol{t}}(\vec{\boldsymbol{y}})\}\{\boldsymbol{t}(\boldsymbol{y}^s) - \bar{\boldsymbol{t}}(\vec{\boldsymbol{y}})\}^\top.$$

A body of literature exists on computational methods for finding $\hat{\boldsymbol{\theta}}$ given a starting configuration $\boldsymbol{\theta}^0$; and on approximate techniques suitable for finding such a configuration.

1.1. Techniques for finding the MLE

The currently popular MLE techniques can be broadly classified into two categories: stochastic approximation (SA) and Monte Carlo Maximum Likelihood Estimation (MCMLE). We review them in turn.

1.1.1. Stochastic Approximation Methods

Stochastic approximation methods represented the first attempts to find the actual MLE for ERGMs, starting with Snijders (2002) application of Robbins and Monro (1951) and similar algorithms, and, later, refinements such as those of Okabayashi and Geyer (2012). Given a guess $\boldsymbol{\theta}^t$, these techniques simulate a sample $\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t} = (\boldsymbol{y}^{\boldsymbol{\theta}^t,1}, \ldots, \boldsymbol{y}^{\boldsymbol{\theta}^t,S})$ from ERGM$(\boldsymbol{\theta}^t)$ and update the guess to

$$\boldsymbol{\theta}^{t+1} \stackrel{\text{def}}{=} \boldsymbol{\theta}^t - \boldsymbol{\alpha}_t \bar{\boldsymbol{U}}_{\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t}}(\boldsymbol{\theta}^t),$$

for $\boldsymbol{\alpha}_t$ a scalar or a $q \times q$ matrix that is decreasing in t. (The gradient methods cited are all specified for non-curved ERGMs, but this is a direct extension.) Robbins–Monro implementation as used by Snijders (2002) and the `PNet` software suite for ERGM inference (Wang et al., 2014) uses a scalar multiple of the inverse of the diagonal of $\widetilde{\text{Var}}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t})$ in particular.

Methods of this type require an initial guess, $\boldsymbol{\theta}^0$. In the context of network models in particular, a poor initial guess may induce a near-degenerate distribution concentrated on the edge of the convex hull of the set of attainable statistics $\text{Conv}(\{\boldsymbol{g}(\boldsymbol{y}') : \boldsymbol{y}' \in \mathcal{Y}\})$ (often an empty network or a complete graph). (Rinaldo et al., 2009; Hunter et al., 2012, and others) While $\boldsymbol{U}(\boldsymbol{\theta}^0)$ itself may not be on the edge of this convex hull, its sample value $\boldsymbol{U}_{\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^0}}(\boldsymbol{\theta}^0)$ could very well be, leaving the gradient-based methods without an unambiguous direction of ascent. And, if $\boldsymbol{\Theta}_{\text{N}} \neq \mathbb{R}^q$, MCMC sampling for $\boldsymbol{\theta}^0 \notin \boldsymbol{\Theta}_{\text{N}}$ will diverge in the first place, and locating a $\boldsymbol{\theta}^0 \in \boldsymbol{\Theta}_{\text{N}}$ may itself be a challenge. (Krivitsky, 2012)

Choice of $\boldsymbol{\theta}^0$ can affect estimation in other ways as well: while one can represent a network $\boldsymbol{y}$ as an $n \times n$ matrix of relationship values, most large networks studied tend to be sparse, and sparse matrix representations are used in implementations. Then, storing and processing a network with more ties is more costly in both memory and time, and if a poor choice of $\boldsymbol{\theta}^0$

induces very dense networks, computation can be slowed down severely or fail.

SA algorithms also tend to be relatively computationally inefficient: every new guess $\boldsymbol{\theta}^t$ requires a burn-in period and a sample to estimate $\boldsymbol{U}(\boldsymbol{\theta}^t)$, and optimal length of each step is unknown, so relatively many such steps are typically required.

1.1.2. Monte Carlo Maximum Likelihood Estimation

Introduced by Geyer and Thompson (1992), and applied to curved ERGMs by Hunter and Handcock (2006), MCMLE draws on importance sampling integration, observing that

$$\begin{aligned}\frac{\kappa(\boldsymbol{\theta}')}{\kappa(\boldsymbol{\theta})} &= \sum_{\boldsymbol{y}\in\mathcal{Y}} \exp[\{\boldsymbol{\eta}(\boldsymbol{\theta}') - \boldsymbol{\eta}(\boldsymbol{\theta})\}^\top \boldsymbol{g}(\boldsymbol{y})] \frac{h(\boldsymbol{y})\exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top \boldsymbol{g}(\boldsymbol{y})\}}{\kappa(\boldsymbol{\theta})} \\ &= \mathrm{E}\left(\exp[\{\boldsymbol{\eta}(\boldsymbol{\theta}') - \boldsymbol{\eta}(\boldsymbol{\theta})\}^\top \boldsymbol{g}(\boldsymbol{Y})]; \boldsymbol{\theta}\right),\end{aligned}$$

and proposes to estimate this expectation for values of $\boldsymbol{\theta}'$ near $\boldsymbol{\theta}$ based on a sample from the model with configuration $\boldsymbol{\theta}$: given a sample $\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t}$ from ERGM($\boldsymbol{\theta}^t$), update the guess

$$\begin{aligned}\boldsymbol{\theta}^{t+1} &= \arg\max_{\boldsymbol{\theta}'} \left(\{\boldsymbol{\eta}(\boldsymbol{\theta}') - \boldsymbol{\eta}(\boldsymbol{\theta}^t)\}^\top \boldsymbol{g}(\boldsymbol{y}^{\mathrm{obs}}) - \log \frac{1}{S} \sum_{s=1}^{S} \exp[\{\boldsymbol{\eta}(\boldsymbol{\theta}') - \boldsymbol{\eta}(\boldsymbol{\theta}^t)\}^\top \boldsymbol{g}(\boldsymbol{y}^{\boldsymbol{\theta}^t,s})]\right) \\ &= \arg\max_{\boldsymbol{\theta}'} \log \frac{1}{S} \sum_{s=1}^{S} \exp[-\{\boldsymbol{\eta}(\boldsymbol{\theta}') - \boldsymbol{\eta}(\boldsymbol{\theta}^t)\}^\top \boldsymbol{z}(\boldsymbol{y}^{\boldsymbol{\theta}^t,s})]. \qquad (2)\end{aligned}$$

This is equivalent to solving (again, assuming a unique maximum)

$$\hat{\boldsymbol{U}}_{\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t}}(\boldsymbol{\theta}^{t+1}) \overset{\text{def}}{=} -\boldsymbol{\eta}'(\boldsymbol{\theta}^{t+1})^\top \hat{\mathrm{E}}_{\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t}}\{\boldsymbol{z}(\boldsymbol{Y}); \boldsymbol{\theta}^{t+1}\} = \boldsymbol{0},$$

the MCMLE approximation of the score equation, where, for a statistic $\boldsymbol{t}(\cdot)$,

$$\hat{\mathrm{E}}_{\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t}}\{\boldsymbol{t}(\boldsymbol{Y}); \boldsymbol{\theta}'\} \overset{\text{def}}{=} \frac{\sum_{s=1}^{S} \exp[\{\boldsymbol{\eta}(\boldsymbol{\theta}') - \boldsymbol{\eta}(\boldsymbol{\theta}^t)\}^\top \boldsymbol{g}(\boldsymbol{y}^{\boldsymbol{\theta}^t,s})] \boldsymbol{t}(\boldsymbol{y}^{\boldsymbol{\theta}^t,s})}{\sum_{s=1}^{S} \exp[\{\boldsymbol{\eta}(\boldsymbol{\theta}') - \boldsymbol{\eta}(\boldsymbol{\theta}^t)\}^\top \boldsymbol{g}(\boldsymbol{y}^{\boldsymbol{\theta}^t,s})]},$$

the importance sampling approximation of $\mathrm{E}\{\boldsymbol{t}(\boldsymbol{Y}); \boldsymbol{\theta}'\}$.

The MCMLE approach has the benefit of making very efficient use of the simulated sample, compared to the SA methods (Geyer and Thompson, 1992, Sec. 1.3): it uses the entire distribution of $\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t}$, rather than just its first two moments, incorporates nonlinear effects of $\boldsymbol{\theta}$ on $\mathrm{E}\{\boldsymbol{g}(\boldsymbol{Y});\boldsymbol{\theta}\}$ in determining the next guess, and automatically determines the optimal (or close) step length, requiring much fewer sampling runs before convergence.

This efficiency comes at a cost: MCMLE is highly sensitive to a poor initial guess $\boldsymbol{\theta}^0$. Whereas SA methods only fail if the sample lies entirely on the edge of the convex hull (or $\boldsymbol{\theta}^0 \notin \Theta_{\mathrm{N}}$), MCMLE for non-curved ERGMs will also fail whenever the convex hull of the simulated statistics, $\mathrm{Conv}\{\boldsymbol{g}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^0})\}$, does not contain $\boldsymbol{g}(\boldsymbol{y}^{\mathrm{obs}})$ (or, equivalently, $\mathbf{0} \notin \mathrm{Conv}\{\boldsymbol{U}_{\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^0}}(\boldsymbol{\theta}^0)\}$). Then, $\boldsymbol{\theta}^{t+1}$ does not exist. (Hummel et al., 2012, p. 926)

Hummel et al. (2012) proposed two major modifications to the MCMLE algorithm that ameliorate this. The first is the lognormal approximation: if $\boldsymbol{g}(\vec{\boldsymbol{Y}}^{\boldsymbol{\theta}^t})$ is approximately normal, $\exp[\{\boldsymbol{\eta}(\boldsymbol{\theta}') - \boldsymbol{\eta}(\boldsymbol{\theta}^t)\}^\top \boldsymbol{g}(\vec{\boldsymbol{Y}}^{\boldsymbol{\theta}^t})]$ is lognormal, and its expectation gives an approximation

$$\ell(\boldsymbol{\theta}') - \ell(\boldsymbol{\theta}^t) \approx \{\boldsymbol{\eta}(\boldsymbol{\theta}') - \boldsymbol{\eta}(\boldsymbol{\theta}^t)\}^\top \{-\bar{\boldsymbol{z}}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t})\} - \\ \{\boldsymbol{\eta}(\boldsymbol{\theta}') - \boldsymbol{\eta}(\boldsymbol{\theta}^t)\}^\top \widetilde{\mathrm{Var}}\{\boldsymbol{z}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t})\}\{\boldsymbol{\eta}(\boldsymbol{\theta}') - \boldsymbol{\eta}(\boldsymbol{\theta})\}/2, \quad (3)$$

whose maximizer in $\boldsymbol{\theta}'$ depends only on the first two moments of $\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t}$ and has a closed form for non-curved ERGMs—the version derived by Hummel et al. (2012)—extending directly to curved models, though the maximizer no longer has a closed form (as implemented in the R (R Core Team, 2015) package `ergm` (Hunter et al., 2008; Handcock et al., 2015)).

The second is the Partial Stepping technique, where a step length $0 < \gamma \leq 1$ is selected, and $\boldsymbol{g}(\boldsymbol{y}^{\mathrm{obs}})$ is replaced with $\gamma \boldsymbol{g}(\boldsymbol{y}^{\mathrm{obs}}) + (1-\gamma)\bar{\boldsymbol{g}}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t})$ in the calculation of $\hat{\boldsymbol{U}}_{\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t}}(\cdot)$. In other words, the vector of observed statistics is shifted towards the centroid of the simulated statistics, reducing the length of the step while preserving its general direction. Hummel et al. choose γ adaptively, selecting a safety margin (1.05) and finding the highest $\gamma \leq 1$ such that

$$1.05\gamma \boldsymbol{g}(\boldsymbol{y}^{\mathrm{obs}}) + (1 - 1.05\gamma)\bar{\boldsymbol{g}}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t}) \in \mathrm{Conv}\{\boldsymbol{g}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t})\}. \quad (4)$$

While this approach survives poor starting values (provided $\boldsymbol{\theta}^0 \in \Theta_{\mathrm{N}}$), it is not immune to them, in that a poor starting value is likely to result in a tiny γ and a very long optimization. And so, we turn to the question of obtaining good values for $\boldsymbol{\theta}^0$.

1.2. Techniques for Finding Starting Values

Although there have been some recent developments on asymptotic approximations for ERGMs (He and Zheng, 2015), they have only been derived for a very specific set of models, and may or may not generalize. The two major techniques for obtaining $\boldsymbol{\theta}^0$ are the *maximum pseudo-/composite likelihood estimation* (MPLE/MCLE) and the more recently proposed *contrastive divergence* (CD). (It is also possible to instead fit just the intercept parameter of the model and initialize the remaining elements of $\boldsymbol{\theta}$ to 0, as is done by `PNet` (Wang et al., 2014).)

1.2.1. Composite Likelihood

Before simulation-based methods were proposed, the only practical way to fit ERGMs with intractable normalizing constants was using pseudolikelihood (Besag, 1974; Strauss and Ikeda, 1990), approximating

$$\mathrm{L}(\boldsymbol{\theta}) \approx \tilde{\mathrm{L}}(\boldsymbol{\theta}) \overset{\text{def}}{=} \prod_{(i,j)\in\mathbb{Y}} \Pr(Y_{i,j} = y_{i,j}^{\text{obs}} | \boldsymbol{Y}_{\neg(i,j)} = \boldsymbol{y}_{\neg(i,j)}^{\text{obs}}; \boldsymbol{\theta}), \tag{5}$$

where $y_{i,j}$ is the indicator of the presence of a tie from actor i to actor j and $\boldsymbol{y}_{\neg(i,j)}$ is the set of all ties in $\boldsymbol{y}$ excluding (i, j). The pseudolikelihood is then maximized to produce the maximum pseudolikelihood estimator (MPLE) $\tilde{\boldsymbol{\theta}}$.

For binary ERGMs, this gives an estimating equation

$$\tilde{\boldsymbol{U}}(\boldsymbol{\theta}) \overset{\text{def}}{=} \boldsymbol{\eta}'(\tilde{\boldsymbol{\theta}})^\top \sum_{(i,j)\in\mathbb{Y}} [\boldsymbol{y}_{i,j}^{\text{obs}} - \text{logit}^{\text{-1}}\{\boldsymbol{\eta}(\tilde{\boldsymbol{\theta}})^\top \boldsymbol{\Delta}_{i,j}\boldsymbol{g}(\boldsymbol{y}^{\text{obs}})\}]\boldsymbol{\Delta}_{i,j}\boldsymbol{g}(\boldsymbol{y}^{\text{obs}}) = \boldsymbol{0},$$

a (nonlinear) logistic regression, with "covariates" $\boldsymbol{\Delta}_{i,j}\boldsymbol{g}(\boldsymbol{y}) \overset{\text{def}}{=} \boldsymbol{g}(\boldsymbol{y}\cup\{(i,j)\}) - \boldsymbol{g}(\boldsymbol{y}\backslash\{(i,j)\})$, the effect of adding the tie (i,j) to the network $\boldsymbol{y}$ on $\boldsymbol{g}(\boldsymbol{y})$, all other ties being equal. (We are not aware of any existing implementations of MPLE for curved ERGMs, however.)

MPLE can be quite different from the MLE, however, (van Duijn et al., 2009) so, with growing computing power making methods of Section 1.1 feasible, today it is mainly used to initialize them. Even in that capacity, it has practical limitations. For example, consider a network drawn from a process in which the total number of ties that can be observed is fixed at c, used in the application by Hunter and Handcock (2006). That is, $\mathcal{Y} = \{\boldsymbol{y} \in 2^{\mathbb{Y}} : |\boldsymbol{y}| = c\}$. One Metropolis–Hastings algorithm for exploring such a sample space selects one tie and one non-tie in $\boldsymbol{y}^s$ at random and proposes to toggle both of them, thus preserving the total number of ties. Using this algorithm to sample $\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t}$ for either MCMLE or SA would result in the MLE on the constrained sample space.

In contrast, MPLE, and its generalization, maximum composite likelihood estimate (MCLE) (Lindsay, 1988), would require an algorithm to enumerate, rather than explore, the set of possible pairs of toggles, and the resulting pseudolikelihood would no longer be a binary logistic regression, but rather a multinomial model. In practice, this creates an additional burden on the implementer. Other constraints—such as conditioning on the degree sequence of a graph—require as many as 4 or 6 toggles in the proposal. (Rao et al., 1996) The resulting combinatorial explosion can be addressed by sampling, but the problem of requiring a reimplementation of MPLE remains.

In valued ERGMs, $\Pr(Y_{i,j} = y_{i,j}^{\mathrm{obs}} | \boldsymbol{Y}_{\neg(i,j)} = \boldsymbol{y}_{\neg(i,j)}^{\mathrm{obs}}; \boldsymbol{\theta})$ might, itself, be intractable, such as when CMP (Shmueli et al., 2005) is used, whereas MCMC-based methods require no additional implementational or computational effort. (Krivitsky, 2012)

1.2.2. Contrastive Divergence

In a model whose log-likelihood gradient could only be obtained by an MCMC simulation, Hinton (2002) proposed not to run the MCMC simulation to convergence but rather to make a series of parallel MCMC updates, each starting at the observed data, and calculate the gradient based on that. As applied to ERGMs by Asuncion et al. (2010), given an MCMC sampling algorithm for $\mathrm{ERGM}(\boldsymbol{\theta})$, let $\mathrm{ERGM}_{\mathrm{CD}_k}(\boldsymbol{\theta})$ be the distribution of random graphs produced after k MCMC transitions starting with $\boldsymbol{y}^{\mathrm{obs}}$. Call its expectation $\mathrm{E}_{\mathrm{CD}_k}(\cdot; \boldsymbol{\theta})$. Then, a CD_k estimate $\tilde{\boldsymbol{\theta}}^k$ solves

$$\boldsymbol{U}_{\mathrm{CD}_k}(\tilde{\boldsymbol{\theta}}^k) \overset{\mathrm{def}}{=} \boldsymbol{\eta}'(\tilde{\boldsymbol{\theta}}^k)^\top [- \mathrm{E}_{\mathrm{CD}_k}\{\boldsymbol{z}(\boldsymbol{Y}); \tilde{\boldsymbol{\theta}}^k\}] = \boldsymbol{0}, \tag{6}$$

shown by Hyvrinen (2006) to be equivalent to the MPLE if only one variable (i.e. edge) is updated and the updates are full-conditional Gibbs. Asuncion et al. (2010) noted that CD_1 (the MPLE) and CD_∞ (the MLE) were endpoints of a continuum of increasingly close approximations to the latter and showed that if k variables are block-updated in each MCMC step (*blocked contrastive divergence* (BCD)), CD_1 estimate is equivalent to maximizing the composite likelihood with block size of k. Asuncion et al. then applied CD_k to a number of exponential families, including ERGMs, using SA (with $\boldsymbol{\alpha}_t$ a scalar) to find the MCLE. Carreira-Perpiñan and Hinton (2005) proposed using the CD_k estimates to seed MCMLE.

No burn-in phase is required for CD_k estimates, which means that some of the inefficiency of the SA algorithms is not as problematic, but the issues of step length remain: Asuncion et al. (2010) used very short steps, for example. Also, the sampling algorithm required is distinct from the one that one might use for MCMLE, so using BCD as initial values for MCMLE may require additional effort on the part of the implementer.

Notice, however, that CD_k sampling alleviates the sensitivity issues of MCMLE: if $\boldsymbol{\eta}(\boldsymbol{\theta}^0) = \mathbf{0}$, then for sampling $\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^0,k}$ from $\mathrm{ERGM}_{\mathrm{CD}_k}(\boldsymbol{\theta}^0)$ is very unlikely to produce realizations such that $\boldsymbol{g}(\boldsymbol{y}^{\mathrm{obs}}) \notin \mathrm{Conv}\{\boldsymbol{g}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^0,k})\}$, and it is also immune to the problem of $\boldsymbol{\theta}^0 \notin \boldsymbol{\Theta}_{\mathrm{N}}$. We therefore propose to combine the two approaches.

Fellows (2014) described a framework for contrastive divergence as a variational approximation, provided some guidelines on what proposal kernels are likely to perform well, and advocated using a more efficient Newton-like update of the form

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - [\widetilde{\mathrm{Var}}\{\boldsymbol{z}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t})\}]^{-1}\bar{\boldsymbol{z}}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t,k}), \tag{7}$$

for the special case of non-curved ERGMs. This approach is equivalent to lognormal approximation of Hummel et al. (2012) (with step length γ fixed at 1) and "Robbins–Monro" with $\boldsymbol{\alpha}_t = [\widetilde{\mathrm{Var}}\{\boldsymbol{z}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t})\}]^{-1}$. The author has also recently become aware of a thesis by Hummel (2011) that also discussed ERGM CD inference. Hummel focused on exploring different MCMC kernels, but some computational considerations were also discussed, and we note the overlap where it occurs.

Notably, a Newton-style update like (7) can be approximated for curved ERGMs as well by replacing $\boldsymbol{z}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t})$ with $\boldsymbol{U}_{\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t}}(\boldsymbol{\theta}^t)$: differentiating (3) with respect to $\boldsymbol{\theta}'$ gives

$$\hat{\boldsymbol{U}}_{\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t}}(\boldsymbol{\theta}') \approx -\boldsymbol{\eta}'(\boldsymbol{\theta}')^\top\{\bar{\boldsymbol{z}}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t})\} - \boldsymbol{\eta}'(\boldsymbol{\theta}')^\top\widetilde{\mathrm{Var}}\{\boldsymbol{z}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t})\}\{\boldsymbol{\eta}(\boldsymbol{\theta}') - \boldsymbol{\eta}(\boldsymbol{\theta})\},$$

and differentiating again and treating $\boldsymbol{\eta}'(\boldsymbol{\theta}')$ as constant in $\boldsymbol{\theta}'$ gives $\partial\hat{\boldsymbol{U}}_{\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t}}(\boldsymbol{\theta}')/\partial\boldsymbol{\theta}' \approx \boldsymbol{\eta}'(\boldsymbol{\theta}')^\top\widetilde{\mathrm{Var}}\{\boldsymbol{z}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t})\}\boldsymbol{\eta}'(\boldsymbol{\theta}') = \widetilde{\mathrm{Var}}\{\boldsymbol{\eta}'(\boldsymbol{\theta}')^\top\boldsymbol{z}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t})\} \approx \widetilde{\mathrm{Var}}\{\boldsymbol{\eta}'(\boldsymbol{\theta}^t)^\top\boldsymbol{z}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t})\}$, giving such a Newton update.

Outline and Contributions

We have, in the introduction, provided a detailed overview of the available frequentist techniques for fitting exponential families with intractable normalizing constants in general and ERGMs in particular, of obtaining their initial values, and the advantages and the disadvantages of these approaches, also noting where and how the approaches can be directly extended to curved exponential families. Next, we extend MCMLE Partial Stepping technique of Hummel et al. (2012) to curved ERGMs in Section 2. Though we note the overlap with Hummel's work, in Section 3, we provide explicit motivation for applying MCMLE-like approach to CD estimation (as opposed to gradient-based techniques of Carreira-Perpiñan and Hinton (2005)). Focusing on CD as a source of initial values for the MLE estimation, we discuss associated practical issues such as impact of algorithmic choices and of tuning parameters, propose stopping criteria, ways to improve the approximation that do not require additional model-specific work from the implementer, and inexpensive ways to select starting values from among several options. In Section 4, we report a computational study, testing these techniques against network data and models previously considered in ERGM computation literature, providing a systematic comparison between the popular methods for obtaining initial values for ERGM estimation and compare and contrast the proposed technique's variants, while gaining some intuition for the tuning parameters they require and the limitations of the proposed approaches.

2. Partial Stepping for Curved ERGMs

Hummel et al. (2012) derive Partial Stepping and the adaptive selection of the step length γ for non-curved ERGMs. Using their approach with curved models is likely to result in unnecessarily conservative step lengths, however. To see why, consider a popular Geometrically Weighted Degrees (GWD) (Hunter and Handcock, 2006, eq. 4.8) ERGM term. In our notation, this term has two free parameters, θ_1 (the strength of the effect) and θ_2 (decay rate), which map to $(n-1)$-subvectors of $\boldsymbol{\eta}(\cdot)$ and $\boldsymbol{g}(\cdot)$ having elements

$$\begin{aligned}\eta_i(\boldsymbol{\theta}) =& \theta_1 \exp(2\theta_2)[\{1-\exp(-\theta_2)\}^i - 1 + i\exp(-\theta_2)] \\ g_i(\boldsymbol{y}) =& \textstyle\sum_{j=1}^n 1_{|\boldsymbol{y}_j|=i},\end{aligned}$$

for $i = 1, \ldots, n-1$, with $|\boldsymbol{y}_j|$ being the degree of actor j. That is, for every degree value i, $\boldsymbol{\eta}(\boldsymbol{\theta})$ has an element with a coefficient proportional to θ_1 and decaying in i at a rate controlled by θ_2, and $\boldsymbol{g}(\boldsymbol{y})$ has an element with the count of actors with degree exactly i.

The sufficient statistic therefore includes the full degree distribution of the network. A necessary, though not sufficient, requirement for (4) to hold for a given γ is that

$$\min_s g_i(\boldsymbol{y}^{\boldsymbol{\theta}^t,s}) < 1.05\gamma g_i(\boldsymbol{y}^{\text{obs}}) + (1-1.05\gamma)\bar{g}_i(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t}) < \max_s g_i(\boldsymbol{y}^{\boldsymbol{\theta}^t,s})$$

holds for every degree value i, and applying Partial Stepping to $\boldsymbol{g}(\cdot)$ itself would select γ accordingly, as if every element of $\boldsymbol{\eta}$ were a free parameter, even though the actual dimension of $\boldsymbol{\theta}$ is much smaller.

To address this problem, we observe that (1) can be expressed as

$$\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{\eta}'(\boldsymbol{\theta})^\top \boldsymbol{g}(\boldsymbol{y}^{\text{obs}}) - \boldsymbol{\eta}'(\boldsymbol{\theta})^\top \operatorname{E}\{\boldsymbol{g}(\boldsymbol{Y});\boldsymbol{\theta}\},$$

which suggests that for curved ERGMs, we might use γ such that

$$1.05\gamma\boldsymbol{\eta}'(\boldsymbol{\theta}^t)^\top\boldsymbol{g}(\boldsymbol{y}^{\text{obs}}) + (1-1.05\gamma)\boldsymbol{\eta}'(\boldsymbol{\theta}^t)^\top\bar{\boldsymbol{g}}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t}) \in \operatorname{Conv}\{\boldsymbol{\eta}'(\boldsymbol{\theta}^t)^\top\boldsymbol{g}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t})\}.$$

Our generalization does not provide the same guarantees as using the raw $\boldsymbol{g}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t})$, since $\boldsymbol{\eta}'(\boldsymbol{\theta}^t)$ is not constant in $\boldsymbol{\theta}^t$, but it gives each element of $\boldsymbol{g}(\cdot)$ its due weight.

3. Contrastive Divergence via Monte Carlo MLE

3.1. Motivation

Just as the algorithm in Section 1.1.2 solves the score equations (1), we can apply the importance sampling paradigm to solving (6). For the special case of CD_1 with a Metropolis–Hastings sampler with proposal density $q(\cdot|\cdot)$,

$$\operatorname{E}_{\text{CD}_1}\{\boldsymbol{z}(\boldsymbol{Y});\boldsymbol{\theta}\} = \sum_{\boldsymbol{y}'\in\mathcal{Y}\backslash\{\boldsymbol{y}^{\text{obs}}\}} q(\boldsymbol{y}'|\boldsymbol{y}^{\text{obs}}) \min\left[1, \frac{q(\boldsymbol{y}^{\text{obs}}|\boldsymbol{y}')}{q(\boldsymbol{y}'|\boldsymbol{y}^{\text{obs}})}\exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top\boldsymbol{z}(\boldsymbol{y}')\}\right]\boldsymbol{z}(\boldsymbol{y}'),$$

since for rejections, $\boldsymbol{y}' \equiv \boldsymbol{y}^{\text{obs}}$, so $\boldsymbol{z}(\boldsymbol{y}') = \boldsymbol{0}$. Since

$$\operatorname{E}_{\text{CD}_1}\{\boldsymbol{z}(\boldsymbol{Y});\boldsymbol{\theta}'\} = \operatorname{E}_{\text{CD}_1}\left(\frac{\min\left[\frac{q(\boldsymbol{Y}|\boldsymbol{y}^{\text{obs}})}{q(\boldsymbol{y}^{\text{obs}}|\boldsymbol{Y})}, \exp\{\boldsymbol{\eta}(\boldsymbol{\theta}')^\top\boldsymbol{z}(\boldsymbol{Y})\}\right]}{\min\left[\frac{q(\boldsymbol{Y}|\boldsymbol{y}^{\text{obs}})}{q(\boldsymbol{y}^{\text{obs}}|\boldsymbol{Y})}, \exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top\boldsymbol{z}(\boldsymbol{Y})\}\right]}\boldsymbol{z}(\boldsymbol{Y});\boldsymbol{\theta}\right),$$

the importance sampling estimator for $\operatorname{E}_{\text{CD}_1}\{\boldsymbol{z}(\boldsymbol{Y});\boldsymbol{\theta}'\}$ based on a sample $\vec{\boldsymbol{y}}^{\boldsymbol{\theta},1} = (\boldsymbol{y}^{\boldsymbol{\theta}^t,1,1},\dots,\boldsymbol{y}^{\boldsymbol{\theta}^t,1,S})$ drawn from $\text{ERGM}_{\text{CD}_1}(\boldsymbol{\theta})$ is

$$\hat{\operatorname{E}}_{\vec{\boldsymbol{y}}^{\boldsymbol{\theta},1}}\{\boldsymbol{z}(\boldsymbol{Y});\boldsymbol{\theta}'\} = \frac{1}{S}\sum_{s=1}^{S}\frac{\min\left[\frac{q(\boldsymbol{y}^{\boldsymbol{\theta},1,s}|\boldsymbol{y}^{\text{obs}})}{q(\boldsymbol{y}^{\text{obs}}|\boldsymbol{y}^{\boldsymbol{\theta},1,s})}, \exp\{\boldsymbol{\eta}(\boldsymbol{\theta}')^\top\boldsymbol{z}(\boldsymbol{y}^{\boldsymbol{\theta},1,s})\}\right]}{\min\left[\frac{q(\boldsymbol{y}^{\boldsymbol{\theta},1,s}|\boldsymbol{y}^{\text{obs}})}{q(\boldsymbol{y}^{\text{obs}}|\boldsymbol{y}^{\boldsymbol{\theta},1,s})}, \exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top\boldsymbol{z}(\boldsymbol{y}^{\boldsymbol{\theta},1,s})\}\right]}\boldsymbol{z}(\boldsymbol{y}^{\boldsymbol{\theta},1,s}). \tag{8}$$

If the ratios of $q(\cdot|\cdot)$ are recorded during the sampling, this could be implemented directly; and similarly—although with complications—for $k > 1$. In practice, MCMLE weights $(\exp[\{\boldsymbol{\eta}(\boldsymbol{\theta}') - \boldsymbol{\eta}(\boldsymbol{\theta})\}^\top \boldsymbol{z}(\boldsymbol{y}^{\boldsymbol{\theta},1,s})])$ can be used instead: the importance weight in (8) for a given $\boldsymbol{y}^{\boldsymbol{\theta},1,s}$ is monotonically increasing in $\boldsymbol{\eta}(\boldsymbol{\theta}')^\top \boldsymbol{z}(\boldsymbol{y}^{\boldsymbol{\theta},1,s})$, with the weights being equal (to 1) if $\boldsymbol{\theta}' = \boldsymbol{\theta}$, so using the MCMLE weights will, at worst, make the approximation somewhat worse when $\boldsymbol{\theta}'$ is far away from $\boldsymbol{\theta}$, but if (8) evaluated at $\boldsymbol{\theta}' = \boldsymbol{\theta}^{t+1}$ is close to $\mathbf{0}$ and $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t$, we can be confident that the optimization has converged. For higher k, the distribution $\mathrm{ERGM}_{\mathrm{CD}_k}(\boldsymbol{\theta})$ of the sample will be closer to $\mathrm{ERGM}(\boldsymbol{\theta})$, so this approximation will only improve.

3.2. Algorithm

This leads to a CD update of the form

$$\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}'} \log \frac{1}{S} \sum_{s=1}^{S} \exp[-\{\boldsymbol{\eta}(\boldsymbol{\theta}') - \boldsymbol{\eta}(\boldsymbol{\theta}^t)\}^\top \boldsymbol{z}(\boldsymbol{y}^{\boldsymbol{\theta}^t,k,s})].$$

(Hummel (2011, eq. 4.3) used a similar update in the context of CD for non-curved ERGMs, but did not motivate the use of MCMLE importance sampling weights explicitly.) It has a number of appealing properties. From the implementation point of view, the only change required to turn MCMLE into CD is modifying the MCMC sampler to revert the chain to $\boldsymbol{y}^{\mathrm{obs}}$ every k steps, and any improvements to the sampling algorithm also improve the estimator.

From the computational cost point of view, in MLE methods, every new guess $\boldsymbol{\theta}^t$ requires a long burning-in period, a fixed cost that cannot be reduced by parallel processing, and $\boldsymbol{y}^{\boldsymbol{\theta}^t,s}$ tend to be autocorrelated, which encourages using a large S and fewer iterations. But, as $\boldsymbol{\theta}'$ in (2) moves farther away from $\boldsymbol{\theta}^t$, the accuracy of the estimate decreases. On the other hand, $\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t,k}$ is a random sample, requiring a total of Sk MCMC steps per iteration, and the sampling is an embarrassingly parallel problem. This means that a series of relatively short, inexpensive CD steps can be used to obtain an initial value.

To ameliorate potential problems with using MCMLE weights rather than true weights, we propose to use the Hummel et al. (2012) Partial Stepping technique with a more conservative γ safety margin than the Hummel et al. (2012) default of 1.05. (Hummel (2011, p. 77) CD implementation also uses 1.05. We explore its effects in Section 4.) Whether their lognormal approximation should be used is less clear. Its Newton-like update (7) is optimal if

$\boldsymbol{g}(\vec{\boldsymbol{Y}}^{\boldsymbol{\theta}^t,k,s})$ is well approximated by the multivariate normal distribution and the relationship between $\boldsymbol{\theta}$ and $\boldsymbol{U}_{\mathrm{CD}_k}(\boldsymbol{\theta})$ is well approximated by linear over the magnitude of the update, but, for modest k, this is highly unlikely to be the case: for example, if $g(\boldsymbol{y}) = |\boldsymbol{y}|$, the number of edges in the network, for any MCMC step that toggles one potential tie at a time $\boldsymbol{g}(\boldsymbol{y}^{\boldsymbol{\theta}^t,1,s})$ can be one of only three values: $\boldsymbol{g}(\boldsymbol{y}^{\mathrm{obs}}) - 1$, $\boldsymbol{g}(\boldsymbol{y}^{\mathrm{obs}})$, or $\boldsymbol{g}(\boldsymbol{y}^{\mathrm{obs}}) + 1$.

At the same time, although every MCMC step reduces the Kullback–Leibler divergence between $\mathrm{ERGM}(\boldsymbol{\theta})$ and $\mathrm{ERGM}_{\mathrm{CD}_k}(\boldsymbol{\theta})$ (Cover and Thomas, 1991, Thm. 15.1.10, for example), a full-conditional Gibbs sampler is likely to do so faster than a Metropolis–Hastings sampler with the same block size, at least at first. MPLE is equivalent to CD with full-conditional Gibbs sampling (Hyvrinen, 2006), while Metropolis–Hastings is more practical for ERGMs (Hunter et al., 2008), so it is likely that MPLE will outperform CD_1, and Fellows (2014), in particular, focuses on full-conditional Gibbs.

3.3. Artificial multiplicity

Fellows (2014) also shows that increasing k alone may not be sufficiently effective at improving the estimators, and suggests that CD kernels should instead be designed to "focus" on the dependencies in the model: if blocked contrastive divergence (Asuncion et al., 2010) is used for, say, a network model with triadic closure, the "blocks" should include triads.

Unfortunately, specialized proposals negate the advantage of CD as a source of initial values: it is no longer a drop-in replacement for MCMC. Therefore, we propose an *ad hoc* remedy by modifying the Metropolis–Hastings algorithm to create artificial blocks of proposals. Recall that, given a proposal distribution $q(\cdot|\cdot)$, the acceptance probability

$$\alpha(\boldsymbol{y}^\star|\boldsymbol{y}) = \min\left(1, \{q(\boldsymbol{y}|\boldsymbol{y}^\star)/q(\boldsymbol{y}^\star|\boldsymbol{y})\} \exp[\boldsymbol{\eta}(\boldsymbol{\theta})^\top\{\boldsymbol{g}(\boldsymbol{y}^\star) - \boldsymbol{g}(\boldsymbol{y})\}]\right).$$

For MCMC, a simple proposal that toggles only one dyad, or the minimal number of dyads needed to preserve a constraint, generally suffices. A more complex proposal can be emulated by chaining m simple proposals, i.e., $\boldsymbol{y}^{\star 1} \sim q(\boldsymbol{y}^{\star 1}|\boldsymbol{y}), \boldsymbol{y}^{\star 2} \sim q(\boldsymbol{y}^{\star 2}|\boldsymbol{y}^{\star 1}), \ldots, \boldsymbol{y}^{\star m} \sim q(\boldsymbol{y}^{\star m}|\boldsymbol{y}^{\star m-1})$, then accepting $\boldsymbol{y}^{\star m}$ with probability

$$\alpha(\boldsymbol{y}^{\star m}|\boldsymbol{y}) = \min\left(1, \frac{q(\boldsymbol{y}|\boldsymbol{y}^{\star 1})}{q(\boldsymbol{y}^{\star 1}|\boldsymbol{y})} \cdots \frac{q(\boldsymbol{y}^{\star m-1}|\boldsymbol{y}^{\star m})}{q(\boldsymbol{y}^{\star m}|\boldsymbol{y}^{\star m-1})} \exp[\boldsymbol{\eta}(\boldsymbol{\theta})^\top\{\boldsymbol{g}(\boldsymbol{y}^{\star m}) - \boldsymbol{g}(\boldsymbol{y})\}]\right),$$

remaining at $\boldsymbol{y}$ otherwise. This is not the correct acceptance probability (because a correct one would consider all possible ways to propose $\boldsymbol{y}^{\star m}$ from

$\boldsymbol{y}$), so m is a trade-off between the correctness of the stationary distribution and incorporation of the dependence in the model.

But, an approximation is what we require. We will use $\tilde{\boldsymbol{\theta}}^{(m,k)}$ to refer to a $\mathrm{CD}_{(m,k)}$ estimate, taking k steps with artificial multiplicity m.

3.4. Stopping Criterion

We briefly turn to the question of when to consider the optimization to be concluded. The stopping criterion of Hummel et al. (2012) is not well-suited to CD, because for small $m \times k$ in particular, it may not be possible for $\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t,(m,k)}$ to draw sufficiently far away from $\boldsymbol{y}^{\text{obs}}$ for Hummel et al. for $\text{Conv}\{\boldsymbol{g}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t,(m,k)})\}$ to not contain $\boldsymbol{g}(\boldsymbol{y}^{\text{obs}})$, no matter how bad $\boldsymbol{\theta}^t$ is.

The forms of the estimating equations (1) and (6) suggest another straightforward method to determine whether a particular $\boldsymbol{\theta}^t$ is sufficiently close to $\tilde{\boldsymbol{\theta}}^{(m,k)}$ to stop. For each guess $\boldsymbol{\theta}^t$, CD draws a simple random sample $\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t,(m,k)}$ from $\text{ERGM}_{\mathrm{CD}_{(m,k)}}(\boldsymbol{\theta}^t)$, $\bar{\boldsymbol{g}}(\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t,(m,k)})$ is an unbiased estimator of $\mathrm{E}_{\mathrm{CD}_{(m,k)}}\{\boldsymbol{g}(\boldsymbol{Y});\boldsymbol{\theta}^t\}$, and premultiplication by $\boldsymbol{\eta}'(\boldsymbol{\theta}^t)$ is a linear transformation, so $\bar{\boldsymbol{U}}_{\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t,(m,k)}}(\boldsymbol{\theta}^t)$ is unbiased for $\boldsymbol{U}_{\mathrm{CD}_{(m,k)}}(\boldsymbol{\theta}^t)$.

Therefore, we can use a Hotelling's T^2-Test (Hotelling, 1931) to test $H_0 : \boldsymbol{U}_{\mathrm{CD}_{(m,k)}}(\boldsymbol{\theta}^t) = \boldsymbol{0}$ based on a sample $\boldsymbol{U}_{\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t,(m,k)}}(\boldsymbol{\theta}^t)$, stopping upon a failure to reject. The decision to terminate entails accepting a null hypothesis, but this can be ameliorated in practice by setting a very high α, because the cost of a Type I error is small: setting $\alpha = 0.5$ only entails running on average $1/\alpha = 2$ more iterations than necessary.

3.5. Choice of k and m

The choice of k is a trade-off: higher k leads to $\tilde{\boldsymbol{\theta}}^{(m,k)}$ being closer to $\hat{\boldsymbol{\theta}}$, but the computing cost increases in proportion to it, and sensitivity to poor $\boldsymbol{\theta}^0$ does as well, and a similar trade-off (up to a point) applies for m.

Our goal is to maximize the utility of the CD estimate as the starting value of MCMLE, and a simple one-dimensional metric of this utility is available: the Hummel et al. (2012) adaptive step length for the first MCMLE iteration (4). This is, essentially, a measurement of how deep in the convex hull of $\boldsymbol{\eta}'(\tilde{\boldsymbol{\theta}}^{(m,k)})^\top \boldsymbol{g}(\vec{\boldsymbol{y}}^{\tilde{\boldsymbol{\theta}}^{(m,k)}})$ is $\boldsymbol{\eta}'(\tilde{\boldsymbol{\theta}}^{(m,k)})^\top \boldsymbol{g}(\boldsymbol{y}^{\text{obs}})$. An estimated step length of 1 or close implies that only a few steps of full MCMLE will be required, while a step length close to 0 implies that the starting value is practically useless.

We therefore propose to evaluate $\tilde{\boldsymbol{\theta}}^{(m,k)}$ for a series of (m, k) configurations, then, for each estimate $\tilde{\boldsymbol{\theta}}^{(m,k)}$, draw an MCMC sample, evaluate the adaptive step length, and initialize the MCMLE with the one giving the highest γ such that (4) holds. Because MCMLE step requires a long burn-in, this is likely to be computationally expensive, but we can, instead, use a proxy in the form of a short MCMC run that would nonetheless have a burn-in period much longer than the highest value of $m \times k$ used.

4. Examples

In this section, the proposed techniques are illustrated by replicating examples found in the ERGM computational methods literature. We list the examples here, identifying the computational challenge of each; more details about the data and the models are given in the Appendix.

Lazega, a collaboration network of lawyers, was used by Hunter and Handcock (2006) to demonstrate inference for curved ERGMs, fitting a curved ERGM conditional having a specific number of ties—a complex constraint. (We also replicate the curved fit without the constraint, modeling edge count.)

E. coli, a transcriptional regulation network, was selected by Hummel et al. (2012) for being particularly difficult to fit.

Kapferer, a network of workers in a tailor shop in Zambia, was also used by Hummel et al. (2012).

Zachary, a valued network of counts of contexts of interactions among members of a university karate club, which we use to to demonstrate immediate applicability to models for valued networks, fitting a Binomial- and a Poisson-reference ERGM. (For the latter, we include the CMP (Shmueli et al., 2005) term, deliberately initializing CD with a starting value outside of $\mathbf{\Theta}_{\mathrm{N}}$ to test the algorithm's robustness.)

4.1. Procedure

We have implemented the proposed techniques in the R (R Core Team, 2015) package `ergm` (Hunter et al., 2008; Handcock et al., 2015) and released them on an experimental basis. The source code for the required packages, code to reproduce this study, and the datasets in machine-readable format can be found in the supplementary materials.

We refrain from tuning the algorithms to each specific dataset, and unless otherwise stated, we use default settings of the `ergm` package. For CD, we use $S = 1024$, start the estimation at $\mathbf{0}_q$ (unless otherwise noted), and allow 60 iterations. For each example, we evaluate the the intercept-based estimate and

MPLE (where available: for binary ERGMs without complex constraints), and CD for each combination of $k = 1, 2, 4, 16, 128$ and $m = 1, 2, 4, 8$ such that $k \times m \leq 256$. For each (m, k) combination, we estimate $\tilde{\boldsymbol{\theta}}^{(m,k)}$ using MCMLE and using stochastic approximation updates. For MCMLE, we try every combination of likelihood approximation type ("IS MCMLE", the importance-sampling (2), and "Lognormal" (3)) and γ margins 1.05 (used by Hummel et al. (2012)), 1.5 and 2 (more conservative). For stochastic approximation, we consider a Newton-like update (7) generalized to curved ERGMs, i.e., $\boldsymbol{\alpha}_t \equiv [\widetilde{\mathrm{Var}}\{\boldsymbol{U}_{\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t,(m,k)}}(\boldsymbol{\theta}^t)\}]^{-1}$, and two Robbins–Monro regimes suggested by Snijders (2002): $\boldsymbol{\alpha}_t \equiv (a_0 t^{-c})[\widetilde{\mathrm{Var}}\{\boldsymbol{U}_{\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t,(m,k)}}(\boldsymbol{\theta}^t)\}]^{-1}$ for $(a_0, c) = (0.1, 0.5)$ and $(a_0, c) = (0.5, 0.5)$. (If the sample variance–covariance matrix is singular for a given iteration, we use the Moore–Penrose pseudoinverse.)

Having found the $\tilde{\boldsymbol{\theta}}$ according to each method and parameters, we measure its utility as a starting value in two ways:

$\boldsymbol{\gamma}_{\mathbf{S}}$: To test the suggestion of Section 3.5, we generate an MCMC sample from ERGM($\tilde{\boldsymbol{\theta}}$) starting at $\boldsymbol{y}^{\mathrm{obs}}$ with burn-in 8192, sample size 1024, and interval 8 (longer than any $k \times m$, but still shorter than the defaults for `ergm`), then evaluate adaptive step length γ as proposed by Hummel et al. (2012) or our extension in Section 2.

MLE: We also proceed to use these values to initialize full MLE estimation and record the reliability of the estimation and the number of iterations it took before convergence.

4.2. Results

For a given (m, k), all CD algorithms considered are estimating the same quantity, so they can mainly be compared on speed and reliability. Table 1 gives the effects of the update type and its parameters on those outcomes. (All use the same convergence criterion.) Importance sampling MCMLE updates as opposed to the Newton-like lognormal updates appear to be a trade-off between speed and stability, with MCMLE making more efficient steps, at a greater risk of making a poor step. A more conservative γ margin alleviates this, while retaining the efficiency improvement. For $S = 1024$, using even a aggressive Robbins–Monro regime appears to be counterproductive, though Newton's update performs about as well as lognormal with a small γ margin (by virtue of being nearly mathematically equivalent), while showing an advantage in overall run time, likely because of its simplicity.

Turning to comparing the distinct estimators, the effects of m and k and their comparison with the intercept method and the MPLE are visualized

Table 1: Aggregate effects of the update type and parameters (approximation type and γ margin for MCMLE and initial gain (α_0) and decay rate (c) for Robbins–Monro) on quality, speed, and reliability. Means are taken after standardizing each value by its example's overall mean and standard deviation. Failed fits are treated as having taken 60 iterations. Error usually means that the estimation procedure was stuck in a very poor configuration, and a procedure was considered Unconverged if it did not meet the convergence criterion after 60 iterations.

Monte Carlo MLE updates

Settings		Cost (mean)		Failures	
Approximation	γ mar.	Iter.	$\frac{\text{sec.}}{m\times k}$	Error	Unconv.
IS MCMLE	1.05	−0.47	−0.10	1%	13%
IS MCMLE	1.50	−0.51	−0.15	0%	6%
IS MCMLE	2.00	−0.38	−0.04	0%	6%
Lognormal	1.05	−0.47	−0.17	1%	5%
Lognormal	1.50	−0.39	−0.09	0%	5%
Lognormal	2.00	−0.27	0.02	0%	7%

Stochastic Approximation updates

Settings		Cost (mean)		Failures	
α_0	c	Iter.	$\frac{\text{sec.}}{m\times k}$	Error	Unconv.
0.10	0.50	1.68	0.65	0%	100%
0.50	0.50	1.25	0.36	0%	58%
Newton		−0.47	−0.50	1%	5%

in Figure 1, using number of iterations taken by a subsequent MCMC MLE fit as a proxy for quality of starting values, imputing 20 (the maximum) if the estimation fails. The general pattern appears to be that MPLE, where available, outperforms CD with small k and m, but CD eventually matches it, except in particularly hard-to-sample models such as the *E. coli* with no self-loops. At the same time, there are diminishing returns as $m \times k$ increases, and, in valued ERGMs, they actually perform worse. (Interestingly, in the harder-to-sample models, the MPLE is not that much better than the intercept method.)

`ergm` with default settings appears to have difficulty given *any* starting value for the full Kapferer model. For the hard-to-sample *E. coli* with loops, higher artificial multiplicities seem to outperform lower for the same $m \times k$, but the results are less consistent for other ERGMs, and, in particular, for the valued ERGMs and the fixed-edges model, whose proposal is already multiplicitous; this may be because there are many more possible ways to a given $\boldsymbol{y}^{\star m}$ from $\boldsymbol{y}$ in those cases, which $\alpha(\boldsymbol{y}^{\star m}|\boldsymbol{y})$ ignores. Nevertheless, using $m = 2$ seems to be safe and an improvement in all cases.

The relationship between γ_{S} and the quality of the starting value is given in Figure 2. It appears to be highly predictive of the success of the estimate for all cases except for *E. coli*, including picking out the only two successful initial values for the full Kapferer model. At the same time, the *E. coli* fits suggest that for hard-to-sample models, too short pilot runs may result in selecting a poor start. It may be possible to use a burn-in diagnostic to determine when this is the case.

In the CMP model, CD using MCMLE was able to locate an adequate $\boldsymbol{\theta}^0 \in \boldsymbol{\Theta}_{\mathrm{N}}$ in 95% of the trials.

5. Conclusion

We have reviewed the available techniques for obtaining initial values for the simulation-based MLE methods for exponential family models with intractable normalizing constants, and, combining the approaches of Monte Carlo MLE and contrastive divergence, we proposed a fairly universal algorithm for obtaining these values, providing an empirical comparison of different approaches to the problem. In addition, we have extended to curved ERGMs the existing techniques for improving the stability of MCMLE.

Our examples demonstrate the viability and versatility of our approach: adequate starting values are produced for a wide variety of datasets and

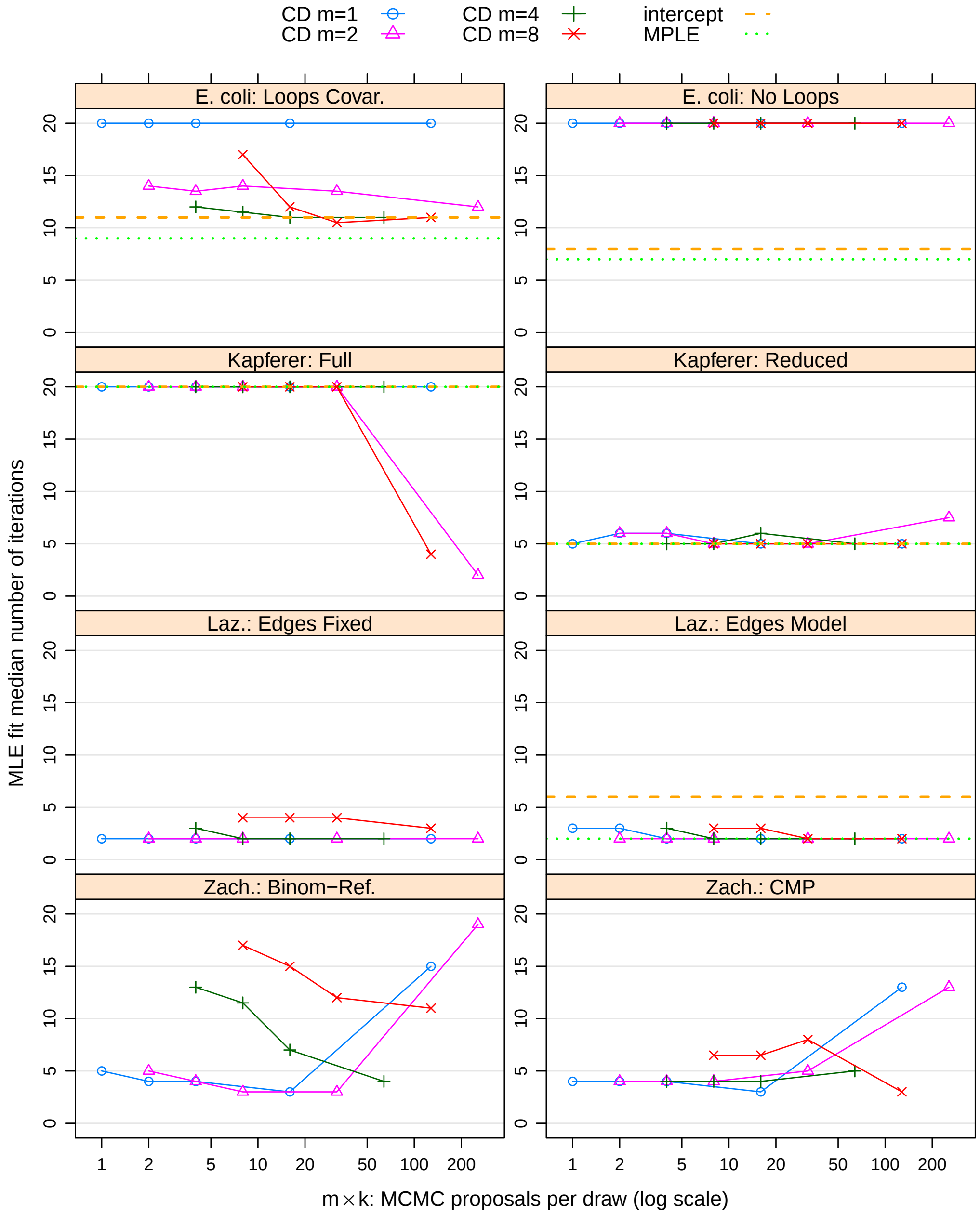

Figure 1: Effect of (m, k) on the quality of the starting value as measured by the number of iterations taken by the subsequent call to MLE estimation to converge (with 20 recorded on failure). Values are medians of iteration counts pooled over the most *reliable* settings (IS MCMLE with γ margin 1.5 or higher, lognormal, and stochastic approximation with Newton-style updates).

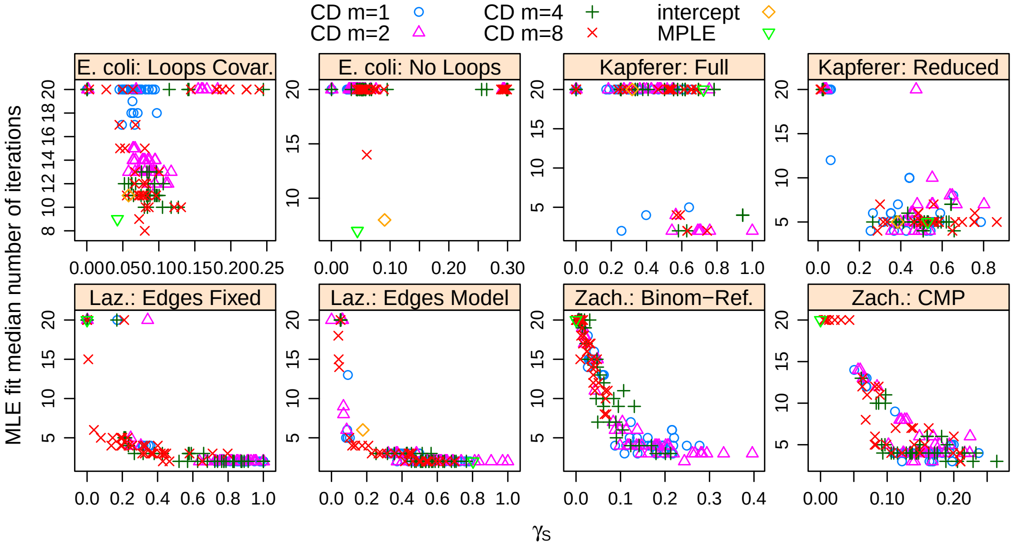

Figure 2: Using Hummel step length from a short run to predict the number of iterations for the full MCMC MLE run. The median value is given for each distinct source and configuration of starting values.

models—some designed to be difficult—with an algorithm agnostic to the specifics of the model. In practice, this means that any implementation of MCMLE for a new valued or constrained ERGM class (e.g., rank or signed networks) acquires a source of starting values without additional effort.

At the same time, we exposed limitations of this approach: barring further improvements in CD estimation, if an MPLE implementation is available, it should probably be preferred as a source of starting values. On the other hand, we have shown that short pilot MCMC runs can be used to select an adequate starting value for the MCMC out of several candidates, which are themselves inexpensive to fit. Thus, one may include MPLE as one of an ensemble of initial value methods, then pick the most promising ones to seed the much more time-consuming MCMC MLE.

An alternative approach to selecting (m, k) may be to use an increasing sequence of ks, initializing each at the previous one's solution as its stopping criterion is met. This approach should be used with caution, however, because $\tilde{\boldsymbol{\theta}}$ based on a small k can be worse than $\tilde{\boldsymbol{\theta}} = \mathbf{0}$. This is subject for future research.

We had focused on the case where the networks were fully observed.

Handcock and Gile (2010) formulated a framework for modeling of partially observed networks—networks that have missing ties—and expressed the log-likelihood as $\ell(\boldsymbol{\theta}) = \log \Pr(\boldsymbol{Y} \in \mathcal{Y}(\boldsymbol{y}^{\text{obs}}); \boldsymbol{\theta}) = \log \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{y}^{\text{obs}})} \Pr(\boldsymbol{Y} = \boldsymbol{y}'; \boldsymbol{\theta})$, where $\mathcal{Y}(\boldsymbol{y}^{\text{obs}})$ is defined as the set of networks whose partial observation could have produced $\boldsymbol{y}^{\text{obs}}$: essentially, all of the ways to impute the missing ties in $\boldsymbol{y}^{\text{obs}}$. They then proposed to maximize this likelihood by taking advantage of the fact that, if $\kappa_{\mathcal{Y}'}(\boldsymbol{\theta}) \overset{\text{def}}{=} \sum_{\boldsymbol{y}' \in \mathcal{Y}'} h(\boldsymbol{y}') \exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top \boldsymbol{g}(\boldsymbol{y}')\}$, log-likelihood can be expressed as $\ell(\boldsymbol{\theta}) = \log \kappa_{\mathcal{Y}(\boldsymbol{y}^{\text{obs}})}(\boldsymbol{\theta}) - \log \kappa_{\mathcal{Y}}(\boldsymbol{\theta})$, resulting in

$$\boldsymbol{U}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ell(\hat{\boldsymbol{\theta}}) = \boldsymbol{\eta}'(\hat{\boldsymbol{\theta}})^\top [\mathrm{E}_{\mathcal{Y}(\boldsymbol{y}^{\text{obs}})}\{\boldsymbol{g}(\boldsymbol{Y}); \hat{\boldsymbol{\theta}}\} - \mathrm{E}_{\mathcal{Y}}\{\boldsymbol{g}(\boldsymbol{Y}); \hat{\boldsymbol{\theta}}\}] = \boldsymbol{0},$$

with MCMLE approximation also possible for the first term by sampling $\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t}|\boldsymbol{y}^{\text{obs}}$ from $\mathrm{ERGM}_{\mathcal{Y}(\boldsymbol{y}^{\text{obs}})}(\boldsymbol{\theta}^t)$. For CD, this creates a problem: while $\vec{\boldsymbol{y}}^{\boldsymbol{\theta}^t}|\boldsymbol{y}^{\text{obs}}$ depends on $\boldsymbol{y}^{\text{obs}}$ only through the observed dyads and information about which dyads are missing due to the ergodic property of MCMC, sampling from $\mathrm{ERGM}_{\mathrm{CD}_{(m,k)}}(\boldsymbol{\theta}^t)$ requires a specific initial network and depends on it strongly. In the context of CD, these problems can be partially addressed by using higher ks: the longer the MCMC chain, the less important $\boldsymbol{y}^{\text{obs}}$, but more efficient and stable approaches are subject for research. (A similar issue exists for the MPLE: the composite likelihood is a sum of (5) over possible imputations of missing dyads in $\boldsymbol{y}^{\text{obs}}$, and simply excluding the unobserved dyads from the product (5) still conditions on them.)

A network might not be observed at all, only its sufficient statistic vector $\boldsymbol{g}^{\text{obs}}$ along with its sample space $\mathcal{Y}$. By sufficiency, MLE is unaffected by this. (Hummel et al., 2012) MPLE, MCLE, and CD are, however. A simple practical solution is to use simulated annealing to construct a network $\boldsymbol{y}^{\text{sim}}$ such that $\boldsymbol{g}(\boldsymbol{y}^{\text{sim}}) \approx \boldsymbol{g}^{\text{obs}}$ and use it as a surrogate for $\boldsymbol{y}^{\text{obs}}$. It may not be possible to obtain a perfectly matched network, but this can be addressed in the same way as with missing data.

Lastly, we have focused on ERGMs in particular, but these methods are agnostic to the nature of the data, operating only on sufficient statistics, so this development is equally applicable to other domains. In particular, the problem of a complex $\boldsymbol{\Theta}_{\mathrm{N}}$ is present in Strauss and related point processes as well (Geyer and Thompson, 1992, for example).

Acknowledgements

Calculations were performed on the National Institute for Applied Statistics Research Australia (NIASRA) High Performance Computing cluster. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The author wishes to thank the editor, the associate editor, and two anonymous reviewers for their invaluable comments; and Dr. Ian Fellows for helpful discussions, particularly for pointing out to the author that the problem of conditioning on missing dyads mentioned in the Discussion applies to MPLE as well.

References

Asuncion, A. U., Liu, Q., Ihler, A. T., Smyth, P., 2010. Learning with blocks: Composite likelihood and contrastive divergence. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-10).
URL `http://machinelearning.wustl.edu/mlpapers/papers/AISTATS2010_AsuncionLIS10`

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion). Journal of the Royal Statistical Society, Series B 36, 192–236.

Carreira-Perpiñan, M. A., Hinton, G., 2005. On contrastive divergence learning. In: Cowell, R. G., Ghahramani, Z. (Eds.), Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, Jan 6-8, 2005, Savannah Hotel, Barbados. Society for Artificial Intelligence and Statistics, pp. 33–40.
URL `http://www.gatsby.ucl.ac.uk/aistats/`

Cover, T. M., Thomas, J. A., 1991. Elements of Information Theory. Wiley Series in Telecommunications. John Wiley & Sons, Inc.

Fellows, I. E., 2014. Why (and when and how) contrastive divergence works. arXiv preprint arXiv:1405.0602.
URL `https://arxiv.org/abs/1405.0602`

Geyer, C. J., Thompson, E. A., 1992. Constrained Monte Carlo maximum likelihood for dependent data (with discussion). Journal of the Royal Statistical Society. Series B 54 (3), 657–699.

Handcock, M. S., Gile, K. J., 2010. Modeling social networks from sampled data. Annals of Applied Statistics 4 (1), 5–25.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., Morris, M., 2015. `ergm`: Fit, Simulate and Diagnose Exponential-Family Models for Networks. The Statnet Project (`http://www.statnet.org`), R package version 3.4.0.
URL `http://CRAN.R-project.org/package=ergm`

He, R., Zheng, T., 2015. GLMLE: Graph-limit enabled fast computation for fitting exponential random graph models to large social networks. Social Network Analysis and Mining 5 (1).

Hinton, G. E., 2002. Training products of experts by minimizing contrastive divergence. Neural computation 14 (8), 1771–1800.

Hotelling, H., Aug. 1931. The generalization of student's ratio. Annals of Mathematical Statistics 2 (3), 360–378.

Hummel, R. M., may 2011. Improving estimation for exponential-family random graph models. Ph.D. thesis, The Pennsylvania State University.
URL `https://etda.libraries.psu.edu/paper/11493/`

Hummel, R. M., Hunter, D. R., Handcock, M. S., 2012. Improving simulation-based algorithms for fitting ergms. Journal of Computational and Graphical Statistics 21 (4), 920–939.

Hunter, D. R., Handcock, M. S., 2006. Inference in curved exponential family models for networks. Journal of Computational and Graphical Statistics 15 (3), 565–583.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., Morris, M., May 2008. `ergm`: A package to fit, simulate and diagnose exponential-family models for networks. Journal of Statistical Software 24 (3), 1–29.
URL `http://www.jstatsoft.org/v24/i03`

Hunter, D. R., Krivitsky, P. N., Schweinberger, M., 2012. Computational statistical methods ror social network models. Journal of Computational and Graphical Statistics 21 (4), 856–882.

Hyvrinen, A., Oct. 2006. Consistency of pseudolikelihood estimation of fully visible boltzmann machines. Neural Computation 18 (10), 2283–2292.

Kapferer, B., 1972. Strategy and Transaction in an African Factory: African Workers and Indian Management in a Zambian Town. Manchester University Press.

Krivitsky, P. N., 2012. Exponential-family random graph models for valued networks. Electronic Journal of Statistics 6, 1100–1128.

Lazega, E., Pattison, P. E., 1999. Multiplexity, generalized exchange and cooperation in organizations: a case study. Social Networks 21 (1), 67–90.

Lindsay, B. G., 1988. Composite likelihood methods. Contemporary Mathematics 80, 221–239.

Morris, M., Handcock, M. S., Hunter, D. R., May 2008. Specification of exponential-family random graph models: Terms and computational aspects. Journal of Statistical Software 24 (4), 1–24.
URL `http://www.jstatsoft.org/v24/i04`

Okabayashi, S., Geyer, C. J., 2012. Long range search for maximum likelihood in exponential families. Electronic Journal of Statistics 6, 123–147.

R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `http://www.R-project.org/`

Rao, A. R., Jana, R., Bandyopadhyay, S., Jun. 1996. A markov chain Monte Carlo method for generating random (0, 1)-matrices with given marginals. Sankhyā: The Indian Journal of Statistics, Series A 58 (2), 225–242.

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

Robbins, H., Monro, S., Sep. 1951. A stochastic approximation method. The Annals of Mathematical Statistics 22 (3), 400–407.

Robins, G., Pattison, P., Wasserman, S. S., 1999. Logit models and logistic regressions for social networks: III. Valued relations. Psychometrika 64 (3), 371–394.

Shen-Orr, S. S., Milo, R., Mangan, S., Alon, U., May 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nature Genetics 31 (1), 64–68.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., Boatwright, P., Jan. 2005. A useful distribution for fitting discrete data: Revival of the Conway–Maxwell–Poisson distribution. Journal of the Royal Statistical Society: Series C 54 (1), 127–142.

Snijders, T. A. B., 2002. Markov chain Monte Carlo estimation of exponential random graph models. Journal of Social Structure 3 (2).

Snijders, T. A. B., Pattison, P. E., Robins, G. L., Handcock, M. S., 2006. New specifications for exponential random graph models. Sociological Methodology 36 (1), 99–153.

Strauss, D., Ikeda, M., 1990. Pseudolikelihood estimation for social networks. Journal of the American Statistical Association 85 (409), 204–212.

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

Wang, P., Robins, G., Pattison, P., Koskinen, J., Jun. 2014. MPNet User Manual. Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, Australia.
URL `http://sna.unimelb.edu.au/PNet`

Wasserman, S. S., Pattison, P., 1996. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^*. Psychometrika 61 (3), 401–425.

Zachary, W. W., 1977. An information flow model for conflict and fission in small groups. Journal of Anthropological Research 33 (4), 452–473.

Appendix A. Details of the Examples

Appendix A.1. Lazega and Pattison's Law firm

Hunter and Handcock (2006), in their development of inference for curved ERGMs, used data collected by Lazega and Pattison (1999), describing patterns of collaboration of lawyers in a firm. The model they fit included covariates such as the effect of seniority, type of practice, whether the two lawyers had the same practice, were of the same gender, and worked in the same office; and it modeled triadic closure using Alternating k-triangles (also known as Geometrically-Weighted Edgewise Shared Partners (GWESP)), a curved ERGM term.

We fit two variants of their Model 2 to these data: a variant whose sample space was restricted to have the same edge count as the observed network (which is what was fit by Hunter and Handcock) and a variant not conditioned on edge count, but using edge count as an additional model statistic.

Appendix A.2. E. coli *transcriptional regulation network*

Hummel et al. (2012), in illustrating their computational methods on a difficult model, used the *E. coli* transcriptional regulation network of Shen-Orr et al. (2002). Here, we fit two variants demonstrated by Hummel et al., referred to as "Model 2": edge count, counts of actors with degree 2–5 (separately), and Geometrically-Weighted Degree (GWD) term with decay coefficient fixed at 0.25) and "Model 2 plus self-edges", contains all of the above terms and, in addition, nodal covariates indicating whether a node has a non-self-edge and whether it has a self-edge.

Appendix A.3. Kapferer's sociational data

Hummel et al. (2012) also demonstrated their approach on a well-known dataset collected by Kapferer (1972) on workers in a tailor shop in Zambia, and we reproduce the two models they had fit. The first model had, as its terms, count of edges, and the GWD, the GWESP, and the Geometrically-Weighted Dyadwise Shared Partners (GWDSP) statistics, the latter three having their decay coefficient fixed at 0.25. The second model dropped the GWD term.

Appendix A.4. Valued ties in a Zachary's Karate club

For valued ERGMs, the possible intractability of the pseudolikelihood and the possibly complex shape of $\mathbf{\Theta}_{\mathrm{N}}$ for models with infinite sample spaces make the problem of finding $\boldsymbol{\theta}^0$ particularly difficult. We illustrate the contrastive divergence approach to it on data collected by Zachary (1977), who reported observations of social relations in a university karate club with membership that varied between 50 and 100. The actors—32 ordinary club members and officers, the club president ("John A."), and the part-time instructor ("Mr. Hi")—were the ones who consistently interacted outside of the club. Over the course of the study, the club divided into two factions, and, ultimately, split into two clubs, one led by Hi and the other by John and the original club's officers. The split was driven by a disagreement over whether Hi could unilaterally change the level of compensation for his services.

Zachary reported, for each pair of actors, the count of social contexts in which they interacted. The 8 contexts considered were academic classes at the university; Hi's private karate studio in his night classes; Hi's private karate studio where he taught on weekends; student-teaching at Hi's studio; the university rathskeller (bar) located near the karate club; a bar located near the university campus; open karate tournaments in the area; and intercollegiate karate tournaments. The highest number of contexts of interaction for a pair of individuals that was observed was 7.

In Model 1, we model the distribution of counts as a binomial-reference ERGM, i.e., $\mathbb{S} = 0..8$ and $h(\boldsymbol{y}) = \prod_{(i,j)\in\mathbb{Y}} \binom{8}{y_{i,j}}$, zero-modified by adding a term of the form $g_{\text{nonzero}}(\boldsymbol{y}) = \sum_{(i,j)\in\mathbb{Y}} 1_{y_{i,j}\neq 0}$.

In Model 2, we instead use a Poisson-reference ERGM (i.e., having dyadwise sample space of $\mathbb{S} = \{0, 1, 2, \dots\}$) with $h(\boldsymbol{y}) \equiv 1/\prod_{(i,j)\in\mathbb{Y}} y_{i,j}!$, and we include two statistics to affect the dyadwise distribution of counts: g_{nonzero} to control the overall propensity to have ties (i.e., have interactions in more than 0 contexts) and a statistic of the form $g_{\text{CMP}}(\boldsymbol{y}) = \sum_{(i,j)\in\mathbb{Y}} \log(y_{i,j}!)$, which, added to a geometric- or Poisson-reference ERGM models each relationship value as distributed Conway–Maxwell–Poisson (CMP) (Shmueli et al., 2005; Krivitsky, 2012). A linear ERGM with this term—for example, with sufficient statistic $\boldsymbol{g}(\boldsymbol{y}) = (\sum_{(i,j)\in\mathbb{Y}} y_{i,j}, \sum_{(i,j)\in\mathbb{Y}} \log(y_{i,j}!))$, has a constrained natural parameter space $\mathbf{\Theta}_{\mathrm{N}} = \{\boldsymbol{\theta} \in \mathbb{R}^2 : \theta_2 = 1 \wedge \theta_1 < 0 \vee \theta_2 < 1\}$, making it neither regular nor steep (Krivitsky, 2012, App. B). For this reference, we use a Tie-Non-Tie (TNT) (Morris et al., 2008) augmentation of the zero-inflated Poisson algorithm of Krivitsky (2012, Alg. 1).

We model the structure of the network using two more terms: the faction leader effects, $\sum_{(i,j)\in\mathbb{Y}} y_{i,j} 1_{i=\text{Mr. Hi}\vee j=\text{Mr. Hi}}$ and $\sum_{(i,j)\in\mathbb{Y}} y_{i,j} 1_{i=\text{John A.}\vee j=\text{John A.}}$, and transitivity, the statistic described by Krivitsky (2012, eq. 12).

Unlike other fits, where we start the optimization at $\boldsymbol{\theta}^0 = \mathbf{0}_q$, in Model 2, we start the optimization at $\theta^0_{\text{CMP}} = +2$, deliberately outside the parameter space. Also, we change a few non-CD-specific tuning parameters to accommodate non-binary data.