Title

Bayesian Inference for Exponential-family Random Graph Models

ISBN

9798247944201

Author

Resch, Joseph

Publication Date

2026-05-28

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Bayesian Inference for Exponential-family Random Graph Models

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

Joseph Resch

2026

ABSTRACT OF THE DISSERTATION

Bayesian Inference for Exponential-family Random Graph Models

by

Joseph Resch
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2026
Professor Mark S. Handcock, Chair

In network data, such as those representing social and economic relationships, the connections between entities depend on one another. Exponential-family random graph models (ERGMs) are the standard tool for capturing this dependence. They are difficult to fit because their parameter space has a challenging geometry. Large regions of it produce networks that are either nearly empty or nearly complete, neither of which is wanted by practitioners. Common Bayesian methods either ignore this geometry or work around it inefficiently. This dissertation develops methodology that respects it. The first chapter shows that a recent variational method for ERGM estimation introduces a formulation that neglects tie dependence and works poorly in realistic settings. The second chapter proposes a new prior distribution for Bayesian ERGM inference. Unlike standard priors, it concentrates on parameter values that produce realistic networks. The third chapter shows that the new prior generalizes to modern sampling methods that use auxiliary samples. The methods are demonstrated through simulations and applied to a tailor shop network from a 1972 study.

The dissertation of Joseph Resch is approved.

Xiaowu Dai

Ying Nian Wu

Oscar Hernan Madrid Padilla

Mark S. Handcock, Committee Chair

University of California, Los Angeles

2026

To my family and friends.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to acknowledge my advisor Mark Handcock for his support and mentorship throughout this journey. I'm lucky to have been his student, and can easily say that I enjoyed each interaction we had working together. Mark's guidance has had invaluable influence on me as a scientist, both professionally and personally. I thank my committee members Oscar Hernan Madrid Padilla, Ying Nian Wu, and Xiaowu Dai for their time and feedback on this project. Thank you Brooke Pearson, my best friend and partner, for being with me every step of the way. I would not trade your warmth and friendship for the world. I'm grateful for my friends, the reasons that Los Angeles feels like home. Thank you to Tanvi, Alex, Andrew, Davis, Kyla, John, Sophie, Hector, Thomas, Shuchi, Noah, Rose, Rushil, Megan, Advika, and Elizabeth. I'd also like to thank the staff members Jose, Laurie, Kyle, Charlotte, and Chie. Big thank you to my big gorgeous cat, Meater. Thank you to my parents, Coco and Franz, for supporting my interests, without whom none of this would have been possible.

VITA

2016–2020 B.A. in Economics, University of Georgia.

2016–2020 B.S. in Statistics, University of Georgia.

2020–2024 C.Phil. in Statistics, University of California, Los Angeles.

2020– Ph.D. Student in Statistics, University of California, Los Angeles.

PUBLICATIONS

Resch, J. & Handcock, M.S. Assessing Variational Likelihood Estimation in Network Formation Models. *Under Review.*

Resch, J., Baugh, S., Duan, H., Tang, J., Madison, M.J., Cotterell, M. & Jeon, M. (2025). Bayesian Transition Diagnostic Classification Models with Polya-Gamma Augmentation. Psychometrika, 90(4), 1368-1399.

Ranjan, P., Resch, J. & Mandal, A. (2023). Solving an Inverse Problem for Time-Series-Valued Computer Simulators via Multiple Contour Estimation. Journal of Statistical Theory and Practice, 17, 23.

CHAPTER 1

Introduction

Networks are a natural representation of relationships between actors in the social and economic sciences. Friendships, partnerships, and transactions can all be described as relational ties between pairs of individuals or organizations. The configuration of these ties and the dependence among them influence network behaviors, from the spread of information to the formation of communities. Statistical models for networks aim to capture this dependence in an interpretable way that is grounded in theory for the generating process.

Exponential-family random graph models (ERGMs) are one of the most widely used frameworks for this purpose (FS86; HHB03). An ERGM specifies the probability of observing a network as a function of network statistics, such as the number of edges, the tendency for triangles, or the degree of homophily on a nodal attribute. These statistics are chosen to reflect economic or social theory about why ties form. Because the ERGM is an exponential family, it is the unique maximum-entropy distribution on the space of networks whose expected statistics match a given set of moment constraints (FH17). This characterization gives ERGMs a strong theoretical foundation as models for network data.

The main obstacle to working with ERGMs is computational. The normalizing constant of the model requires summing over all possible networks on a fixed set of nodes, a number that grows as $2^{\binom{n}{2}}$ for binary undirected networks on n nodes. Even for moderately sized networks this sum is intractable. Likelihood-based inference therefore depends on approximation, and the choice of approximation method has large consequences for the quality of the resulting estimates. This dissertation addresses three distinct problems for

that approximation.

The first concerns estimation. The dominant approach to maximum likelihood estimation for ERGMs uses Markov chain Monte Carlo (MCMC) to approximate the intractable normalizing constant (GT92; HHB03). MCMC-based maximum likelihood estimation (MCMC-MLE) has a long track record and benefits from decades of methodological development (HHB08a; HHH12; KHM21). An alternative is maximum pseudo-likelihood estimation (MPLE), which replaces the true likelihood with a tractable surrogate formed from the conditional distributions of individual ties (DHG09). MPLE is fast and often accurate when the dependence among ties is weak, but it can be biased when the model exhibits realistic levels of transitivity and clustering. More recently, Mele and Zhu (MZ23) proposed a variational mean-field approximation to the ERGM likelihood. Their method replaces the full ERGM's complex tie dependence by a product of independent Bernoulli distributions over dyads, and optimizes the resulting lower bound on the log-normalizing constant. This approach is deterministic and avoids MCMC entirely, which is appealing. The authors reported that the mean-field variational likelihood estimator (MFVLE) performed comparably to MCMC-MLE in simulation experiments, and they proved that the approximation error converges to zero as the number of nodes grows. However, the theoretical convergence result requires assumptions about how the model parameters behave as the network grows, and these assumptions are difficult to verify for realistic networks. The simulation experiments, as we show in Chapter 2, contain an error that invalidates them.

A second difficulty with ERGMs is model degeneracy. Large regions of the ERGM parameter space correspond to models that concentrate nearly all their probability mass on a small number of extreme networks, typically the empty graph or the complete graph (Han03; Sch11). A model in a degenerate region does not produce networks that resemble anything observed in practice, and the problem has consequences for both estimation and inference. Optimization algorithms can wander into degenerate regions during fitting, and prior distributions specified in the natural parameter space can place substantial density on

degenerate parameter values without the modeler intending or realizing it. Standard prior choices for Bayesian ERGMs, most commonly multivariate normal or uniform distributions over the natural parameters, do not account for this geometry. A normal prior centered at zero with variance 100, which is the default in widely used software, places nearly half its density on parameter values that generate near-complete or near-empty graphs. This is almost never the belief that a researcher means to express. The problem is that the mapping from natural parameters to the space of sufficient statistics is highly nonlinear, and what looks like a diffuse, uninformative prior in θ-space can be strongly informative in the space of network statistics. Notably, it can be informative in exactly the wrong direction, concentrating density on degenerate configurations. Despite the recognized importance of the degeneracy problem, there are few principled prior specifications that directly address it. Existing work on improved ERGM specifications, such as geometrically weighted statistics (SPR06) and tapered ERGMs (FH17; BH22), addresses degeneracy through the model's sufficient statistics rather than through the prior. These tools prove useful in this dissertation, but they do not solve the prior specification problem.

Even with a well-specified prior, Bayesian inference for ERGMs remains doubly intractable. The posterior depends on the same normalizing constant that makes the likelihood intractable, and standard Metropolis-Hastings updates require ratios of normalizing constants at proposed and current parameter values. A variety of methods have been developed to handle this difficulty, including auxiliary variable methods (MPR06; MGM06; Lia10; LJS16), noisy exchange algorithms (LJ13; AFE16), and noisy Hamiltonian Monte Carlo (SBF19). All of these methods require simulating auxiliary networks from the ERGM at proposed parameter values, and the quality of the posterior approximation depends on the number and quality of these auxiliary simulations. Noisy Hamiltonian Monte Carlo (HMC), which we use in Chapter 3, is one of the recent advancements in these methods. It exploits gradient information through a leapfrog trajectory and uses a telescoping estimator of the normalizing constant ratio that is more stable than single-point estimators for long

trajectories. But each iteration of noisy HMC requires auxiliary network simulations at every leapfrog step, and trajectories of 40 to 50 steps are typical for networks of even moderate size. This makes the method expensive and potentially limits its applicability to larger networks or more complex models. Stochastic gradient Langevin dynamics (SGLD), adapted for ERGMs by Zhang and Liang (ZL24), offers a cheaper alternative. SGLD takes a single gradient step with injected noise at each iteration, with no accept-reject step and no normalizing constant ratio to estimate. The per-iteration cost is a single batch of auxiliary network simulations rather than a full trajectory's worth. Zhang and Liang show that SGLD converges to the true posterior in 2-Wasserstein distance, a stronger guarantee than the approximate stationarity targeted by noisy HMC. Because the non-degeneracy prior is evaluated from the same auxiliary network draws that these samplers already generate, it integrates into auxiliary-sample-based methods with minimal additional computation. Chapter 4 demonstrates this by pairing the prior with SGLD and validating the resulting posterior against exact ground truth.

This dissertation makes three contributions to the methodology of ERGM inference, organized as three chapters following this introduction. Chapter 2 assesses the variational mean-field likelihood approximation proposed by Mele and Zhu (MZ23). We identify an error in their simulation study in which both the MCMC-MLE and MFVLE algorithms were initialized at the true parameter values rather than at data-driven starting points. When initialized at the MPLE, which is standard practice, the MFVLE performs poorly and is dramatically outperformed by MCMC-MLE. We further show that the parameter values used in their study produce models that are close to tie-independent, making the estimation problem unrealistically easy. When we increase the transitivity of the model to more realistic levels using the mean-value parameterization framework of van Duijn et al. (DHG09) and tapered ERGMs (BH22), the performance gap widens. The MFVLE is also computationally more expensive than alternatives at every network size tested. We provide conceptual reasoning for why the mean-field approximation will be poor in exactly

the circumstances where ERGM estimation is most needed, that is, when there is strong tie dependence and realistic transitivity. The mean-field family is too restrictive to serve as a default likelihood approximation for models with meaningful network structure.

Chapter 3 proposes the non-degeneracy prior for Bayesian ERGMs. The prior assigns density to each parameter value in proportion to the probability that a network generated at that parameter has sufficient statistics in the interior of the convex hull of achievable statistics, rather than on its boundary. A strength hyperparameter p controls how strongly the prior excludes degenerate regions, with $p = 1$ corresponding to the information of a single observed non-degenerate network. We compute the prior exactly for all networks on seven nodes, where the normalizing constant and the convex hull can be enumerated, and we show that the commonly used normal prior places 43% of its density on near-full graph regions and 29% on near-empty graph regions. The non-degeneracy prior removes this concentration. For larger networks, we develop an importance-reweighting approximation that reuses auxiliary network draws already generated by the posterior sampler, adding negligible computational cost. We validate the approximation against exact ground truth on the seven-node network, and apply it to the Kapferer tailor shop network with 39 nodes. On a model with degeneracy-prone statistics, the non-degeneracy prior at $p = 1$ substantially tightens the posterior predictive distribution compared to normal and uniform priors, with the improvement concentrated on the full-graph side where the prior practically eliminates the tendency to generate near-complete networks. On a model with geometrically weighted statistics that already control degeneracy, the prior has no measurable effect, confirming that it does not interfere when it is not needed.

Chapter 4 replaces noisy HMC with stochastic gradient Langevin dynamics as the posterior sampler for the non-degeneracy prior. The SGLD update requires only a single batch of auxiliary network simulations per iteration rather than a full leapfrog trajectory, reducing the per-iteration cost by a factor equal to the trajectory length. The non-degeneracy prior gradient integrates into the SGLD update at zero additional simulation cost, because it is

computed from the same auxiliary draws used for the likelihood gradient. We derive the prior gradient in closed form and develop a hybrid resampling strategy that spreads the cost of auxiliary network simulation by reusing each batch across multiple Langevin steps with importance-sampling reweighting.

The three chapters address different aspects of ERGM inference but are connected by a recurring observation. Models for networks are useful precisely because they can represent complex dependence among ties. That same dependence makes inference hard. Approximations that simplify the dependence structure, whether by replacing the likelihood with a mean-field representation or by placing priors that ignore the geometry of the parameter space, tend to fail where the dependence is strong and the model is most needed. The methods developed in this dissertation are motivated by this tension, and by the practical goal of making ERGM inference reliable for the kinds of networks that researchers actually observe.

CHAPTER 2

Assessing Variational Likelihood Estimation in Network Formation Models

2.1 Introduction

There has been increasing interest in models for complex stochastic phenomena. Exponential-family models are able to represent complex structure in an interpretable way (Bar78; HHB08b). This is particularly true of stochastic models for networks, where the nodes represent actors and the edges relational ties between those actors. The configuration of these ties and their dependence on attributes of the nodes influence how processes unfold on networks. Exponential-family random graph models (ERGMs) offer a flexible framework for representing these dependencies (DI10; KH14; BM17). There is a substantial literature on these network models and their applications across disciplines (BM17; De 17; Pol19).

2.1.1 Exponential-family Random Graph Models

The ERGM framework for network data is reviewed in the Introduction (Chapter 1). We recall here the elements specific to this chapter. An observed network on n nodes is a realization $y \in \mathcal{Y}_n$ of random network Y, with probability

$$p_\theta(Y = y) = \frac{\exp\{\theta^\top g(y)\}}{Z(\theta)} \quad y \in \mathcal{Y}_n, \tag{2.1}$$

where $g(y) \in \mathbb{R}^d$ is a vector of sufficient statistics, $\theta \in \mathbb{R}^d$ is the natural parameter, and $Z(\theta)$ is the normalizing constant. The heterogeneous nodes, indexed by $i = 1, \ldots, n$, often

have an associated vector x_i of covariates. For notational convenience, dependence on the covariates is suppressed.

The tie variables $\{Y_{i,j}\}$ are jointly distributed and in general not independent. Capturing this dependence is why ERGMs are used in place of simpler dyadic models. When $g(y)$ is a sum of single-tie statistics, the joint distribution reduces to a product of independent Bernoulli ties, and in this special case the MLE coincides with the MPLE.

The sufficient statistics used in this chapter are motivated by economic and social theory about the generating process (FS86; GKM09) and measure the network propensity to form edges ('Edge', $\sum_{i,j}^{n,n} y_{i,j}$), clustering of edges ('Two-Star', $\sum_{i,j,k}^{n,n} y_{i,j}y_{i,k}$), transitivity ('Triangle', $\sum_{i,j,k}^{n,n} y_{i,j}y_{i,k}y_{j,k}$), and homophily on a nodal attribute x_i ('Homophily', $\sum_{i,j=1:y_{i,j}=1}^{n,n} \mathbb{1}\{x_i = x_j\}$). There are many other choices for these statistics (MHH08).

2.1.2 Likelihood-based inference for ERGM

The normalizing constant $Z(\theta) = \sum_{y\in\mathcal{Y}_n} \exp\{\theta^\top g(y)\}$ is computationally intractable for all but the smallest networks. Two established approaches to likelihood-based inference, reviewed in the Introduction (Chapter 1), are MCMC-based maximum likelihood estimation (MCMC-MLE) (GT92; HHB03; HHB08a; HHH12; KHM21) and maximum pseudo-likelihood estimation (MPLE) (DHG09). MCMC-MLE has become the standard method with much research and practical knowledge behind it, while MPLE offers computational simplicity at the cost of potential bias when tie dependence is strong.

Over time, substantial efforts have been made to improve likelihood-based inference for ERGM (HHB08a; HHH12; KHM21). In that effort, Mele and Zhu (MZ23) proposed a method to approximate ERGM likelihood using a variational mean-field and demonstrated its effectiveness through microeconomics application. Their method minimizes the Kullback–Leibler (KL) divergence between a tractable approximating distribution and the true ERGM likelihood by maximizing a lower bound on the intractable normalizing constant. Unlike MCMC

approximation of the intractable likelihood, maximum mean-field variational likelihood estimates (MFVLE) are deterministic.

A key feature of the mean-field approximation is that it replaces the full ERGM, with its complex dependence among ties, by a model with conditionally independent ties whose parameters are chosen to be as close as possible to the ERGM under the KL criterion. This factorization results in a variational problem whose solution provides an approximate log-likelihood without requiring any simulation. Mele and Zhu (MZ23) show that the approximation error of this mean-field log-constant converges to zero as the number of nodes increases, using tools from nonlinear large deviations theory. Consequently, the mean-field log-likelihood uniformly approximates the true ERGM log-likelihood in large networks, and the MFVLE converges to the MLE when the underlying ERGM likelihood is well behaved. They also provide explicit bounds on the approximation error, and their studies suggest that, even for moderately sized networks, the approximation performs better than these bounds indicate.

These results suggest that the variational mean-field approach provides a computationally efficient and theoretically grounded alternative to standard simulation-based ERGM estimation.

2.1.3 Outline of the paper

This paper focuses on the mean-field ERGM likelihood approximation proposed by Mele and Zhu (MZ23) and examines its performance within a comparison framework. While variational approximations have been successful and useful in many fields, their properties are not assured. We show that the estimation experiments on finite networks in Mele and Zhu (MZ23) contain an error that invalidates them. We correct this error and replicate the study on $1,000$ sampled networks, then show that the empirical performance of the method is poor and dramatically worse than the MCMC-MLE. In particular, our analysis adjusts the initialized points for the approximation methods, while focusing on how well the mean-field

likelihood approximation functions under certain levels of network transitivity. A useful tool here comes from van Duijn et al. (DHG09), where transitivity in ERGM natural parameters can be systematically adjusted by way of one-to-one mapped mean-value parameterization. In turn, manually increased transitivity allows us to stress test the estimation methods under more complex network structures. Our findings are accompanied by conceptual reasoning on why the mean-field method will be poor in circumstances where it is most needed.

The paper is organized as follows. Section 2.2 reviews ERGM specifications, likelihood-based inference methods, and other conceptual issues related to the mean-field likelihood approximation for ERGM. Section 2.3 discusses the simulation study framework and compares the mean-field variational method against current methods. Section 2.4 draws conclusions on the contents of this paper.

2.2 Theoretical considerations

2.2.1 Specification of the ERGM class

Mele and Zhu (MZ23) used a microeconomics model to explore best-response equilibrium dynamics in potential games by measuring players' payoffs. We use the same model to simulate networks here. Real-world networks, such as those in social science and biology, tend to have tie dependency that causes clustering of the edges and transitivity. The microeconomics model reflects this by including the formation of Two-stars and Triangles as its terms.

One aspect of model specification that is important to take into consideration is that certain combinations of terms result in the model not placing sufficient probability mass on realistic graphs and networks. This is called the model degeneracy problem (Han03; Sch11; SKB20). For example, using a Triangle term as a measure of transitivity will lead to degeneracy in most cases. This is because a single edge can form the basis to a large number of triangles so using the unweighted count of the number of triangles is problematic: an ERGM with a positive coefficient on a Triangle term will usually exhibit degeneracy.

An alternative measure of the tendency for an edge to close triangles is the geometrically-weighted shared partnership (GWESP) term (SPR06). GWESP is a measure of transitivity of the network y. It is motivated by the need to have weaker conditional independence restrictions than earlier models (FS86). Specifically, near-degeneracy is the related condition where the model exhibits unrealistic probabilistic behavior by placing significant mass on the boundary of the sample space of the network statistics. This typically means the model class will have severe lack-of-fit to realistic network data. For the models here we may expect to see unrealistic numbers of two-stars and/or triangles. In the similar sense that the authors incorporated the suggestion from Chatterjee and Diaconis (CD13) of scaling the transitivity term to not explode with a change in network size, the choice to redefine the microeconomics model using GWESP instead of the current Triangle term is likely to provide benefits to ERGM degeneracy as well. Both the rescaling and the choice of GWESP focus on a sub-class of ERGMs. This suggestion of GWESP serves as an aside and this paper will focus on the model studied by Mele and Zhu.

2.2.2 Comparison of methods of estimation

In this section we compare the various forms of estimation: the MLE, MCMC-MLE, MPLE and the MFVLE.

It is well known that MCMC-MLE converges to the MLE as the sampling increases (GT92) and can provide useful results with enough iterations. However, convergence is often not practically available due to near-degeneracy and MCMC variability (Han03; Sch11). This paper compares the methods for likelihood inference with the degeneracy issue in mind. Maximum pseudo-likelihood estimation is a popular choice, and is used in many applications for its advantage in computational time − compared to using MCMC − to produce approximate results. A comparison between the MLE, MCMC-MLE and MPLE was made by van Duijn et al. (DHG09) and we extend their framework here to incorporate the MFVLE.

2.2.3 Theoretical analysis of MFVLE

Mele and Zhu (MZ23) develop their variational approximation in their Section 2C. The core idea is to express the likelihood as a supremum over the set of all probability mass functions (PMFs) on $\mathcal{Y}_n$. As this is not tractable, the likelihood is approximated by the supremum over the set of all tie-independent PMFs on $\mathcal{Y}_n$, which does have a closed form. Variational approximations gain their power from the closeness of their shape to that of the actual log-likelihood. However, in any given setting it is hard to assess how close this approximation is because $Z(\theta)$ is both computationally and theoretically intractable in regions of the parameter space of interest to modelers. Theoretical results in the closely related Ising model show that it can be quite accurate when the Ising graph is complete (EN78) and very poor when it is sparse (DM10). Ricci-Tersenghi (Ric12) show that the MFVLE can perform poorly, especially if the dependence is strong. Lokhov et al. (LVM18) also show poor performance for strongly coupled systems and near degenerate regions of the parameter space. They contrast this with MPLE and other more robust methods. This difference in performance depending on which region of the parameter space the true model is an intrinsic problem for variational approximations. For example, Biondini et al. (BMP22) consider a mean-field approximation of an ERGM based on $k-$star statistics. They find that, under a particular asymptotic scheme as n increases, the mean-field approximation represents macroscopic properties (e.g., edge density) well, while not representing microscopic properties (e.g., dyadic dependence) well.

If the data generating model or the specified model deviates far from a tie-independent model, then the mean-field approximation is likely to be very poor as the two suprema can be very different. In particular, if there is strong tie-dependence ("coupling") we would expect the approximation to be poor and the resulting estimates unreliable.

One way to understand this behavior more formally is via convex duality. The ERGM log-likelihood is a convex function of the natural parameter, with curvature determined

by the covariance structure of the sufficient statistics $g(Y)$. The mean-field approximation replaces the joint ERGM by a product of conditionally independent Bernoulli dyads, whose covariance structure is diagonal. The associated variational objective therefore replaces the true log-partition function by a much simpler function that cannot reproduce the strong curvature induced by higher-order dependence such as triangles and k-stars. In regions of the parameter space where the feasible set of sufficient statistics is sharply curved or nearly non-convex the KL projection onto the mean-field family can lie far from the true model, leading to biased point estimates and substantially underestimated uncertainty. These are precisely the regions where degeneracy and strong transitivity occur, and the reason we focus on these characteristics in our analysis.

Note that even if the data generating model is tie-independent then the specified model class needs to be the data generating model for the method to be exact (that is, the variational approximation equals the data generating likelihood). However, even if the data generating process is tie-independent the MFVLE may not be accurate if the specified model used to estimate the parameters is tie-dependent. For example, if the model specified in the mean-field approximation includes both a Two-star and Triangle term then the MFVLE can be poor even if the data generating process is tie-independent. In general, we expect the model to be misspecified so that the misspecification error will be present, adding to the overall approximation error.

In addition, we argue in Section 2.3.3 that most realistic models have strong tie-dependence and so the range of applicability of the method may cover mostly unrealistic and therefore not very interesting models.

A common problem with estimates based on variational approximations is their under-estimation of the estimation uncertainty. For example, Tan and Friel (TF20) show that in Bayesian inference for ERGM, variational approximations underestimate the posterior variances.

The main result of Mele and Zhu, Proposition 1, shows that as the number of nodes, n, in

the network increases the approximation converges to the ERGM log-likelihood. This result could be interpreted to imply that the MFVLE will be close to the MLE for large network sizes. However, this result presumes the parameter values $\theta_n = (\alpha_n, \beta_n, \gamma_n)$ do not increase too fast as n increases. From a statistical perspective it is useful to distinguish two related but distinct questions. First, for a fixed parameter value θ, how well does the mean-field approximation of the log-normalizing constant approximate the true ERGM log-normalizing constant? Second, when the approximation is used for estimation, how close is the MFVLE to the MLE for finite n, especially under inevitable model misspecification? Proposition 1 of Mele and Zhu addresses the first question in a particular asymptotic regime; the simulation study in this paper is concerned with the second question for realistic finite networks and parameter values.

The asymptotic regime considered by Mele and Zhu is also close in spirit to the dense-graph (graphon) asymptotics that arise when all ERGM sufficient statistics scale as n^2 and the underlying graph becomes increasingly dense (CD13; BCL11). In contrast, most empirical social and economic networks are sparse or semi-sparse, with edge counts growing sub-quadratically in n, and with parameters that typically change with n. In such regimes the qualitative behavior of the ERGM can differ substantially from the dense limit, and there is no guarantee that a mean-field approximation calibrated to the dense regime will accurately reflect the likelihood surface in the parameter regions that are substantively most relevant. Hence, while on the surface the assumptions of the theorem seem reasonable, we have no basis to make them: The number of nodes is a fundamental social / economic characteristic of the phenomena. That is, it is a structural characteristic not under the researcher's control. Hence different n correspond to different generating processes and − almost surely − disparate θ_n. For example, consider social relations in a school classroom with $n = 3$ students in it compared to a classroom with $n = 100$ students. The social dynamics are quite different and the θ_n should be as well. For a broader discussion of this issue, see Krivitsky et al. (KHM11). In sum, for Proposition 1 to imply that the MFVLE

will be close to the MLE for large network sizes, the fundamental generating processes must behave in a certain convenient way. As this is not assessed in the paper theoretically, we are motivated to study it via simulation for specific network sizes and parameter values.

Consider also the potential game used as motivation in their Section 2B. The final term in this specification (their equation (5)) counts the number of common friends. This is commonly referred to as a Triangle term for ERGM and is well known to lead to degenerate equilibria for wide ranges of $\gamma > 0$ (Han03). As such the specification is unlikely to be economically realistic or relevant.

Finally, in the discussion of the closeness of the mean-field approximation of the log-likelihood to the ERGM log-likelihood they say the approximation is similar in spirit to the MCMC-MLE method, where the log-normalizing constant is approximated via MCMC. However, in MCMC n corresponds to the size of the MCMC sample, which is completely under the control of the researcher and can be made arbitrarily large. In their situation, it is an intrinsic characteristic of the problem and fixed (in any specific application) and so is fundamentally different in spirit and impact.

2.2.4 A framework for the comparison of estimation methods

Motivated by the analysis in the previous sub-section, we design a simulation study to assess the accuracy of the MFVLE compared to the MCMC-MLE and MPLE. We will adopt the framework of van Duijn et al. (DHG09) which was developed for this purpose. Handcock (Han03) proposed the consideration of the alternative mean-value parameterization ERGM, $\mu(\theta)$, that has the one-to-one mapping with natural parameters θ characterized by:

$$\mu(\theta) = \mathbb{E}_\theta[g(Y)]. \tag{2.2}$$

The mapping is strictly increasing in the sense that:

$$(\theta_a - \theta_b)^T (\mu(\theta_a) - \mu(\theta_b)) \geq 0 \tag{2.3}$$

with equality only if $P_{\theta_a}(Y = y) = P_{\theta_b}(Y = y)\ \forall y$. The mapping is also injective in the sense that

$$\mu(\theta_a) = \mu(\theta_b) \rightarrow P_{\theta_a}(Y = y) = P_{\theta_b}(Y = y)\ \forall y. \tag{2.4}$$

The mean value parameterization has the main advantage of interpretability due to its direct translation to the network statistics chosen for the model. van Duijn et al. (DHG09) adopted the mean-value parameterization to examine performance of MCMC-MLE and MPLE under different transitivity scenarios. Their results show the MPLE performs as well as the MCMC-MLE when there is little transitivity/tie-dependence in the model. However, where there is realistic transitivity/tie-dependence the MCMC-MLE outperformed MPLE. It is worth noting that the performance gains of MCMC-MLE over MPLE are most pronounced for parameters capturing endogenous tie dependence, whereas the improvement for nodal and dyadic covariates such as homophily is comparatively modest.

An increase in transitivity cannot be easily induced in the natural parameterization, as a direct change in a natural parameter leads to complex changes in the distributions of the network statistics due to interactions. However, transitivity can be added by increasing a corresponding mean-value parameter value. The mean-value parameters with the added transitivity can then be inversely mapped to the vector of natural parameters (See van Duijn et al. (DHG09) for details).

As model transitivity is increased by way of the mean-value parameterization, the superior performance of MCMC-MLE became more noticeable as the bias of the MPLE for both the natural parameters and mean-value parameters increases. Despite this inferior performance in realistic settings, the MPLE serves as a useful initialization for more complex methods, such as the MCMC-MLE. In fact, it is the default initialization for the state-of-the-art

MCMC-MLE in the `ergm` R package (HHB03; RDevelopmentCoreTeam12). One factor to keep in mind is the likelihood of near-degeneracy for the estimates (that is, high proportion of mass placed on the boundary of the convex hull of network statistics). The degeneracy of a model fit can be assessed by the **R** package `ergm` (HHB03; RDevelopmentCoreTeam12).

We will use the framework for comparison introduced by van Duijn et al. (DHG09) to compare the mean-field variational likelihood estimate (MFVLE) of Mele and Zhu (MZ23) to the MCMC-MLE.

2.3 Study design and simulation results

We first replicate the Monte Carlo simulation study of Mele and Zhu (MZ23), Section 4.2. A key feature of their study was initiating the optimization at the true values. They find that the estimates do not vary much from the start point, and interpret this as accuracy. We replicate their study but starting from the MPLE, which is by far the most common initialization point for estimates (HHB08a). We try larger network sizes to see if that improves the MFVLE, and then show that the parameter values they use are close to a tie-independent model. Such parameter values are rarely realistic and so we modify the parameter values to produce a similar model with more realistic transitivity (i.e., an increased level). We then repeat the study for models with this more realistic transitivity.

2.3.1 Correcting the Monte Carlo simulation study of Mele and Zhu (MZ23)

The identical microeconomics model set-up from Mele and Zhu (MZ23) − in which the chosen sufficient statistics are edges, homophily, two-stars, and triangles − is used. The predetermined true parameters corresponding to the models are dependent on network size. To generate artificial networks in practice, a Metropolis–Hastings network sampler is used with the initialization being an iteratively simulated network using the true parameters with attributes drawn from a Bernoulli distribution. Here, $1,000$ networks are simulated for

model fitting while $10,000$ sets of summary statistics are simulated for mapping of mean-value parameterization. This was done using the ergm package to be consistent with Mele and Zhu (MZ23) (although it was mentioned and used, the package was not cited by them).

We note that there is an error in their simulation study. As stated at the end of their Section 4B, "To be sure that our results do not depend on the initialization of the parameters, we start each estimator at the true parameter value." Mele and Zhu (MZ23) set the initialization of model parameters in the algorithm for both MCMC-MLE and MFVLE to be the true parameters that are used to simulate networks. However, these true parameters are not attainable in practice. More so, their use of the true values obscured the fact that the algorithm did not move the parameter values toward the true values, but only started there. This renders the algorithm applied for MFVLE inappropriate and its results at best questionable. Here, we use the MPLE, the current standard method, as the initialized points to validate our claim. To ensure that the mean-field approximation does not stagnate at a non-optimal solution as is by nature of variational approximations, MPLE initial points used were attributed random noise and thereby effectively restarting the variational approximation procedure over the large number of simulated networks. We have tried other initial estimates, including zero vectors and random Gaussian starting points, with similar results for both the MCMC-MLE and MFVLE.

2.3.2 Original study with adjusted initialization

With the initialization adjustment applied, Table 2.1 partially replicates the motivating comparison for $\theta = \{-4, 2, \frac{1}{n}, \frac{1}{n}\}$ for $n = \{50, 100\}$ (See Mele and Zhu (MZ23) for details). Note the parameters displayed by the authors $\{\tilde{\alpha_1}, \tilde{\alpha_2}, \beta, \gamma\} = \{-2, 1, 1, 1\}$ are scaled by $\{2, 2, \frac{1}{n}, \frac{1}{n}\}$, as the authors make explicit in their provided MFVLE code implementation. We produce the MCMC-MLE using R package `ergm` directly from the log-likelihood function rather than the stochastic approximation (Sni02) chosen by the authors. A burn-in of 100,000 iterations and interval of 1,000 iterations are used to be consistent with the original

n = 50	**Edge**	**Homophily**	**Two-Star**	**Triangle**	**Degeneracy**
True Natural parameters	-4.000	2.000	0.020	0.020	-
MCMC-MLE Natural	-3.851	2.050	-0.006	-0.042	4.6 %
MFVLE Natural	-36.532	-14.502	-0.980	-2.202	9.0 %
MPLE Natural	-3.868	2.044	-0.005	-0.020	0.0 %
True Mean-Value parameters	95.80	82.79	363.76	13.961	-
MCMC-MLE Mean-Value	95.99	82.93	365.08	14.02	-
MFVLE Mean-Value	263.05	163.04	5645.46	1189.43	-
MPLE Mean-Value	139.89	104.40	2561.62	743.19	-
n = 100	**Edge**	**Homophily**	**Two-Star**	**Triangle**	**Degeneracy**
True Natural parameters	-4.000	2.000	0.010	0.010	-
MCMC-MLE Natural	-3.916	2.005	0.004	0.005	0.4 %
MFVLE Natural	-7.386	-24.215	-0.253	-0.212	2.5 %
MPLE Natural	-3.915	2.003	0.004	0.008	0.0 %
True mean-value parameters	393.12	340.94	3091.85	117.74	-
MCMC-MLE Mean-Value	393.07	340.86	3091.37	117.76	-
MFVLE Mean-Value	910.53	621.64	33030.41	6623.04	-
MPLE Mean-Value	416.69	352.13	5547.74	936.87	-

Table 2.1: Natural and mean-value parameter estimates for the original model with $\theta = (-4, 2, 1/n, 1/n)$, comparing MCMC-MLE, MFVLE, and MPLE across 1,000 simulated networks at $n \in \{50, 100\}$. Both MCMC-MLE and MFVLE are initialized at the MPLE. The degeneracy column reports the percentage of fits producing degenerate estimates.

comparison. The MFVLE procedure is directly implemented using the R package `mfergm` made available on GitHub by the authors (Mel19). To ensure proper mapping from natural parameters to mean-value parameters, 10,000 networks are simulated to take the expected counts of sufficient statistics.

Table 2.1 displays the following values: true natural parameters, MCMC-MLE natural parameter estimates, MFVLE natural parameter estimates, MPLE natural parameters used

to initialize both methods, and the one-to-one mean-value parameter estimates mapped from the natural parameters for truth, MCMC-MLE, MFVLE, and MPLE. For these results, the mean is used for both natural and mean-value parameterization. The right-side column shows the percentage of degeneracy fits for each of the three methods out of 1,000 simulated networks to be fitted.

The mean-value parameter estimates in Table 2.1 with $n = 100$ display the vector of true mean-value parameters corresponding to the natural parameter. We see that expected number of triangles is 117.74. The MCMC-MLE estimate is 117.76, while the MFVLE estimate of 6623.04 is severely off the mark. There are similar results for the Two-stars parameter with a true value of 3091.85, MCMC-MLE of 3091.37 and MFVLE of 33030.41. Note that the true MLE is exactly equal to the true values, so that the error in the MCMC-MLE is a computational error. This is partially induced by the near-degeneracy of the model (Han03). While MPLE resembles MCMC-MLE in natural parameterization, the difference between the two methods is stark in the expected count of transitivity terms. The bias in MPLE mean-value parameter estimates is induced by a set of samples with extreme values, similar to the results of van Duijn et al. (DHG09).

Table 2.2 provides root mean squared error (RMSE) and mean absolute deviation (MAD) for the results in Table 2.1. The variability of the MFVLE results stands out greatly compared to the standard variability of the MCMC-MLE in both natural and mean-value parameterization. This can be seen more clearly in the middle row of Fig 2.1, where some extreme values in the MFVLE natural parameters have been replaced, in contrast to the top row, which shows the full set of MCMC-MLE natural parameter samples. Using winsorization based on the 15th and 85th percentile cutoffs, 294 of the 975 total edge counts (accounting for degeneracy) were replaced when producing the histograms shown in the middle row of Fig 2.1. While the center MFVLE distributions highlight the regions of highest mass, the majority of samples are spread broadly across the parameter space (e.g., the Edge natural parameter ranges from approximately -1230 to 652, which is unrealistic). The large variability

in MFVLE results was not observed by Mele and Zhu (MZ23); in fact, the authors reported lower mean absolute deviation in the results of MFVLE compared to that of MCMC-MLE. Although we were able to deterministically reproduce the authors' results when initializing at the true parameter values, these findings substantiate our concerns. That is, MFVLE appears to have performed well in Mele and Zhu (MZ23) primarily because the true parameter values were used as the initial values, which is the only modification our study has made relative to the original study.

n = 100		**Edge**	**Homophily**	**Two-Star**	**Triangle**
MCMC-MLE Natural	RMSE	0.335	0.157	0.020	0.101
	MAD	0.273	0.130	0.016	0.082
MFVLE Natural	RMSE	3.904	6.986	0.060	0.137
	MAD	2.655	4.436	0.045	0.108
MPLE Natural	RMSE	0.338	0.156	0.021	0.103
	MAD	0.278	0.130	0.017	0.084
MCMC-MLE Mean-Value	RMSE	26.448	24.186	420.953	28.130
	MAD	21.790	19.902	346.273	23.222
MFVLE Mean-Value	RMSE	584.253	411.468	17061.371	1131.270
	MAD	403.701	308.844	10830.627	685.123
MPLE Mean-Value	RMSE	27.361	24.897	435.168	28.930
	MAD	22.483	20.496	358.306	23.874

Table 2.2: RMSE and MAD of natural and mean-value parameter estimates across 1,000 simulated networks for $n = 100$, corresponding to the results in Table 2.1.

MFVLE mean-value parameter estimates' lack of resemblance to the true parameters can be somewhat alleviated by replacing extreme values, as seen in Fig 2.1. To look at the medians of the mean-value parameters instead of the means, the MFVLE parameter estimates for $n = 100$ become $\{394, 341, 3076, 116\}$. These values lie much closer to the true parameters $\{393.12, 340.94, 3091.85, 117.74\}$ than the mean-value estimates $\{910.53, 621.64, 33030.41, 6623.04\}$.

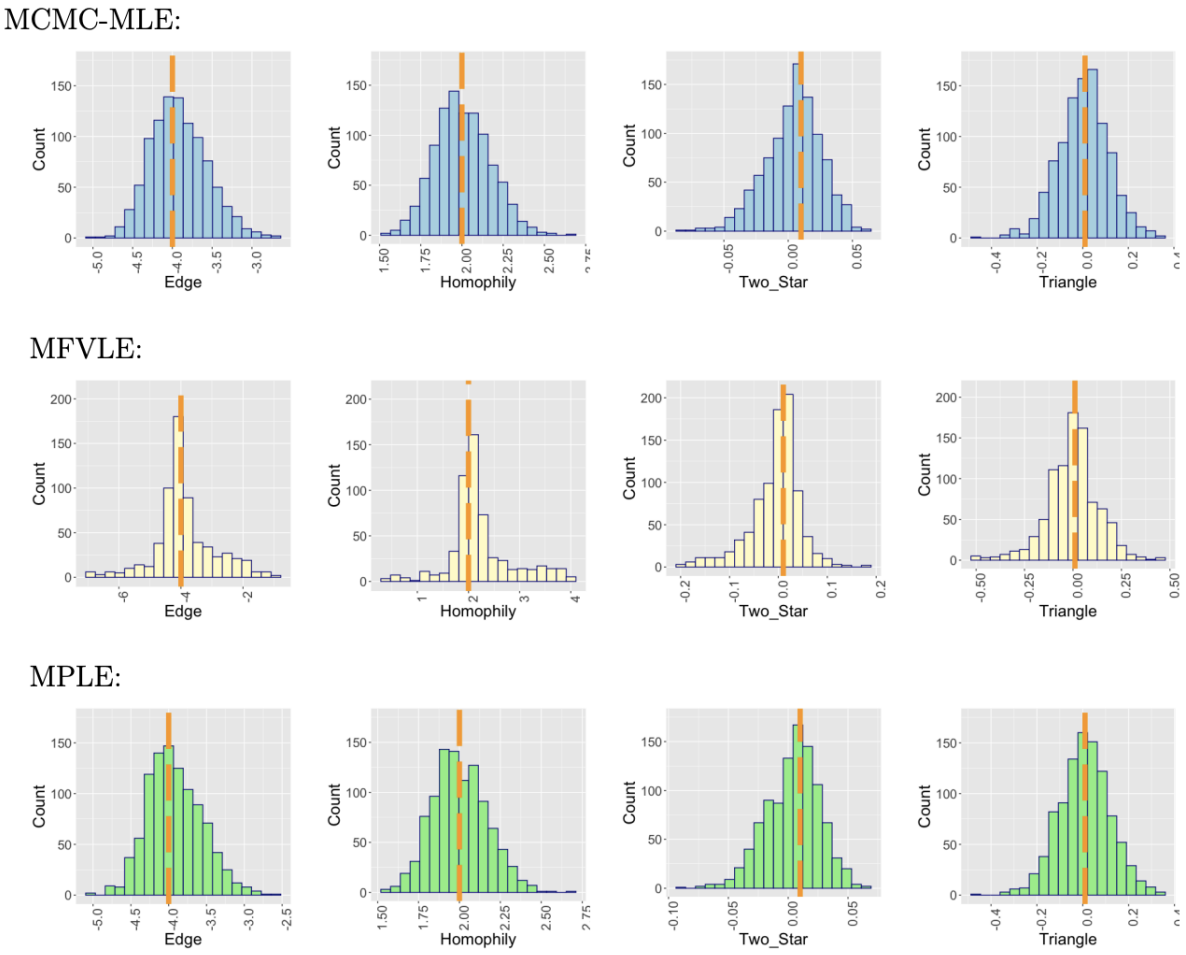


Figure 2.1: Distribution of natural parameter estimates across 1,000 simulated networks for $n = 100$. Top row: MCMC-MLE. Middle row: MFVLE, winsorized at the 15th and 85th percentiles and displayed on a different horizontal scale. Bottom row: MPLE. Orange lines indicate true parameter values. Of the 975 non-degenerate MFVLE fits, 294 edge parameter estimates were replaced by the winsorization cutoffs.

Similarly, the medians of the MFVLE natural parameter estimates $\{3.987, 2.132, 0.000, -0.007\}$ are much closer to true values than the means. Alternative to the full sample errors in Table 2.2, the winsorized samples of Fig 1 can be used to allow slight improvements in MFVLE, leading to its natural parameters having RMSE of $\{2.372, 3.230, 0.051, 0.104\}$ and MAD of $\{1.759, 2.311, 0.038, 0.088\}$. While ignoring extreme values makes MFVLE appear more fa-

vorable, we note again the large spread of MFVLE samples and its nontrivial tendency to produce outliers (roughly 10% of samples using the 1.5× IQR rule of thumb). Moreover, trimming and winsorizing to produce more adequate settings for comparison only result in marginal improvements to RMSE and MAD for MFVLE at the cost of altering a large set of parameter estimates. Even when accounting for MFVLE's pronounced distributional shortcomings, MCMC-MLE still clearly outperforms it in both the natural and mean-value parameterizations.

We now consider if larger network sizes improve the performance of the estimators. For very large network sizes the standard MPLE can be computationally burdensome, so we use a variant of it where selective sampling of the dyads is undertaken. Specifically, ties are selected with probability 1 and non-ties are selected with low probability. Then the pseudo-likelihood is approximated by the inverse-probability estimator based on the sample. This is often called "case-control" or "endogenous" sampling and has been available in **ergm** for two decades. It dramatically speeds up the computation with minimal and controllable loss in precision.

In Table 2.3 we show results for a large network size ($n = 2,000$). This extends the pattern we see for smaller networks: The MCMC-MLE, MPLE and sampled MPLE are very accurate, while the MFVLE is completely unrealistic due to the high proportion of extreme values and high variance. We also note the relative computational costs of these methods. We see that the MPLE and sampled MPLE are quite fast, while the MCMC-MLE is very practical. The MFVLE is extremely slow, in addition to being inaccurate. Finally, we repeated the study using a very large network ($n = 10,000$). For networks of this size the MFVLE algorithm did not finish after two weeks running. It is not computable for large networks. The MPLE took about 560 minutes per estimate and the sample MPLE took about 127 minutes per estimate. Both were very accurate. The sampled MPLE could be made much faster by reducing the number of sampled dyads (here set to very large value, 5 million).

n = 2000	**Edge**	**Homophily**	**Two-Star**	**Triangle**	**CPU time**
True Natural parameters	-4.000	2.000	1.000	1.000	
MPLE Sampled Natural (200 simulations)	-4.001	2.000	1.016	0.784	0.32 min
MPLE Natural (200 simulations)	-4.005	2.000	1.043	0.859	1.2 min
MCMC-MLE Natural (10 simulations)	-4.032	2.004	1.218	0.198	300 min
MFVLE Natural (10 simulations)	-3.462	-15.818	-16.4	-25.5	2775 min

Table 2.3: Natural parameter estimates for n = 2,000 with MPLE initialization and true parameters $\theta = (-4, 2, 1/n, 1/n)$. Two-Star and Triangle values are scaled by n. CPU times are per estimate on an Intel Xeon Platinum 8160 with 256GB RAM.

In summary, these results show that the improved performance of the MFVLE shown in their Table 1 are not correct and are, in fact, due to starting the MFVLE algorithm at the true values. The algorithm does not move from its starting value and hence appears to be accurate. When not started at the truth, MFVLE results appear to be worse than those of MPLE and MCMC-MLE.

From a computational point of view, these experiments also highlight very different scaling behaviors of the various estimators. Each MFVLE fit requires optimization over all dyads at every variational iteration, so its cost grows at least on the order of $O(n^2)$ in the number of nodes, and in practice was dominated by the repeated evaluation of dense expectations of the sufficient statistics. By contrast, MCMC-MLE updates a single dyad at a time and the sampled MPLE operates on a subsample of dyads; for the parameter ranges we consider their effective cost grows much more gently with n, and is largely driven by the number of sampled dyads or MCMC iterations rather than by the total number of possible ties. The CPU times in Table 2.3 are therefore not only case-specific numbers, but also indicative of fundamental scaling limitations of the MFVLE relative to standard likelihood-based methods.

2.3.3 How realistic are these models?

Consider ERGM for which the tie variables are independent. Specifically, where

$$p_\theta(Y = y|X = x) = \prod_{ij} p_\theta(y_{ij}|X = x) = \frac{\exp\{\eta(\theta) \cdot vec(y)\}}{Z(\theta)} \quad y \in \mathcal{Y}_n \tag{2.5}$$

where $\eta(\theta)$ is a known $\binom{n}{2}$-dimensional vector of y_{ij} parameters, itself parameterization by θ and $vec(y)$ is the corresponding vectorized version of y. For such a model the MPLE is identical to the MLE and can be computed deterministically. Such tie-independent models are flexible, but cannot capture the complex interactions that motivate most network modeling in economics and other social sciences. These models are relevant here because standard methods like the MPLE will tend to be both accurate and computationally efficient for network models that are close to tie-independent.

How dependent is our model? The model contains two network statistics that allow it to deviate from tie-independent models - the Two-star and Triangle statistics. A simple test is to generate networks from the two parameter vectors where the Triangle natural parameter is set to zero and thus effectively removed from the model. Then, the mean-value parameters of these altered models can be compared to the original mean-value parameters (that is, the expected numbers of Two-stars and Triangles).

θ	-4	2	$1/n$	$\theta_{Triangle}$
$\theta_{Triangle}$	**Edge**	**Homophily**	**Two-Star**	**Triangle**
0.010	393.043	340.730	3091.731	117.206
0.000	389.376	337.672	3032.674	113.559

Table 2.4: Expected sufficient statistics for the $n = 100$ model with the Triangle natural parameter at its original value (0.01) and set to zero, showing the model is close to tie-independent.

In Table 2.4, the economics model for network size $n = 100$ are adjusted so that only the

Triangle natural term is set to zero. This table shows that setting the Triangle parameter to zero causes minimal change in the expected network statistics. For instance, the ratio of the expected number of triangles ($117.2 \rightarrow 113.6$) formed to the expected number of total edges ($393.0 \rightarrow 389.4$) changes from 0.298 to 0.292. This indicates that the chosen Triangle natural parameter (0.01) is so small that the model is very close to a tie-independent model. Hence the true model parameters set are unlikely to be interesting for economic models, nor challenging to estimate as the MPLE will do well in this case.

From an applied perspective, such parameter settings correspond to networks with very weak clustering relative to what is typically observed in social or economic data. Empirical friendship, collaboration, and trade networks with similar edge densities generally exhibit substantially higher triangle counts and stronger dependence between dyads than the specifications considered by Mele and Zhu. Thus their calibration places the model in a regime that is simultaneously easy for pseudo-likelihood based methods and unrepresentative of the kinds of networks that motivate the use of ERGMs in practice.

2.3.4 Comparison to more realistic models

This analysis suggests that a more relevant and powerful assessment of the estimation methods would come from networks with increased transitivity, as these are more realistic and interesting theoretically. For simplicity, we take as our measure of transitivity of a model its expected number of triangles, with a model producing higher expected numbers of triangles being more transitive.

As noted in Section 2.2.1, the model in Mele and Zhu (MZ23) is prone to degeneracy and will be for parameter values chosen close to realistic networks. The parameter values chosen in the original study are close to tie-independent and hence the degeneracy is not strong. However, if we choose parameter values corresponding in increased transitivity, they will be degenerate or nearly so. This behavior is not realistic, except in rare cases where the degeneracy reflects economic theory. One solution is to use improved terms to measure

degeneracy such as the GWESP (SPR06). However, the triangular term is more intuitive and here we consider the tapered ERGM variant of this model which circumvents the issue of near-degeneracy while maintaining the desirable features of ERGMs (FH17). The tapered version of an ERGM model is itself an ERGM with additional terms to taper the entropy of the original ERGM. This reduces unrealistic behavior in models with increased dependence and transitivity. Tapered ERGM models have been shown to have good statistical and probabilistic properties (For a full treatment, see Blackburn and Handcock (BH22)).

Intuitively, tapering modifies the likelihood by down-weighting network configurations with extreme values of the sufficient statistics via an additional term in the model that penalizes extreme variation. This can be viewed as an entropy-regularization device that keeps the distribution away from the extreme faces of the convex support of $g(Y)$, where standard ERGMs tend to concentrate when they are nearly degenerate. In the parameter regions we consider, the tapering terms are chosen so that the mapping between natural and mean-value parameters in the neighborhood of interest is only mildly perturbed, and the target mean-value parameters are essentially unchanged. Thus tapered ERGMs retain the interpretability of the original statistics while improving the stability and robustness of the models. This stability and robustness extends to MCMC-based inference for them.

The mean-value parameter estimates are unaffected by tapering, and the natural parameter estimates are numerically close to those of the corresponding ERGM and have similar interpretations (BH22). MCMC-MLE for tapered ERGM is implemented by the R package `ergm.tapered` on GitHub (HKF21), and the tapering term is applied to the likelihood function in R package `mfergm` for MFVLE.

To consider models that are close to the original economics model, but with higher transitivity, we can choose tapered ERGM models with similar mean-value parameters, but with increased transitivity. Specifically, we choose models with 50% more transitivity than the original version of the model, thus factoring the original expected count for triangles by 1.5. This is done by solving equation (2) for θ with the left hand side set to the desired

mean-value parameter. This results in natural parameter values for the Triangle term that are substantially above zero. Without tapering, which helps move sampled statistics away from degenerate regions, substantial increases in transitivity are typically infeasible due to model degeneracy. This is particularly true for our model specification, which includes the raw Triangle statistic.

2.3.5 Increased transitivity model

Table 2.5 is a version of Table 1 with the comparison of the estimation methods with the Triangle mean-value increased by approximately 50%, while holding all other terms approximately constant. In the table, we can see this change reflected in the comparison between original mean-values and increased-transitivity mean-values. Taking a large sample with specified target statistics allows for the desired mean-value parameters − and thus the one-to-one mapped natural parameters − to be computed. For $n = 100$ we see the Triangle natural parameter is positive (0.60) to a greater magnitude compared to the original value (0.01). The mean-value parameter reflects the approximately 50% increase in transitivity ($117.5 \rightarrow 176.1$) introduced in this section.

The results in Table 2.5 show that MCMC-MLE provides greater accuracy than MFVLE, and Table 2.6 further demonstrates that the MCMC-MLE attains superior computational efficiency. For $n = 50$, the MFVLE estimate of the natural parameter for the Triangle term is 0.224 compared to the true value of 0.630, while the MCMC-MLE is 0.594. The magnitude of this difference is clarified by comparing the mean-value parameter estimates. The MCMC-MLE mean-value estimate is 20.72 compared to the true value of 20.87, while the MFVLE estimate is 14.61. The MFVLE estimate is unable to capture even the qualitative character of the transitivity, while the MCMC-MLE is very accurate, as would be expected from theory.

Table 2.6 shows that the tapering effect has reduced all RMSE and MAD values compared to Table 2.2 despite the increased transitivity. We see that the MCMC-MLE and MPLE

n = 50	**Edge**	**Homophily**	**Two-Star**	**Triangle**	**Degeneracy**
Increased-Transitivity Tapered θ	-3.732	1.795	-0.033	0.630	-
MCMC-MLE Natural	-3.674	1.836	-0.044	0.594	0.0 %
MFVLE Natural	-17.429	-6.176	-0.463	0.224	10.6 %
MPLE Initialization	-3.614	1.827	-0.048	0.599	0.0 %
Original Mean-Value	95.75	82.71	363.42	13.91	-
Increased-Transitivity Mean-Value	95.31	82.66	363.21	20.87	-
MCMC-MLE Mean-Value	95.39	82.76	362.81	20.72	-
MFVLE Mean-Value	80.44	55.37	369.93	14.61	-
MPLE Mean-Value	95.93	83.13	365.07	20.77	-
n = 100	**Edge**	**Homophily**	**Two-Star**	**Triangle**	**Degeneracy**
Increased-Transitivity Tapered θ	-3.022	1.530	-0.066	0.600	-
MCMC-MLE Natural	-3.009	1.552	-0.068	0.589	0.0 %
MFVLE Natural	-15.194	-1.217	-0.245	0.401	7.2 %
MPLE Initialization	-2.985	1.548	-0.068	0.590	0.0 %
Original Mean-Value	393.05	341.02	3092.06	117.48	-
Increased-Transitivity Mean-Value	393.45	341.16	3092.28	176.12	-
MCMC-MLE Mean-Value	393.53	341.20	3093.09	175.92	-
MFVLE Mean-Value	277.97	201.90	2720.45	141.20	-
MPLE Mean-Value	393.85	341.36	3096.08	176.01	-

Table 2.5: Natural and mean-value parameter estimates for the tapered model with 50% increased transitivity at $n \in \{50, 100\}$, comparing MCMC-MLE, MFVLE, and MPLE across 1,000 simulated networks.

$\theta = \{-3.022, 1.530, -0.066, 0.600\}$		**Edge**	**Homophily**	**Two-Star**	**Triangle**
MCMC-MLE Natural	RMSE	0.511	0.282	0.059	0.217
	MAD	0.433	0.244	0.051	0.185
MFVLE Natural	RMSE	9.951	15.428	0.185	0.393
	MAD	9.368	14.580	0.176	0.370
MPLE Natural	RMSE	0.523	0.283	0.061	0.222
	MAD	0.443	0.244	0.052	0.191
MCMC-MLE Mean-Value	RMSE	3.819	5.084	22.463	4.128
	MAD	3.364	4.331	19.076	3.513
MFVLE Mean-Value	RMSE	47.300	54.517	206.291	14.563
	MAD	34.567	40.567	157.697	11.892
MPLE Mean-Value	RMSE	3.845	5.138	22.517	4.158
	MAD	3.388	4.389	19.160	3.545

Table 2.6: RMSE and MAD of natural and mean-value parameter estimates for the $n = 100$ increased transitivity model in Table 2.5, across 1,000 simulated networks.

remain accurate and stable estimators for this model with the RMSE and MAD slightly better for the MCMC-MLE than the MPLE. However, the MFVLE performs very poorly. To start, the MFVLE estimate fails to converge in 10.6% of the fits (despite the starting value being very accurate) whereas tapering has allowed MCMC-MLE to always produce a result. Compared to the original study, where MFVLE had a 9% degeneracy in Table 2.1, the slight increase observed here can be explained by two opposing effects. Increased transitivity pushes MFVLE toward more degenerate configurations, whereas the tapering effect pulls the model away from degeneracy. While setting a stronger tapering parameter would lead to less degeneracy for MFVLE, the results produced here (using default tapering parameter of 2, detailed in Blackburn and Handcock (BH22)) highlight its behavior. Secondly, both the natural and mean-value parameters are biased. Third, the RMSE and MAD are quite poor compared to those of MCMC-MLE and MPLE. For the mean-value Triangle parameter, MFVLE shows a RMSE of 14.56 where MCMC-MLE has RMSE of 4.128 and MPLE has

RMSE 4.158. It is worth noting that tapering ERGM allows MPLE to have a greater tolerance for transitivity, and become almost as useful as MCMC-MLE in many scenarios. We have separately computed MFVLE without tapering, and those natural parameters yielded RMSE of $\{17.023, 22.456, 0.144, 0.199\}$ and MAD of $\{15.414, 20.708, 0.128, 0.183\}$. Both of these measures are noticeably worse than those for its tapered counterpart. As a strict comparison between the variational methods, MFVLE is dominated by the MPLE on all measures including computation time.

In our Section 2.3 findings, we see that the formation of triangles simulated from the models used by Mele and Zhu (MZ23) do not differ significantly from the same model with the transitivity term is set to zero. Under a model with no transitivity, MPLE produces MLE results and therefore leads to good mean-field approximation results due to initialization. Furthermore, the increase in transitivity for the microeconomics model led to the mean-field method producing dramatically fewer edges compared to true simulated networks.

2.4 Discussion

The fitting of a complex model that realistically represent economic and social systems is challenging. The dependence structure that makes these models appealing also makes likelihood-based inference computationally demanding. As such, the search for accurate and computationally tractable methods is ongoing. While likelihood-based estimation methods can be computational expensive, they do have a strong track record of success and a long history of development.

In this paper we have shown that there is little empirical and conceptual support for the mean-field variational estimator. We show that there is a computational error in the original paper (MZ23) that invalidates their empirical results. We also argue that, even if those results had been correct, they would not have provided much support for the method as the models they were assessed on are unrealistically close to tie-independent models. When

we alter the model to make it more realistic by inducing stronger dependence among ties, the estimator performs poorly and is more computationally expensive compared to both MCMC-MLE and the cheap sampled MPLE.

Variational methods do have value when they are applied to appropriate model classes with parameter values in regions where the approximation is not bad. This is challenging for ERGMs, which constitute a highly flexible class of models: depending on the choice of sufficient statistics, they can induce strong and complex dependence among ties that is difficult for variational approximations to capture. In addition, variants that aim to capture realistic and complex network structure tend to be those where the approximation does poorly. Our results suggest that mean-field variational approximation should not be used as a default method for ERGMs, and a more robust strategy would be to use MPLE for initialization and MCMC-MLE for final inference.

Our analysis also illustrates the value of tapered ERGM both for model specification and estimation. They are ERGM that retain the expressiveness of traditional ERGM while allowing practical MCMC-based methods of inference. In sum, when using ERGM, researchers should usually prefer the tapered ERGM. For inference, they should use MPLE for initialization and MCMC-MLE, if computationally feasible, for reliable inference. MFVLE is not recommended for any model.

A theoretical question left open by this analysis is whether MPLE admits convergence guarantees analogous to those Mele and Zhu (MZ23) establish for MFVLE. Their Proposition 1 shows that the mean-field log-normalizing constant converges to the true log-normalizing constant under a dense-graph asymptotic regime. The corresponding result for MPLE has not been settled in the literature. Because the MPLE is the full-conditional maximum likelihood, and the full-conditional reduces to the joint under tie independence, we conjecture that MPLE inherits convergence properties at least as favorable as those of MFVLE in the regime covered by Mele and Zhu. MPLE also remains well-defined when tie dependence is strong, a regime in which our simulations show that MFVLE fails. A formal treatment of

this conjecture is deferred to future work.

The analysis in this paper was produced using the ergm (HHB03) and ergm.tapered (HKF21; BH22) **R**-packages within the statnet suite (HHB08b). This is sophisticated, accessible, user-friendly and open-source software for fitting and simulating ERGM. The **R** code to reproduce the results of this paper using these packages is available on GitHub (Res24) and will be published Zenodo on acceptance of this paper (RH25).

For completeness it is worth emphasizing that the fully factorized mean-field family studied by Mele and Zhu is only one member of a broader class of deterministic approximations to ERGMs. Composite or pseudo-likelihood based methods, corrected pseudo-likelihoods (DHG09), and variational approximations based on richer dependence structures (for example, Bethe approximations (Ric12) and loopy belief propagation (JKM18)) retain more of the local dependence among ties than a product-Bernoulli mean-field family. Our negative results therefore should not be interpreted as a blanket indictment of all deterministic or variational approaches for ERGMs, but rather as evidence that a tie-independent mean-field family is too restrictive to serve as a default likelihood approximation for models with substantively meaningful transitivity and clustering.

CHAPTER 3

Reference Prior Distribution for ERGM: Non-Degeneracy Prior

3.1 Introduction

Exponential-family random graph models (ERGMs) are a widely used class of statistical models for network data (FS86; HHB03). The ERGM framework is reviewed in the Introduction (Chapter 1). We recall that an observed network on n nodes is treated as a realization $y \in \mathcal{Y}_n$ of a random network Y, with probability

$$p_\theta(Y = y) = \frac{\exp\{\theta^\top g(y)\}}{Z(\theta)}, \qquad y \in \mathcal{Y}_n, \tag{3.1}$$

where $g(y) \in \mathbb{R}^d$ is a vector of network statistics, $\theta \in \mathbb{R}^d$ is the natural parameter, and $Z(\theta) = \sum_{y \in \mathcal{Y}_n} \exp\{\theta^\top g(y)\}$ is the normalizing constant. Common choices of $g(y)$ capture density, degree structure, and triadic dependence such as transitivity, and can incorporate nodal covariates through terms representing homophily or mixing (MHH08; FS86; GKM09).

Since $Z(\theta)$ is intractable, ERGM inference depends on simulation-based or approximate methods, as reviewed in the Introduction (Chapter 1) and Chapter 2. These computational challenges are compounded in Bayesian ERGMs because the posterior depends on the same intractable normalizing constant. In particular, standard Metropolis–Hastings updates require ratios of intractable normalizing constants, which is known as a doubly intractable posterior.

Bayesian inference for ERGMs has often relied on default priors, mostly Gaussian and

uniform distributions on the natural parameter space. While these choices simplify computation and implementation, they can be unintuitive as representations of substantive prior beliefs about network formation. This means that they do not directly reflect the well known fact that large regions of ERGM parameter space correspond to near-degeneracy behavior, in which the model concentrates most of its density on a small subset of extreme (i.e. near-empty or near-complete) networks (Han03). Such priors are commonly used less for their interpretability and effectiveness than for their convenience and generality. This reflects, in part, that widely applicable and principled prior specifications for ERGMs remain limited. In this chapter we propose a prior distribution defined through the ERGM's non-degeneracy region, motivated by the idea that a meaningful prior for network formation should place greater density on parameter values that generate realistic graphs. We define the non-degeneracy prior, compute it exactly in a small setting where ERGM normalizing constant is known, and then develop an approximation strategy for the practically relevant setting in which the normalizing constant is unavailable.

For posterior computation we use an existing state-of-the-art method for doubly intractable targets, noisy Hamiltonian Monte Carlo (noisy HMC; SBF19). Noisy HMC serves here purely as a practical sampling framework for Bayesian ERGMs under the proposed prior, and we note that other approaches for doubly intractable posteriors could be used instead. In our results, we contribute detailed analysis of the proposed prior distribution for all networks on seven actors, and conduct an empirical analysis using the Kapferer tailor shop data with 39 actors.

The remainder of the chapter is organized as follows. Section 3.2 reviews Bayesian inference for ERGMs and details noisy Hamiltonian Monte Carlo (noisy HMC), which we use as our posterior sampling framework. Section 3.4 defines the non-degeneracy prior, develops practical approximations plus exact computation when feasible, and illustrates the construction for the space of all networks on seven actors. Section 3.5 computes posterior distributions under the non-degeneracy prior and compares them to posteriors obtained un-

der common alternative prior specifications. Section 3.7 applies the proposed prior to the Kapferer tailor shop network. Section 3.8 concludes with a discussion of findings and final remarks.

3.2 Bayesian Methods for ERGM

Bayesian inference for ERGMs is challenging because the likelihood computation is limited by its normalizing constant. As a result, the posterior is doubly intractable and standard MCMC updates require ratios of intractable normalizing constants in the acceptance probability. This difficulty has motivated a large body of work on Bayesian computation for ERGMs, which we survey in this section.

Bayesian computation for ERGMs with intractable normalizing constants is often divided into two classes of strategies, based on how the normalizing constant $Z(\theta)$ is handled in posterior calculations. The first class comprises approximation-based strategies that approximate $Z(\theta)$ itself, ratios such as $Z(\theta)/Z(\theta')$, or derivatives such as $\nabla_\theta \log Z(\theta)$. As an early example, (GT92) introduced Monte Carlo maximum likelihood estimation (MCMLE), using importance sampling under a trial distribution centered at a specified parameter value to approximate $Z(\theta)$. Related developments include the adaptive kernel-smoothing approaches of (Lia07) and (ALR13), which treat $Z(\theta)$ as a marginal density associated with the unnormalized exponential-family form $\exp[\theta^\top g(y)]$. For Bayesian posterior sampling, (LJ13) proposed the Monte Carlo Metropolis–Hastings (MCMH) algorithm, which estimates $Z(\theta)/Z(\theta')$ at each iteration using auxiliary networks simulated from $p(y \mid \theta)$ or $p(y \mid \theta')$ via an inner Markov chain. Alquier et al. (AFE16) subsequently characterized MCMH as a noisy exchange method and derived approximation rates in terms of the inner-chain length. The Bayesian stochastic approximation Monte Carlo (SAMC; JL14) and marginal MCMC (Eve12) procedures also belong to this noisy-exchange family and typically require generating many auxiliary networks per iteration for stable convergence. More recently, Alquier et

al. (AFE16) developed a noisy Langevin scheme that uses inner-chain samples to approximate $\nabla_\theta \log Z(\theta)$. The noisy Hamiltonian Monte Carlo (SBF19) adapts the same principle within Hamiltonian Monte Carlo updates by using auxiliary simulations to approximate both $\nabla_\theta \log Z(\theta)$ and $Z(\theta)/Z(\theta')$. Along similar noisy gradient direction for ERGMs, Zhang and Liang (ZL24) study stochastic gradient Langevin dynamics (SGLD), where a short inner MCMC run is used to estimate $E_\theta[g(y)]$ and thereby obtain an estimate of $\nabla_\theta \log Z(\theta)$.

The second category consists of auxiliary variable based methods, which aim to cancel the intractable normalizing constant ratio by augmenting either the target distribution or the proposal distribution with auxiliary variables drawn from the likelihood at proposed parameter values. Early auxiliary-variable constructions include the auxiliary-variable MCMC method of Moller et al. (MPR06) and the exchange algorithm of Murray et al. (MGM06), which in their original forms relied on perfect sampling and were therefore limited in scope. Practical ERGM implementations typically replace perfect sampling with finite inner MCMC runs, yielding approximately-correct methods such as double Metropolis–Hastings (DMH; Lia10) and related approaches for social networks. To reduce the bias induced by short inner chains, adaptive exchange (AEX; LJS16) generates auxiliary samples using an importance sampling procedure that runs in parallel to the target chain.

While these methods differ computationally, they all face the practical need to simulate from ERGMs under proposed parameters, which motivates careful prior specification to avoid degenerate regions of the parameter space. In this chapter, we use noisy HMC as a practical posterior sampling framework for Bayesian ERGMs under the proposed non-degeneracy prior. We detail in the following section the noisy HMC algorithm to fix notation and clarify how auxiliary sampling is used to estimate both gradients and acceptance probabilities in the doubly intractable setting.

3.3 Noisy Hamiltonian Monte Carlo

In a recent advancement of approximation-based Bayesian inference for ERGMs, noisy Hamiltonian Monte Carlo (noisy HMC; SBF19) follows the standard HMC construction by augmenting $\theta \in \mathbb{R}^d$ with an auxiliary momentum $r \sim \mathcal{N}(0, M)$, defining the Hamiltonian

$$H(\theta, r) = -\log \pi(\theta \mid y) + \tfrac{1}{2} r^\top M^{-1} r, \tag{3.2}$$

and proposing (θ', r') via a leapfrog discretization of Hamiltonian dynamics. The key distinction is that the leapfrog updates require $\nabla_\theta \log \pi(\theta \mid y)$, which is not directly available for ERGMs. Using the identity

$$\nabla_\theta \log Z(\theta) = \mathbb{E}_\theta[g(Y)], \tag{3.3}$$

the log-posterior gradient can be written as

$$\nabla_\theta \log \pi(\theta \mid y) = g(y) - \mathbb{E}_\theta[g(Y)] + \nabla_\theta \log p(\theta). \tag{3.4}$$

Noisy HMC replaces $\mathbb{E}_\theta[g(Y)]$ with an empirical average computed from auxiliary draws $u^{(1,\theta)}, \ldots, u^{(N,\theta)} \sim p(\cdot \mid \theta)$, yielding the Monte Carlo gradient estimate

$$\widehat{\nabla}_\theta \log \pi(\theta \mid y) = g(y) - \frac{1}{N} \sum_{n=1}^{N} g\big(u^{(n,\theta)}\big) + \nabla_\theta \log p(\theta). \tag{3.5}$$

These estimated gradients are used at each leapfrog step to evolve (θ, r) over a trajectory of L steps with step size ϵ.

A second difficulty is the HMC accept–reject step, which for doubly intractable targets involves the unknown ratio $Z(\theta)/Z(\theta')$. A common importance-sampling identity implies that if $U \sim p(\cdot \mid \theta')$, then

$$\mathbb{E}\left[\frac{\exp\{\theta^\top g(U)\}}{\exp\{(\theta')^\top g(U)\}}\right] = \frac{Z(\theta)}{Z(\theta')}, \tag{3.6}$$

but (SBF19) show that the corresponding "point" estimator can become unreliable when $\|\theta - \theta'\|$ is large, precisely the regime where HMC is most useful. Their solution is the

leapfrog estimator, which exploits the full sequence of intermediate parameters visited along the numerical trajectory. Let $\theta_0 = \theta$ and $\theta_L = \theta'$ denote the start and end of a leapfrog trajectory, and let $\theta_0, \theta_1, \ldots, \theta_L$ be the intermediate parameter values. The ratio of normalizing constants is decomposed into local ratios,

$$\frac{Z(\theta_0)}{Z(\theta_L)} = \prod_{\ell=0}^{L-1} \frac{Z(\theta_\ell)}{Z(\theta_{\ell+1})}, \tag{3.7}$$

and each local ratio is estimated using auxiliary draws generated at $\theta_{\ell+1}$:

$$\widehat{\frac{Z(\theta_\ell)}{Z(\theta_{\ell+1})}} = \frac{1}{N} \sum_{n=1}^{N} \frac{\exp\{\theta_\ell^\top g(u^{(n,\theta_{\ell+1})})\}}{\exp\{\theta_{\ell+1}^\top g(u^{(n,\theta_{\ell+1})})\}}. \tag{3.8}$$

Multiplying these local estimates produces an estimator of $Z(\theta)/Z(\theta')$ that is substantially more stable for long trajectories because each factor compares nearby parameter values (SBF19). Importantly, the auxiliary draws required for Equation 3.8 are the same draws used to form the gradient estimates Equation 3.5, so the improved ratio estimation is obtained with essentially no additional simulation beyond evaluating the unnormalized likelihood at adjacent states along the path.

Putting these components together, one noisy HMC iteration proceeds as follows: sample $r \sim \mathcal{N}(0, M)$; generate a leapfrog trajectory using the Monte Carlo gradient estimates Equation 3.5 at each θ_ℓ; compute the leapfrog estimator of $Z(\theta)/Z(\theta')$ via Equation 3.7–Equation 3.8; and accept the proposed endpoint (θ', r') with probability

$$\widehat{\alpha} = 1 \wedge \exp\left\{ \log q_{\theta'}(y) - \log q_\theta(y) + \log p(\theta') - \log p(\theta) + \log \widehat{\frac{Z(\theta)}{Z(\theta')}} - \tfrac{1}{2}\left(r'^\top M^{-1} r' - r^\top M^{-1} r\right) \right\}, \tag{3.9}$$

where $q_\theta(y) = \exp\{\theta^\top g(y)\}$ and $\widehat{Z(\theta)/Z(\theta')}$ is obtained from the leapfrog estimator. As emphasized by (SBF19), the resulting Markov kernel is approximate for finite N (and, in practice, for finite inner MCMC used to obtain approximately-distributed auxiliary networks). Nevertheless, increasing N and improving the quality of auxiliary simulation reduces the noise in both Equation 3.5 and Equation 3.9, and noisy HMC can provide an effective practical sampler in doubly intractable models.

In our implementation, noisy HMC also aligns naturally with the proposed non-degeneracy prior detailed in the following section. The prior approximation developed in Section 3.4 is itself driven by Monte Carlo summaries of networks simulated from $p(\cdot \mid \theta)$ at parameter values visited by the sampler. We therefore reuse the same auxiliary networks needed for Equation 3.5–Equation 3.8 when evaluating and differentiating the prior approximation, so that posterior sampling under the non-degeneracy prior does not require additional simulation tools beyond the auxiliary ERGM draws already required by noisy HMC.

3.4 The Non-Degeneracy Prior

Standard prior choices for Bayesian ERGMs, most commonly the multivariate normal or uniform distributions on the natural parameter space, do not account for the geometry of the model. Large regions of the ERGM parameter space correspond to near-degenerate behavior, in which the model concentrates most of its density on a small subset of extreme (near-empty or near-complete) networks (Han03). This means that a standard prior in θ-space implicitly places substantial density on parameter values that generate unrealistic networks, although almost no practical scenario would intend such a belief. This reflects an artifact of specifying priors without regard to the structure of the sufficient statistic space.

We propose a prior distribution that places greater density on parameter values that generate non-degenerate networks. The construction follows the idea that a sensible and interpretable prior for network formation should place greater density on parameter values under which the ERGM generates networks whose sufficient statistics fall in the interior of the convex hull of realizable statistics, rather than at its boundary. In the following, we define the non-degeneracy prior and its strength-adjusted generalization (Sections 3.4.1–3.4.2), present exact computation when all graph configurations can be enumerated (Section 3.4.3), and develop an importance-reweighting approximation for the setting in which they cannot (Section 3.4.4).

3.4.1 Key Definitions

Let $\mathcal{C} = \text{conv}\{g(y) : y \in \mathcal{Y}_n\}$ denote the convex hull of sufficient statistic vectors over all networks on n nodes, with boundary $\partial\mathcal{C}$ and interior $\text{int}(\mathcal{C})$. Networks such as the empty graph or the complete graph whose sufficient statistics lie on $\partial\mathcal{C}$ correspond to extreme configurations that are characteristic of degenerate model behavior. Conversely, networks with $g(y) \in \text{int}(\mathcal{C})$ represent the non-degenerate regime in which the model distributes density across a diverse set of graph structures.

For a given parameter value θ, the probability that a network drawn from the ERGM has sufficient statistics in the interior is

$$P_\theta\big(g(Y) \in \text{int}(\mathcal{C})\big) = \sum_{y \in \mathcal{Y}_n} \frac{\exp\{\theta^\top g(y)\}}{Z(\theta)} \, \mathbb{1}\big\{g(y) \in \text{int}(\mathcal{C})\big\}. \tag{3.10}$$

This quantity is close to one when the model under θ generates realistic, non-degenerate networks, and close to zero when θ lies in a degenerate region of the parameter space. Hence, a prior distribution based on this probability could be a desirable candidate for encoding substantive beliefs about network formation.

The non-degeneracy prior is then defined such that each parameter value receives density in proportion to its non-degeneracy probability:

$$p(\theta) \propto P_\theta\big(g(Y) \in \text{int}(\mathcal{C})\big), \tag{3.11}$$

where the proportionality constant ensures that $p(\theta)$ integrates to one over the natural parameter space. The effect here is to concentrate prior density on θ-space regions that produce non-degenerate networks, while putting less density on the degenerate regions that would otherwise dominate under standard prior choices.

It follows that the posterior distribution takes the standard Bayesian form,

$$\pi(\theta \mid y) = \frac{p(\theta)\, p(y \mid \theta)}{c(y)}, \tag{3.12}$$

where $p(y \mid \theta) = \exp\{\theta^\top g(y)\}/Z(\theta)$ is the ERGM likelihood and $c(y) = \int p(\theta)\, p(y \mid \theta)\, d\theta$ is the marginal likelihood. In the majority of published Bayesian ERGM studies, the prior $p(\theta)$ in Equation 3.12 is a multivariate normal distribution, and occasionally a multivariate uniform distribution. The exchange of the non-degeneracy prior for these standard choices is straightforward once computed or approximated. Thus, no modification to the likelihood or the posterior sampling algorithm is required beyond substituting for $p(\theta)$.

3.4.2 Strength-Adjusted Non-Degeneracy Prior

The standard non-degeneracy prior as defined in Equation 3.11 is configured such that the prior belief is equivalent to a single observed network whose sufficient statistics do not lie on $\partial\mathcal{C}$. In many cases, a modeler could have stronger or weaker confidence that the network formation process is non-degenerate. The strength of this belief can be adjusted proportional to the number of prior observed networks, resulting in the class of non-degeneracy priors with strength hyperparameter p and countering hyperparameter q for the rare event of previously observed networks being on the hull:

$$p(\theta) \propto P_\theta\big(g(Y) \in \operatorname{int}(\mathcal{C})\big)^p \, P_\theta\big(g(Y) \in \partial\mathcal{C}\big)^q, \tag{3.13}$$

where $p \geq 1$ denotes the amount of prior information, measured as the equivalent of p observed networks whose sufficient statistics do not lie on the boundary of the convex hull. Similarly, $q \geq 0$ represents q observed networks with statistics that lie on the boundary, on the rare event that any such network was observed. That is, the hyperparameter q defaults to zero in most settings, and the case $p = 1$, $q = 0$ recovers the baseline prior in Equation 3.11.

This hyperparameterization is analogous to the effective prior sample size in conjugate Bayesian models. Just as increasing the concentration parameter of a Dirichlet prior by a factor of p corresponds to having observed p additional datasets, raising the non-degeneracy probability to the power p sharpens the prior's exclusion of degenerate parameter regions

by the same factor. In practice, greater values of p shrink the posterior toward $\text{int}(\mathcal{C})$ to concentrate greater posterior density on parameter values that generate realistic networks while pulling further away from the hull. The choice of p can be thought of as the degree of confidence that the network formation process is non-degenerate. More practically, this choice dictates how willing we are for the non-degeneracy prior to influence the likelihood for our posterior computation.

A reasonable way to select p is via prior predictive analysis, sometimes called the device of imaginary results. One draws θ from the prior and simulates a network from $p_\theta(Y = y)$, repeating to build the prior predictive distribution of $g(Y)$. A practitioner then considers which region of the sufficient statistic space typical observed networks should occupy, and chooses p so that the prior predictive concentrates density on that region. The empirical recommendations developed later in this chapter calibrate p from posterior diagnostics under specific observed networks.

3.4.3 Exact Computation

When n is small enough that all $|\mathcal{Y}_n|$ graph configurations can be enumerated, the non-degeneracy prior can be computed almost exactly. We illustrate this for the edges and two-stars ERGM on $n = 7$ nodes, where the natural parameter is $\theta = (\theta_e, \theta_t)^\top$ and the sufficient statistic is $g(y) = (e(y), t(y))^\top$, with $e(y)$ the edge count and $t(y)$ the two-star count. There are $2^{\binom{7}{2}} = 2,097,152$ graphs on seven nodes, with 144 distinct combinations of edge and two-star counts. Let $f(e, t)$ denote the number of graphs with e edges and t two-stars.

The normalizing constant can be computed by summing over these 144 distinct statistic vectors, such that

$$Z(\theta_e, \theta_t) = \sum_e \sum_t f(e, t)\, \exp(\theta_e\, e + \theta_t\, t). \tag{3.14}$$

For natural ERGM coefficients θ_e and θ_t within the convex hull, the exponential family form guarantees a one-to-one mapping to mean-value parameters $\mu_e(\theta_e, \theta_t)$ for edges and $\mu_t(\theta_e, \theta_t)$ for two-stars (Han03). Hence, we define the mean-value parameters as

$$\mu_e(\theta_e, \theta_t) = \sum_e \sum_t e \, \frac{f(e,t) \, \exp(\theta_e \, e + \theta_t \, t)}{Z(\theta_e, \theta_t)}, \qquad \mu_t(\theta_e, \theta_t) = \sum_e \sum_t t \, \frac{f(e,t) \, \exp(\theta_e \, e + \theta_t \, t)}{Z(\theta_e, \theta_t)}, \tag{3.15}$$

where each μ is summed over the total of 144 distinct combinations of edge and two-star counts.

We are interested in the inverse solution to Equation 3.15 to translate the non-degenerate space, i.e., the translation of the convex hull observed in the mean-value space to the natural parameter space. This can be done by optimizing for θ_e and θ_t in the Euclidean distance between a specific μ to be mapped, μ_{target}, and $\mu(\theta_e, \theta_t)$ to be iterated on. That is, the one-to-one θ mapping of μ_{target} is defined

$$\theta(\mu_{\text{target}}) = \underset{\theta_e, \, \theta_t}{\arg\min} \; \left\| \mu_{\text{target}} - \mu(\theta_e, \theta_t) \right\|^2. \tag{3.16}$$

Of the 144 distinct (e, t) combinations, ten lie on $\partial\mathcal{C}$ and form the convex hull vertices. Under the mapping between μ and θ, the prior using the non-degeneracy standard is defined. Here, the prior is defined as the probability that parameters θ produce networks in the interior of the convex hull, such that

$$P_{\theta_e, \theta_t}\big(g(Y) \in \text{int}(\mathcal{C})\big) = 1 - \sum_{(e,t) \in \partial\mathcal{C}} \frac{f(e,t) \, \exp(\theta_e \, e + \theta_t \, t)}{Z(\theta_e, \theta_t)}, \tag{3.17}$$

which results in the exact non-degeneracy prior density over $\mathbb{R}^2$ after normalization. A caveat of this exact computation is that the mapping between θ- and μ-spaces is highly nonlinear and sensitive, meaning that a small shift in θ values often results in a large and disproportionate change in the corresponding μ. This is a known artifact of ERGMs, and

has consequences in practice for how the prior is computed and visualized, which we detail in Section 3.5.

3.4.4 Approximation via Importance Reweighting

For networks of realistic size, the non-degeneracy prior cannot be computed exactly because the $2^{\binom{n}{2}}$ graph configurations cannot be enumerated. We approximate it using an importance-reweighting estimator that reuses auxiliary networks already generated by the noisy HMC sampler at each leapfrog step.

At each leapfrog step, noisy HMC draws auxiliary networks $x_1, \ldots, x_m \sim p(\cdot \mid \widetilde{\theta})$ at the current parameter $\widetilde{\theta}$ for gradient estimation (Equation 3.5) and normalizing constant ratio estimation (Equation 3.8). To evaluate the non-degeneracy probability at a nearby parameter $\theta \neq \widetilde{\theta}$, we reweight these draws using the exponential family likelihood ratio:

$$w_i(\theta, \widetilde{\theta}) = \frac{p(x_i \mid \theta)}{p(x_i \mid \widetilde{\theta})} = \frac{Z(\widetilde{\theta})}{Z(\theta)} \exp\big\{(\theta - \widetilde{\theta})^\top g(x_i)\big\}. \tag{3.18}$$

Because both the numerator and denominator of the self-normalized estimator use the same importance weights, $Z(\widetilde{\theta})/Z(\theta)$ cancels exactly, giving

$$\widehat{P}_\theta\big(g(Y) \in \mathrm{int}(\mathcal{C})\big) = \frac{\displaystyle\sum_{i=1}^{m} \exp\big\{(\theta - \widetilde{\theta})^\top g(x_i)\big\}\, \mathbb{1}\big\{g(x_i) \in \mathrm{int}(\mathcal{C})\big\}}{\displaystyle\sum_{i=1}^{m} \exp\big\{(\theta - \widetilde{\theta})^\top g(x_i)\big\}}. \tag{3.19}$$

The approximation depends only on sufficient statistic differences and hull membership indicators, and the intractable normalizing constants do not appear. The estimator is consistent as $m \to \infty$ with the usual properties of self-normalized importance sampling. Evaluating the indicator $\mathbb{1}\{g(x_i) \in \mathrm{int}(\mathcal{C})\}$ requires the convex hull $\mathcal{C}$ of the achievable sufficient statistic region. For the $n = 7$ network in Section 3.5, $\mathcal{C}$ is computed exactly from the 144 distinct statistic combinations. For larger networks, the hull must itself be approximated. The adaptive procedure used for this is described in Section 3.6.

The estimator is differentiable with respect to θ, which is necessary for its use within HMC, so the gradient $\nabla_\theta \log p(\theta)$ needed in Equation 3.4 can be computed from the same auxiliary sample. Setting $\widetilde{\theta} = \theta_\ell$ at each leapfrog step gives the prior evaluation and its gradient as a byproduct of the existing trajectory, since sampling auxiliary networks is already the dominant computational cost of each iteration and the prior adds only the indicator evaluations and the reweighting. In particular, $\widetilde{\theta}$ begins at the MPLE used to initialize the chain (Section 3.6) and updates with θ_ℓ at every leapfrog step thereafter. This choice constrains the estimator to use only auxiliary networks drawn at the current $\widetilde{\theta}$. Samples from earlier leapfrog steps or previous iterations cannot be pooled, because the normalizing constant cancellation in Equation 3.18 requires all samples to share the same reference parameter $\widetilde{\theta}$. Mixing samples drawn at different reference parameters would reintroduce the intractable ratios $Z(\widetilde{\theta}_i)/Z(\theta)$ that the self-normalization eliminates. The number of auxiliary draws N at each step therefore determines the quality of the prior approximation and cannot be supplemented by accumulation.

For the strength-adjusted prior Equation 3.13, the approximation generalizes to

$$\widehat{p}(\theta) \propto \widehat{P}_\theta\big(g(Y) \in \mathrm{int}(\mathcal{C})\big)^p \; \widehat{P}_\theta\big(g(Y) \in \partial\mathcal{C}\big)^q, \tag{3.20}$$

where $\widehat{P}_\theta(g(Y) \in \partial\mathcal{C}) = 1 - \widehat{P}_\theta(g(Y) \in \mathrm{int}(\mathcal{C}))$. The gradient with respect to θ follows by differentiating the log of Equation 3.20.

The quality of the approximation in Equation 3.19 depends on the effective sample size, which decreases when θ is far from the reference $\widetilde{\theta}$. Within noisy HMC, the leapfrog step size ϵ controls this distance at each step. The same considerations that govern gradient estimation accuracy (the choice of ϵ, the number of auxiliary draws N, and the inner MCMC chain length) therefore also govern the quality of the prior approximation. This is examined empirically in Sections 3.5 and 3.7.

3.5 Compute and Compare Non-degeneracy Prior

In this section the exact prior is computed for the edges and two-stars ERGM with $n = 7$ nodes, where all 2^{21} graph configurations can be enumerated. We build an ERGM prior distribution using the non-degeneracy probability density and show this choice of prior in both the natural θ- and mean-value μ-spaces. The baseline non-degeneracy prior with strength hyperparameters $p = 1$ and $q = 0$ is used, such that we assume the information we have is a single observed network that is not on the boundary of the convex hull. As noted in Section 3.4.3, the translation between the two parameter spaces is highly irregular and sensitive, meaning that a minuscule shift in θ values often results in a large disproportionate change in the corresponding μ. Given this distortion, computation of the prior using even a very fine grid is infeasible. The much stronger presentation of the prior can be computed using a Metropolis-Hastings sampler with Gaussian proposals. By taking a large and thinned chain to reduce autocorrelation in our samples, the non-degeneracy prior can be produced firstly in the θ-space in Figure 3.1.

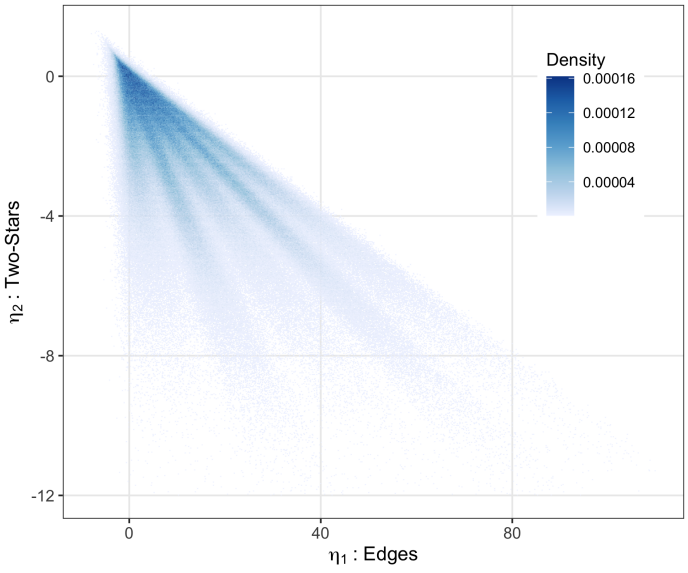


Figure 3.1: Non-degeneracy prior in the θ-space, under strength hyperparameters $p = 1$ and $q = 0$.

We first display the non-degeneracy prior density over θ to illustrate the geometry of the non-degenerate region in the natural parameterization. The density in Figure 3.1 points out a high concentration of values around the top left portion of the figure. This translates to most realistic occurring networks having natural parameters in this region. The wide spread from this concentrated area shows parameter combinations that have greater tendency to produce degenerate networks, and the importance for a prior choice to not put emphasis on the degenerate regions. The "stealth-bomber" pattern illustrated in this figure is a known behavior for ERGM non-degeneracy in the natural parameter space, with much of its density gathered at the "nose" of its shape.

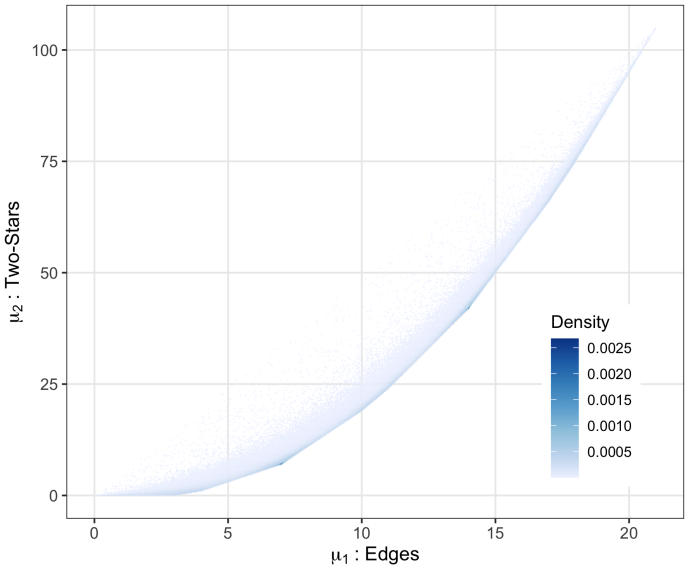


Figure 3.2: Non-degeneracy prior in the μ-space, under strength parameters $p = 1$ and $q = 0$.

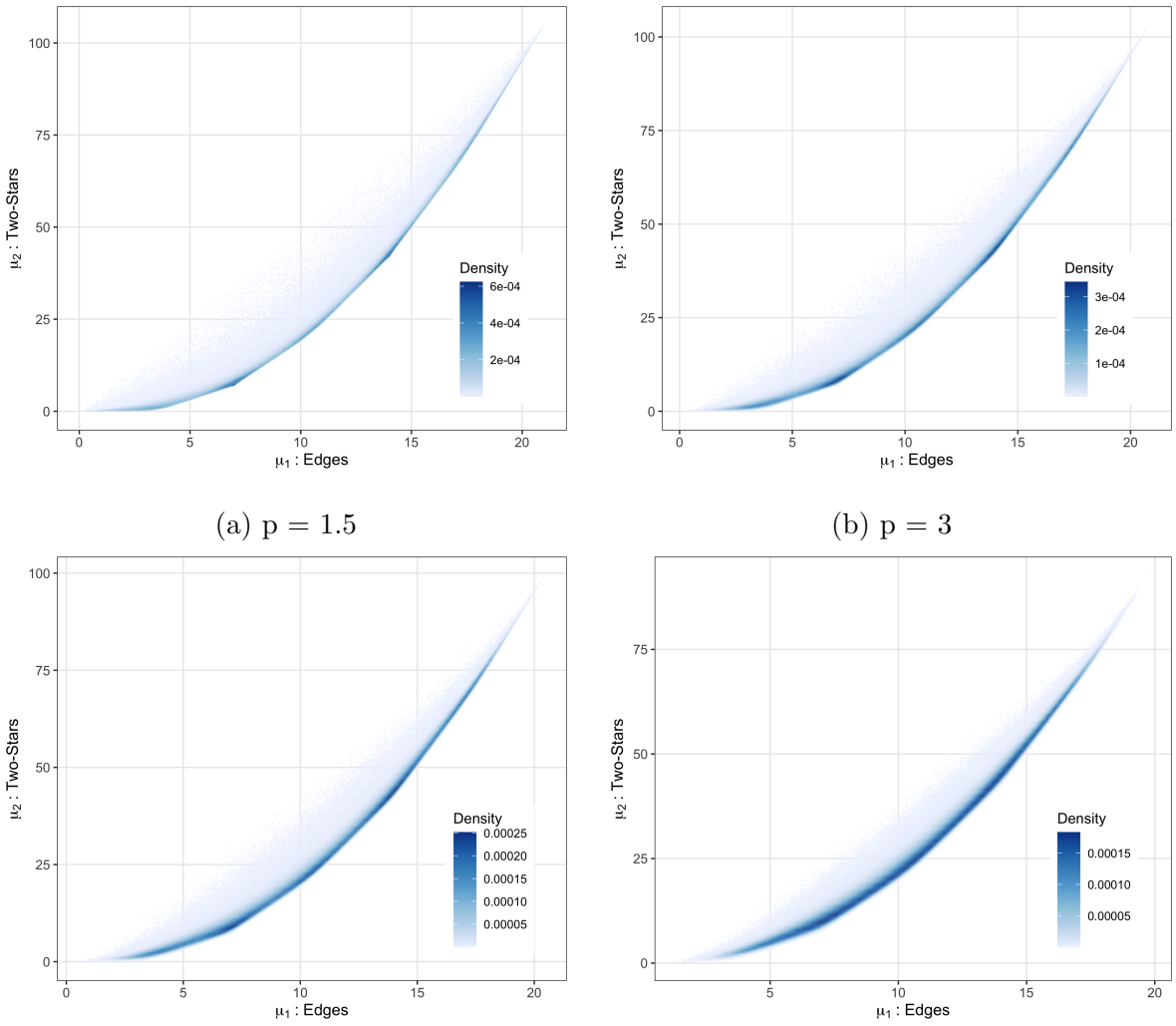
To map the stealth-bomber natural θ distribution to the μ-space, a Metropolis-Hastings sampler with Gaussian proposals is used. The several wings that are observable in the stealth-bomber pattern in Figure 3.1 induce autocorrelation, and thinned MCMC chains are used to address that. Figure 3.2 shows much more smoothness, and well-boundedness in stark contrast to its natural parameter counterpart in Figure 3.1. The mean-value space of the non-degeneracy prior is gathered near the bottom curvature (or the near-degeneracy

region) of the space, and shows which network configurations the nose of the stealth-bomber translate to.

We previously attempted to compute the non-degeneracy prior in the μ-space exactly. Networks of size $n = 7$ contain 144 distinct combinations of edge and two-star counts. Out of those, 10 combinations with those configurations make up the points that form the convex hull of the mean value μ-space. A fine grid can be created for the interior of the μ convex hull, and is then mapped to natural parameter θ-space via optimization. Probability density can be computed for the interior convex hull for θ-space and is then normalized to a full distribution to serve as the prior distribution. However, this approach fails to capture the extreme nonlinearity of the ERGM's transformation between its natural and mean-value parameter spaces. Because the grid gives equal weight to all parts of the μ-space, it failed to capture the concentration of the non-degeneracy prior near the degeneracy boundary, misrepresenting the distribution as approximately uniform.

The concentration of the prior near the degeneracy boundary can be adjusted through the strength hyperparameters p and q. In fact, as p is increased from its default value 1, the distribution spreads more evenly and concentrates less on the near-degeneracy region. The prior densities in the μ-space for $p \in \{1.5, 3, 5, 10\}$ are shown in Figure 3.3. The non-degeneracy prior for $p = 10$ shows much smoother properties than the $p = 1.5$ case, and this is further detailed in the density summaries in Table 3.1.

In comparison to the non-degeneracy prior, the default prior choices of multivariate normal and uniform distributions for Bayesian ERGMs are examined in the same fashion. These priors are defined in the natural parameter θ-space and are translated to the more interpretable mean-value μ-space. As with the non-degeneracy prior, the translation from θ-space to μ-space is highly nonlinear, making a grid-based approach unreliable. Thus, a large sample of 10,000,000 draws in θ-space is used to build the μ-space prior. The prior hyperparameters are set using popular choices. For the bivariate normal distribution prior, the BERGM package default of $N(0, 10\, I_2)$ is used. In the less common case where a uniform distribution is



(a) p = 1.5

(b) p = 3

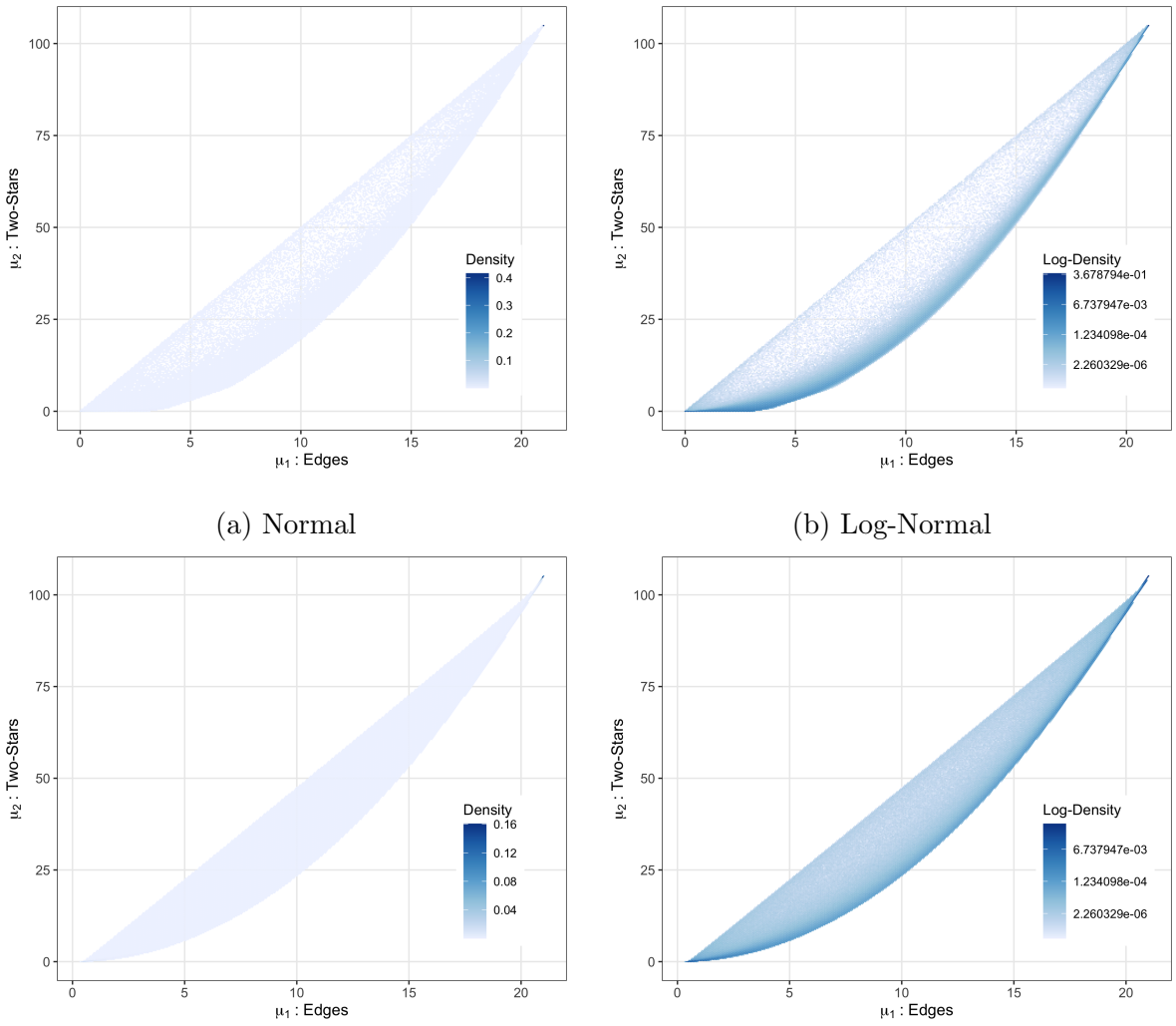
(c) p = 5

(d) p = 10

Figure 3.3: Non-degeneracy prior in the μ-space, under strength parameters $p \in \{1.5, 3, 5, 10\}$ and $q = 0$.

used instead of a normal, we follow (ZL24) in setting the prior to $\text{Unif}\big([-4, 2] \times [-0.05, 1]\big)$.

The overall normal and uniform prior distributions are given in Figure 3.4, where the normal prior is on the left and the uniform prior is on the right. The standard densities are displayed on top, while the log-densities are on the bottom. Both priors exhibit strongly concentrated density in the full and empty graph (i.e., degenerate) regions, although this



(a) Normal

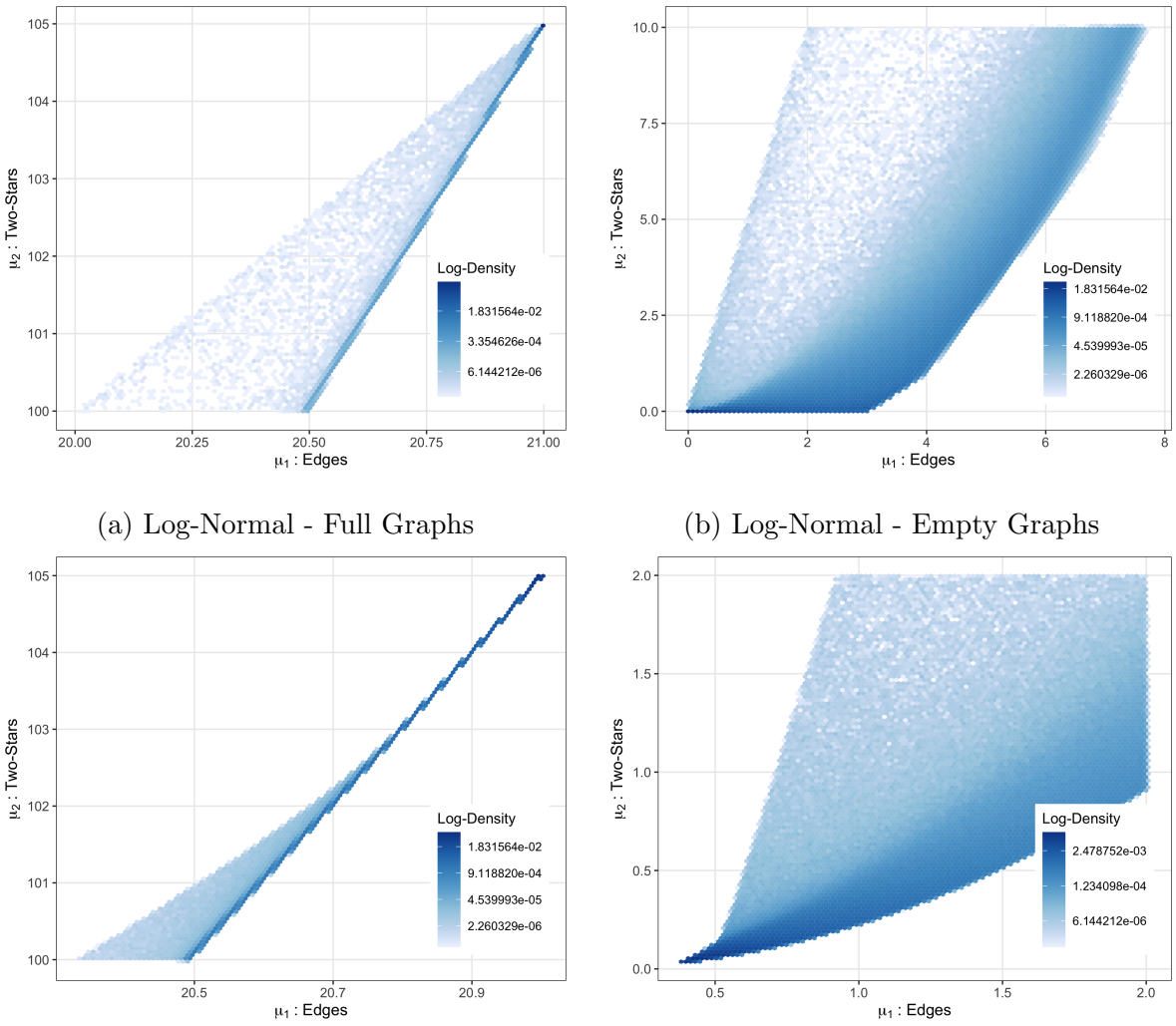
(b) Log-Normal

(c) Uniform

(d) Log-Uniform

Figure 3.4: Normal (left) and uniform (right) priors in the μ-space, with the standard densities on top and log-densities on the bottom.

concentration may be hard to notice given how small those regions are. This concentration is why the standard density legends show an upper bound of approximately 0.4 for the normal prior and 0.16 for the uniform prior. The log-densities give a better sense of where the density concentrates, but still cannot fully capture the magnitude of these concentrations. However, we can examine these regions more closely by zooming in.



(a) Log-Normal - Full Graphs

(b) Log-Normal - Empty Graphs

(c) Log-Uniform - Full Graphs

(d) Log-Uniform - Empty Graphs

Figure 3.5: Normal and Uniform density priors logged mean-value space, zoomed in at the full and empty graph regions.

Figure 3.5 zooms into the near-full graph (left) and near-empty graph (right) regions for the log-densities of the normal and uniform priors, providing more detailed insight into these concentrations. One notable difference between the normal and uniform priors is that the uniform prior is strictly bounded away from regions where the normal prior would otherwise assign density. In the log-density plots for both priors, the density gradients terminate

in sharp concentrations exactly where the empty graph $(0, 0)$ and full graph $(21, 105)$ lie. In fact, these gradients are only an artifact of the log scale and correspond to negligible differences in actual density, as seen in the left column of Figure 3.4. These concentrations are summarized in Table 3.1.

3.5.1 Assessing the impact of prior choice on degeneracy

To compare distributions over the ERGM parameter space, we introduce a predictive measure of non-degeneracy. For a given distribution over θ, the predictive non-degeneracy is defined as

$$P\big(g(Y) \in \text{int}(\mathcal{C})\big) = \mathbb{E}_{p(\theta)}\Big[P_\theta\big(g(Y) \in \text{int}(\mathcal{C})\big)\Big] = \int P_\theta\big(g(Y) \in \text{int}(\mathcal{C})\big)\, p(\theta)\, d\theta. \tag{3.21}$$

This quantity asks: if we draw a parameter θ from the distribution and then generate a network from the ERGM at that θ, what is the probability that the resulting network has non-degenerate sufficient statistics? A prior that places most of its density on realistic parameter regions will have $P(g(Y) \in \text{int}(\mathcal{C}))$ close to one, while a prior that inadvertently emphasizes degenerate regions will have $P(g(Y) \in \text{int}(\mathcal{C}))$ close to zero. That is, $P(g(Y) \in \text{int}(\mathcal{C}))$ summarizes in a single quantity how much of the prior's density sits in the non-degenerate region of the parameter space, once the ERGM likelihood is accounted for.

For the μ-space, the normal and uniform priors can be compared to the non-degeneracy prior under different values of p to examine where the densities of the prior lie. The regions of particular interest are those near the full graphs, those near the empty graphs, and those that are near degenerate and lie along the bottom belly of the μ-space distribution. These comparisons are reported in Table 3.1, where $P(g(Y) \in \text{int}(\mathcal{C}))$ is the prior predictive non-degeneracy defined in Equation 3.21. The degeneracy regions are defined by thresholds of greater than 20 edges and greater than 104 two-stars (near-full graphs), and fewer than 3 or 1 edges with fewer than 1 two-star (near-empty graphs). Notably, the normal prior places

Prior Choice	$\boldsymbol{P(g(Y) \in \text{int}(\mathcal{C}))}$	**< 0.1 away from hull**	**> 20 Edges > 104 Two-stars**	**< 3 Edges < 1 Two-stars**	**< 1 Edges < 1 Two-stars**
Normal	0.154	0.374	0.432	0.286	0.122
Uniform	0.244	0.796	0.346	0.145	0.093
NDG, $p = 1$	0.398	0.931	0.00009	0.0522	0.00546
NDG, $p = 1.5$	0.505	0.944	0.000005	0.037	0.002
NDG, $p = 2$	0.577	0.948	0	0.0272	0.00083
NDG, $p = 3$	0.671	0.951	0	0.014	0.00014
NDG, $p = 5$	0.769	0.954	0	0.00402	0.0000045
NDG, $p = 10$	0.861	0.957	0	0.000166	0
NDG, $p = 20$	0.915	0.960	0	0	0

Table 3.1: Comparison of non-degeneracy prior (NDG) against normal and uniform priors by density inside convex hull $P(g(Y) \in \text{int}(\mathcal{C}))$, density 0.1 Euclidean distance away from hull, and density in the degeneracy regions, i.e. greater than 20 edges and greater than 104 two-stars (near full graphs), and less than 3 or 1 edges and less than 1 two-star (near empty graphs).

far more density on the degenerate regions than any other choice, with almost half (0.432) of its density on near-full graphs and a substantial portion (0.286) on near-empty graphs. The uniform prior alleviates some of normal prior's focus on degenerate ERGM parameters, but still has 0.346 of its density on near full graphs and 0.145 of its density in near empty graphs.

In comparison, the non-degeneracy prior density does not tend to concentrate on the degenerate regions as expected by its definition. Even the weakest setting of $p = 1$, corresponding to a single non-degenerate observed network, already improves over the normal and uniform priors by placing approximately no density in the near-full graph region and only 0.0522 in the near-empty graph region. As expected, these values decrease asymptotically to zero as p increases. However, the $p = 1$ case shows only moderate $P(g(Y) \in \text{int}(\mathcal{C}))$ of 0.398. As seen in Figure 3.2, the non-degeneracy prior under this hyperparameter puts significant

density along the bottom boundary of the convex hull without putting significance in the near empty or full regions. Naturally, this effect is altered by the increase of hyperparameter p, such that $P(g(Y) \in \text{int}(\mathcal{C}))$ reaches 0.915 for $p = 20$.

The results so far show that the non-degeneracy prior indeed puts density away from the degeneracy and convex hull regions especially as the assumption of non-degeneracy strengthens in its hyperparameter specification. More importantly, the commonly used normal and uniform priors put the bulk of their density in the degenerate, full and empty graph, regions. To assume prior information that the network is degenerate is almost certainly nonsensical and unintended by any practitioners, as degenerate networks are almost never observed in nature. In the following, the effects of a prior emphasizing degenerate regions on the posterior are explored in contrast to the posterior for the non-degeneracy prior.

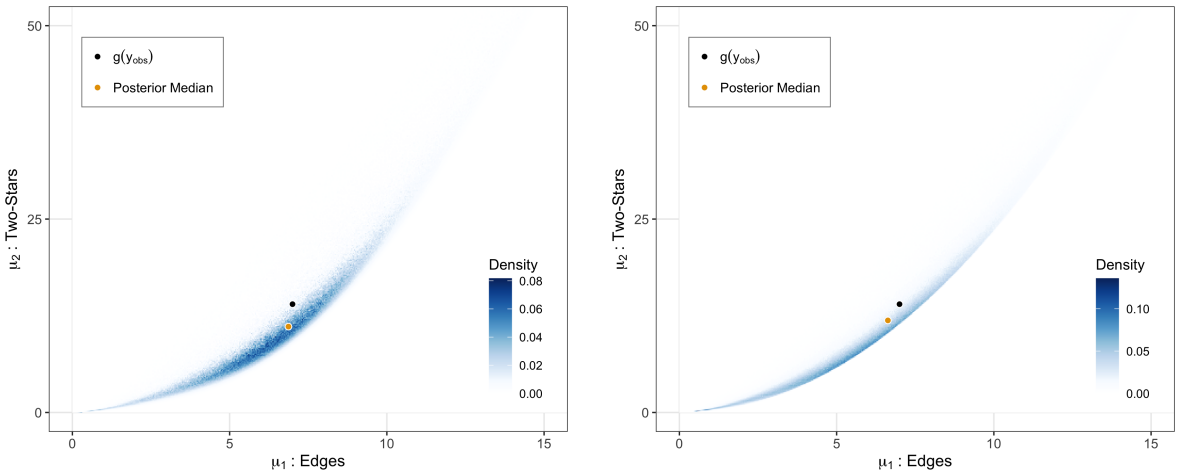
3.5.2 Compare Posteriors

Next, we explore the effect of ERGM prior choice on the posterior. Recall that the goal of this study is to identify a suitable default prior for the ERGM class. That is, a favorable choice should concentrate posterior density on parameter values that generate realistic networks, which can be measured by $P(g(Y) \in \text{int}(\mathcal{C}))$ as defined in Equation 3.21 and applied here to the posterior, and produce accurate predictions of the observed sufficient statistics, measured by predictive MSE, the mean squared error between posterior predictive sufficient statistics and the observed network's statistics. More subtly, the posterior should remain close enough to the prior that the prior's information is meaningfully incorporated rather than overwhelmed by the likelihood. We measure this using the KL divergence $D_{\text{KL}}(\pi \| p)$ between the prior and posterior. The interpretation for this quantity is informative, but varies depending on context. For example, a small value can indicate that the prior was well-placed and the data agreed with it, but it can also indicate that a strong prior overrode the likelihood. We therefore use $D_{\text{KL}}(\pi \| p)$ as a description of how much the posterior departs from the prior rather than as a quality criterion on its own, and read it alongside the

other columns. In the following, the non-degeneracy prior is compared against the standard normal and uniform priors on these criteria, alongside posterior location summaries (medians and means).

Two combinations of observed statistics are examined, split between this section and the next. In the first setting, the posterior is created for an observed network with 7 edges and 14 two-stars, an observation that is somewhat in the center of all observed statistics combinations. In Figure 3.6 the exactly computed model likelihood is combined with the normal and uniform priors in Figure 3.4 to produce posterior distributions, displayed alongside the posterior median and the observed $g(y_{\text{obs}}) = (7, 14)$. All posterior plots in this section and the next are zoomed to $\mu_1 \in [0, 15]$ and $\mu_2 \in [0, 50]$, where the posterior density is concentrated. The normal prior produces a posterior whose density sits noticeably below $g(y_{\text{obs}})$, with posterior median $(6.86, 11.10)$ pulled away from the observation on both coordinates. This pull is reflective of the interaction between the normal prior's density and the likelihood. That is, Table 3.1 in Section 3.5.1 shows the $\mathcal{N}(0, 100\, I_2)$ prior places 0.286 of its density on the near-empty graph region and 0.432 on the near-full graph region. The near-full density has essentially no effect on the posterior, given that the likelihood at $g(y_{\text{obs}}) = (7, 14)$ is far away from and assigns negligible weight to parameter values that generate near-full graphs. By contrast, the near-empty density is close enough to the region where the likelihood is concentrated to exert influence on the posterior, pulling its median toward the empty graph region of the μ-space. In this asymmetric pull, the edge coordinate moves only slightly, from 7 to 6.86, while the two-star coordinate drops from 14 to 11.10. This reflects that the near-empty graph region lies further from $(7, 14)$ along the two-star axis than along the edge axis, such that a prior concentrated near the origin of the μ-space pulls more strongly on μ_2 than on μ_1. Hence, the cost of using $\mathcal{N}(0, 100\, I_2)$ is a concrete bias in the posterior median, which is directly visible in Figure 3.6(a).

The uniform prior's posterior is given in Figure 3.6(b), with its density strictly bounded within a subset of the mean-value parameter space that reflects the prior's own boundedness



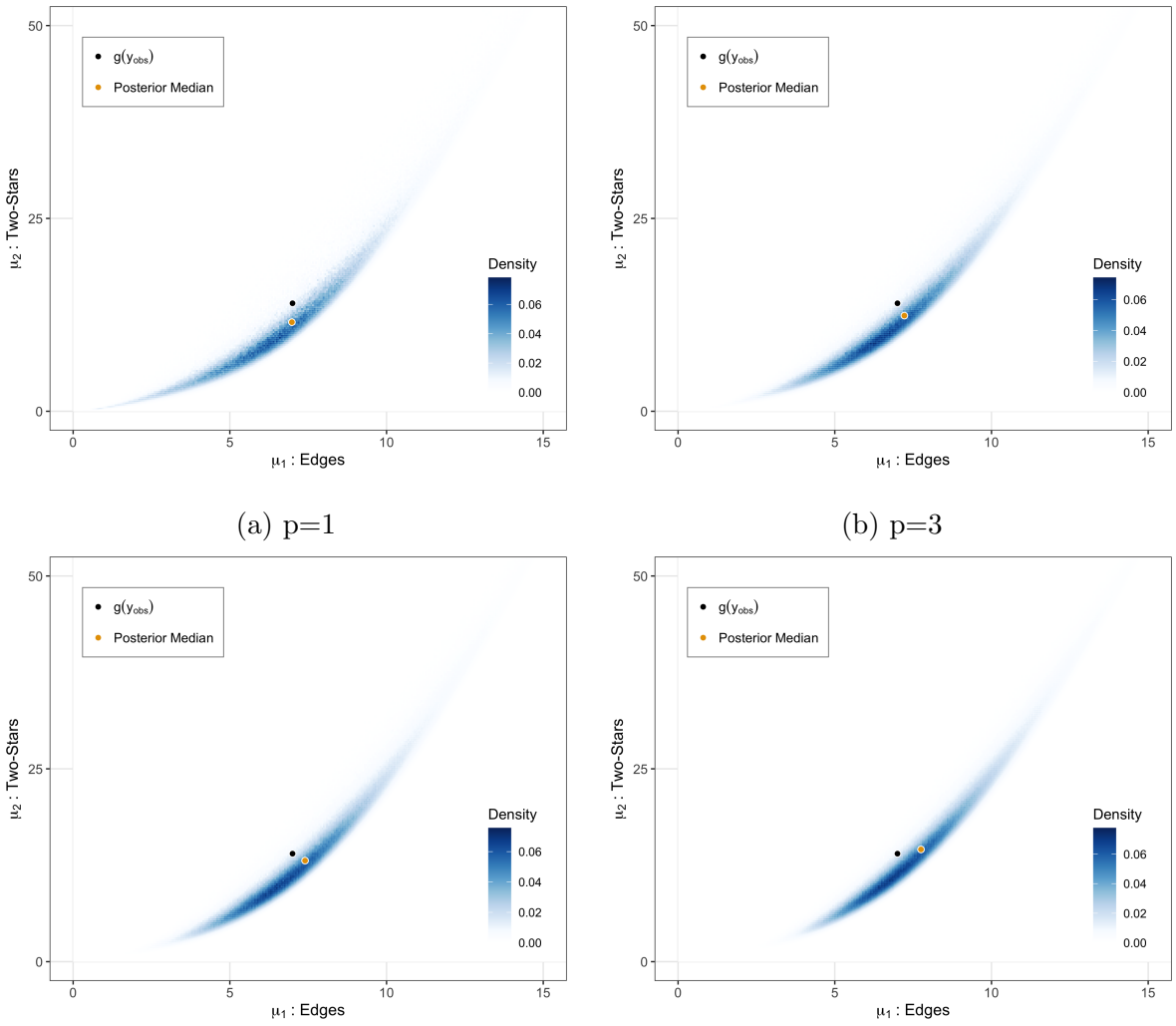
(a) Normal

(b) Uniform

Figure 3.6: Posterior distributions in the μ-space using normal and uniform priors.

on $[-4, 2] \times [-0.05, 1]$ in θ-space. The posterior median $(6.62, 11.90)$ is also pulled below $g(y_{\text{obs}})$, similar to the normal case but with a different asymmetry. That is, the edge coordinate shifts further here, from 7 to 6.62, while the two-star coordinate shifts less, from 14 to 11.90. Table 3.1 shows that the uniform prior places less density on the degenerate regions than the normal prior, with 0.145 on near-empty graphs compared to 0.286 and 0.346 on near-full graphs compared to 0.432.

The non-degeneracy prior with $p \in \{1, 3, 5, 10\}$ is used to produce the posterior distributions in Figure 3.7. Across these four cases, the posterior median approaches $g(y_{\text{obs}}) = (7, 14)$ as p increases from 1 to 10, with $p = 10$ placing the median closest to the observed statistics. Note here that visual inspection of Figure 3.7 can be misleading given that the plots are zoomed with μ_1 spanning 15 units while μ_2 spans 50 units. For larger values of $p \in \{20, 50, 100\}$, the posterior median increases asymptotically to around 10 edges and 25 two-stars, at which point the prior clearly overpowers the likelihood to pull the posterior center closer to the middle region of the μ-space. As seen in this practice, the choice of p therefore controls a tradeoff between how strongly the prior excludes degenerate regions and how much it overrides the likelihood's information about the observed statistics. For the



(a) p=1

(b) p=3

(c) p=5

(d) p=10

Figure 3.7: Posterior distributions in the μ-space using the non-degeneracy prior, under strength hyperparameters $p \in \{1, 3, 5, 10\}$ and $q = 0$.

non-degeneracy prior to balance the belief that the network formation process is not degenerate without this belief dominating the likelihood, setting $p \in [5, 10]$ would be seemingly reasonable for the observed network considered here.

The comparison across prior choices is difficult to capture from the posterior figures alone, and Table 3.2 summarizes the criteria introduced at the beginning of this section to form

Prior Choice	$\boldsymbol{P(g(Y) \in \text{int}(\mathcal{C}))}$	**PMSE**	$\mathbf{D_{KL}}(\pi\|p)$	**Posterior Median**		**Posterior Mean**	
				μ_1	μ_2	μ_1	μ_2
Normal	0.862	(6.857, 127.825)	2.956	6.86	11.10	7.055	14.009
Uniform	0.838	(10.094, 195.984)	1.766	6.62	11.90	6.982	15.793
NDG, $p = 1$	0.874	(6.205, 114.653)	3.057	6.98	11.55	7.179	14.291
NDG, $p = 1.5$	0.881	(5.953, 110.387)	2.624	7.05	11.80	7.269	14.514
NDG, $p = 2$	0.887	(5.724, 105.665)	2.334	7.11	12.00	7.332	14.653
NDG, $p = 3$	0.896	(5.459, 100.343)	1.949	7.22	12.40	7.448	14.954
NDG, $p = 5$	0.907	(5.217, 94.833)	1.514	7.4	13.1	7.635	15.517
NDG, $p = 10$	0.923	(5.187, 91.717)	1.039	7.75	14.55	7.978	16.708
NDG, $p = 20$	0.938	(5.680, 96.324)	0.703	8.24	16.70	8.430	18.476

Table 3.2: Comparison of posterior distributions in the μ-space for a network with observed statistics $g(y_{obs}) = (7, 14)$ using the normal prior, uniform prior, and the non-degeneracy prior under strength hyperparameters $p \in \{1, 1.5, 2, 3, 5, 10, 20, 50, 100\}$ and $q = 0$.

more precise findings. The non-degeneracy prior improves on the normal and uniform priors across every criterion. Even at $p = 1$, the weakest strength setting, $P(g(Y) \in \text{int}(\mathcal{C}))$ reaches 0.874, which marginally exceeds both the normal prior's 0.862 and the uniform prior's 0.838. The predictive MSE shows a sharper gap. The normal prior produces $(6.857, 127.825)$ and the uniform prior produces a much larger $(10.094, 195.984)$, while every non-degeneracy prior row from $p = 1$ through $p = 20$ has smaller PMSE on both components. That is, the prior shape observed in Section 3.5.1, which concentrates density away from degenerate regions, is reflected in tighter predictive fit to the observed sufficient statistics once the posterior is formed.

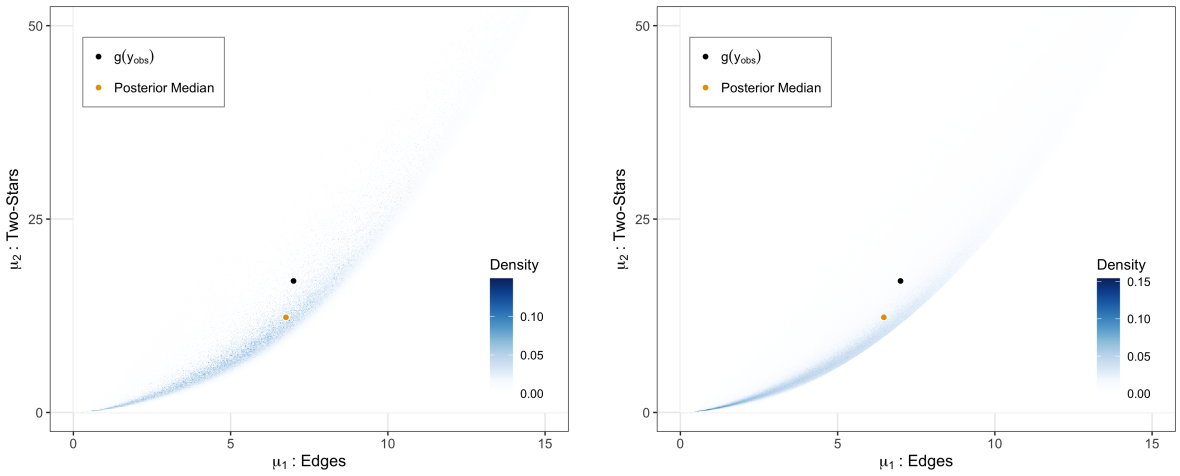
Within the non-degeneracy rows, the PMSE decreases monotonically from $p = 1$ to $p = 10$ and then turns back upward at $p = 20$. Hence, $p = 10$ minimizes PMSE among the values tested. The posterior median follows the same pattern, moving from $(6.98, 11.55)$ at $p = 1$ toward $(7.75, 14.55)$ at $p = 10$ and then past the observation to $(8.24, 16.70)$ at $p = 20$. The posterior mean tracks the same trajectory throughout, sitting slightly above

the median in each row. This pattern however is not reflected in $D_{\mathrm{KL}}(\pi\|p)$, which decreases monotonically across the non-degeneracy rows, from 3.057 at $p = 1$ to 0.703 at $p = 20$. This makes sense by the definition of the non-degeneracy prior. As noted at the beginning of this section, this quantity is informative but ambiguous, given that a small value can reflect either a well-placed prior or a prior that has dominated the likelihood. Read alongside the posterior median drift at $p = 20$, the decreasing KL at larger p is better interpreted as the prior beginning to override the likelihood rather than as agreement with the data. The $P(g(Y) \in \mathrm{int}(\mathcal{C}))$ column, meanwhile, increases monotonically through the non-degeneracy rows and would favor arbitrarily large p on its own, which is why we do not read it in isolation. Hence, these metrics corroborate the $p \in [5, 10]$ range identified from the figures, with PMSE and posterior median accuracy both pointing to the upper end of the range.

3.5.3 Maximally Clustered Network Posterior Comparison, $g(y_{\mathbf{obs}}) = (7, 17)$

In the second example of observed network statistics, the observed network $g(y_{\mathrm{obs}}) = (7, 17)$ has the same edge count as the previous $(7, 14)$ case but the maximum number of two-stars achievable under that edge count. That is, among all networks on $n = 7$ nodes with exactly 7 edges, the largest two-star count is 17. This configuration arises when one vertex is connected to all six others and a single additional edge connects two of the leaves. Unlike the previous case, this maximally clustered observation in the μ-space creates a more demanding test for the posterior.

Figure 3.8 shows the posteriors under the normal and uniform priors combined with the likelihood at $g(y_{\mathrm{obs}}) = (7, 17)$. The normal prior produces a posterior median of $(6.70, 12.05)$, pulled far below the observation on both coordinates and especially on the two-star axis, where the drop from 17 to 12.05 is substantially larger than the drop from 14 to 11.10 seen in the previous case. This reflects the same behavior as before, where the near-empty density of the normal prior imposes influence on the posterior. The effect is more apparent here, with a two-star drop of 4.95 compared to 2.90 in the previous case. This likely reflects that



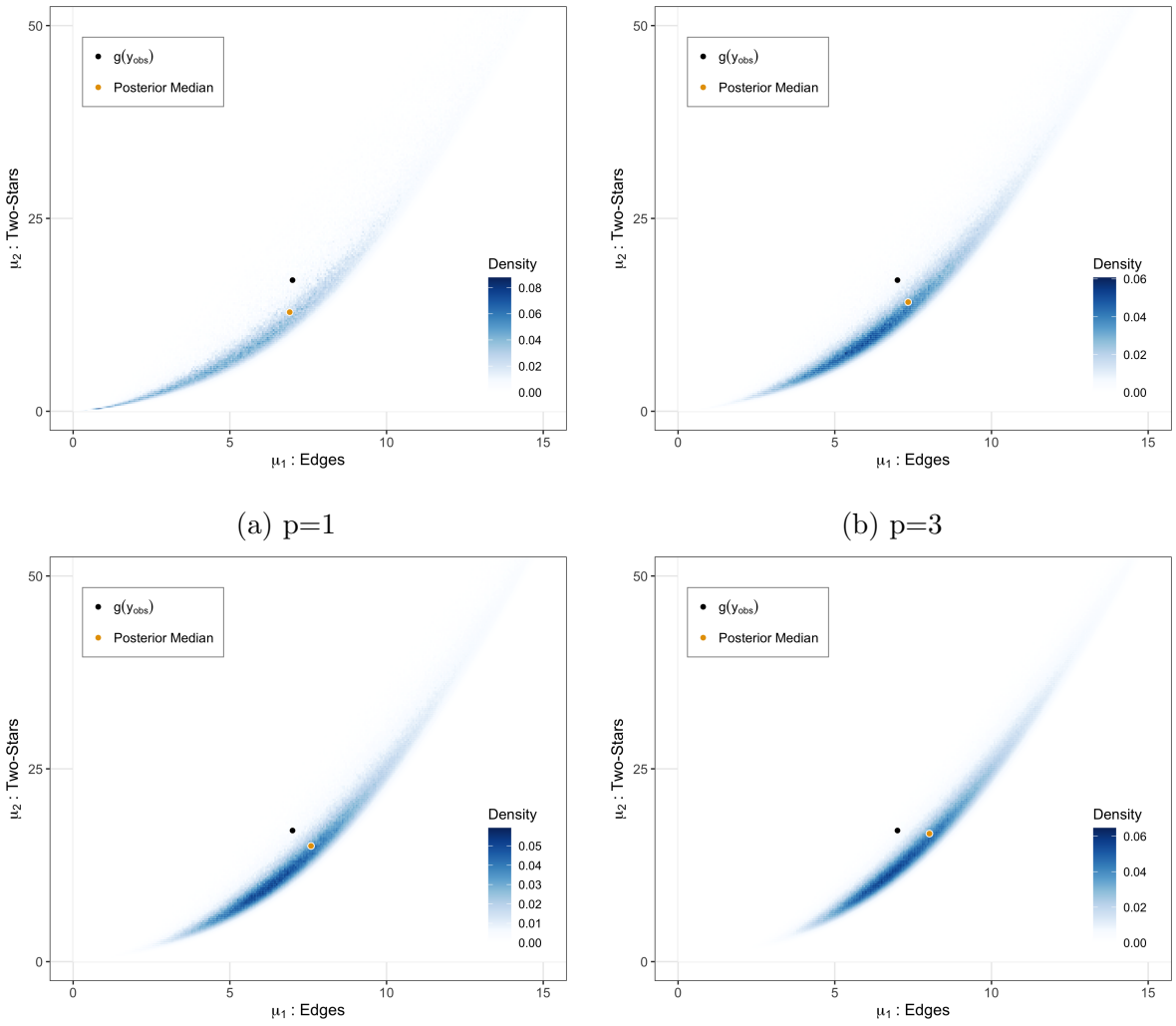
(a) Normal (b) Uniform

Figure 3.8: Posterior distributions in the μ-space using normal and uniform priors for a network with observed statistics $g(y_{obs}) = (7, 17)$.

the likelihood at $(7, 17)$ places less of its density near the belly of the convex hull, leaving the near-empty region of the prior with more agency to shape the posterior. The uniform prior's posterior median $(6.41, 12.10)$ shows the same pattern, pulled below $g(y_{\text{obs}})$ in a similar manner.

The non-degeneracy prior with $p \in \{1, 3, 5, 10\}$ produces the posteriors in Figure 3.9. As in the previous case, the posterior median approaches $g(y_{\text{obs}}) = (7, 17)$ as p increases, moving from $(6.91, 12.85)$ at $p = 1$ toward $(8.02, 16.60)$ at $p = 10$. At $p = 20$ the median drifts to $(8.51, 18.55)$, past the observation on both coordinates. The same behavior seen in the $(7, 14)$ case where the value of p is set too high therefore holds here, with the posterior tracking the observation most closely at moderate p before the prior's influence begins to dominate at larger values. Overall, the non-degeneracy posteriors seem to perform very well compared to using normal and uniform priors, especially for $p \in [5, 10]$.

Table 3.3 summarizes the same criteria used in Section 3.5.2 for the $(7, 17)$ observation. The non-degeneracy prior improves on the normal and uniform priors across every criterion



(a) p=1

(b) p=3

(c) p=5

(d) p=10

Figure 3.9: Posterior distributions in the μ-space using the non-degeneracy prior for a network with observed statistics $g(y_{obs}) = (7, 17)$, under strength hyperparameters $p \in \{1, 3, 5, 10\}$ and $q = 0$.

considered. At the weakest specification $p = 1$, $P(g(Y) \in \text{int}(\mathcal{C}))$ reaches 0.841 and exceeds both the normal prior's 0.803 and the uniform prior's 0.770. The predictive MSE gap is wider than in the previous case. The normal prior produces $(12.373, 259.623)$ and the uniform prior $(14.202, 300.265)$, while every non-degeneracy row through $p = 10$ is smaller on both

components.

Prior Choice	$P(g(Y) \in \text{int}(\mathcal{C}))$	**PMSE**	$D_{KL}(\pi\|p)$	**Posterior Median**		**Posterior Mean**	
				μ_1	μ_2	μ_1	μ_2
Normal	0.803	(12.373, 259.623)	3.496	6.70	12.05	7.081	16.978
Uniform	0.770	(14.202, 300.265)	1.756	6.41	12.10	6.947	17.651
NDG, $p = 1$	0.841	(10.070, 208.498)	3.702	6.91	12.85	7.284	17.086
NDG, $p = 1.5$	0.857	(9.377, 192.693)	3.273	7.06	13.30	7.427	17.329
NDG, $p = 2$	0.869	(8.762, 177.483)	2.980	7.16	13.60	7.516	17.400
NDG, $p = 3$	0.886	(8.065, 159.533)	2.574	7.34	14.15	7.675	17.641
NDG, $p = 5$	0.905	(7.334, 138.232)	2.086	7.59	15.00	7.900	18.078
NDG, $p = 10$	0.926	(6.816, 115.518)	1.497	8.02	16.60	8.270	19.025
NDG, $p = 20$	0.942	(6.938, 101.260)	1.024	8.51	18.55	8.693	20.379
NDG, $p = 50$	0.957	(8.050, 95.725)	0.573	9.23	21.75	9.311	22.722
NDG, $p = 100$	0.963	(9.307, 99.980)	0.342	9.71	24.05	9.741	24.528

Table 3.3: Comparison of posterior distributions in the μ-space for a maximally clustered network with observed statistics $g(y_{obs}) = (7, 17)$ using the normal prior, uniform prior, and the non-degeneracy prior under strength hyperparameters $p \in \{1, 1.5, 2, 3, 5, 10, 20, 50, 100\}$ and $q = 0$.

For the non-degeneracy rows, the edge term PMSE decreases from $p = 1$ to $p = 10$ and turns upward at $p = 20$, matching the pattern seen in the previous example in Table 3.2. The two-star component behaves differently here. It continues to decrease through $p = 20$ and bottoming at $p = 50$ before turning upward at $p = 100$. As a standalone, the two-star PMSE would appear to favor larger p. However, the posterior median reaches its closest approach to $g(y_{\text{obs}}) = (7, 17)$ at $p = 10$ with $(8.02, 16.60)$, and drifts past the observation to $(8.51, 18.55)$ at $p = 20$, $(9.23, 21.75)$ at $p = 50$, and $(9.71, 24.05)$ at $p = 100$. The posterior mean tracks the same trajectory, sitting slightly above the median in each row. $D_{KL}(\pi\|p)$ decreases monotonically across the non-degeneracy rows, from 3.702 at $p = 1$ to 0.342 at $p = 100$. Read alongside the median overshoot at $p = 20$ and beyond, the decreasing KL

at larger p is again better interpreted as the prior's tendency dominate the likelihood rather than as agreement with the data as our prior belief strengthens in p. Hence, median accuracy and the overshoot diagnostic support the same $p \in [5, 10]$ range identified in Section 3.5.2.

Across both observations $g(y_{\text{obs}}) = (7, 14)$ and the maximally clustered $g(y_{\text{obs}}) = (7, 17)$, the non-degeneracy prior improves on the normal and uniform priors on every criterion ($P(g(Y) \in \text{int}(\mathcal{C}))$, PMSE, $D_{KL}(\pi \| p)$, median, and mean) in Tables 3.2 and 3.3, and the $p \in [5, 10]$ range identified in Section 3.5.2 continues to hold for the $(7, 17)$ case. Notably, the criterion that shifts most across the two observations is two-star PMSE, which bottoms at $p = 10$ for $(7, 14)$ but only at $p = 50$ for $(7, 17)$. Hence, the posterior median and the overshoot diagnostic provide a more stable basis for the recommended range than PMSE alone. The results in this section rely on almost exact computation of the non-degeneracy prior, which is feasible for $n = 7$ because all 2^{21} graph configurations can be enumerated. For networks of practical size this enumeration is unavailable, and the prior must instead be approximated using the importance-reweighting estimator developed in Section 3.4.4 alongside noisy Hamiltonian Monte Carlo for posterior sampling. Before applying this approximation to a larger empirical network in Section 3.7, we first revisit the $n = 7$ setting in Section 3.6 using noisy HMC under the approximated non-degeneracy prior. This step validates that the exact posteriors computed here provide a ground truth against which the approximate sampler can be compared, given that any discrepancy between the two reflects approximation error rather than a difference in the underlying prior.

3.6 Approximate Non-degeneracy Prior Using Noisy HMC

The exact non-degeneracy prior in Section 3.5 relies on enumeration of all $2^{\binom{n}{2}}$ graph configurations, which is available only for small n. For networks of realistic size, the prior must be approximated using the importance-reweighting estimator detailed in Section 3.4.4 and paired with a posterior sampler that handles the ERGM's intractable normalizing constant. We use noisy HMC (SBF19) to naturally integrate with the prior approximation since both rely on the same auxiliary network draws at each leapfrog step.

The focus of this section is not noisy HMC itself, whose behavior for doubly intractable posteriors is studied in detail by (SBF19). The sampler is used here as a practical tool, and the purpose of this section is to confirm that the prior comparisons established in Section 3.5 continue to hold when the non-degeneracy prior is approximated rather than computed exactly. We return to the same $n = 7$ setting with observed statistics $g(y_{\text{obs}}) = (7, 14)$, where the exact posterior in Table 3.2 provides a ground truth against which the approximate sampler can be directly compared. In this setup, any difference between the exact and approximate posteriors under the non-degeneracy prior is reflective of approximation error from Equation 3.19 and the sampler, rather than a difference in the underlying prior used.

Implementing the importance-reweighting approximation from Equation 3.19 is straightforward within the noisy HMC framework. At each leapfrog step ℓ, noisy HMC already draws auxiliary networks $\{x_1, \ldots, x_N\} \sim p(\cdot \mid \theta_\ell)$ for the gradient estimate in Equation 3.5 and the leapfrog ratio in Equation 3.8. Setting $\widetilde{\theta} = \theta_\ell$, the same draws are reused to evaluate $\widehat{P}_\theta(g(Y) \in \text{int}(\mathcal{C}))$ and its gradient. Sampling of auxiliary networks is by far the most computationally expensive component of each iteration, so reusing these samples for the prior practically nullifies the additional cost needed by the non-degeneracy prior. Evaluating the indicator $\mathbb{1}\{g(x_i) \in \text{int}(\mathcal{C})\}$ in Equation 3.19 requires the convex hull $\mathcal{C}$. Rather than pre-computing $\mathcal{C}$ from full graph enumeration, the hull is built adaptively. It is initially approximated from $100,000$ networks simulated at the MPLE, then grows during sampling

as each auxiliary network's sufficient statistics are added. The hull only ever expands, and for the $n = 7$ case with 144 distinct statistic combinations it converges to the true hull within the early iterations of the chain. The same $100,000$ networks that seed the hull also provide the initial importance-sampling pool for the prior, which is replaced by the leapfrog auxiliary draws once sampling begins. This warm-start stores only the sufficient statistic pairs, not the reference parameter $\widetilde{\theta}$. The importance-reweighting estimator in Equation 3.19 sets $\widetilde{\theta} = \theta_\ell$ fresh at each leapfrog step, so no fixed reference is needed. For efficiency, auxiliary draws use a custom tie-no-tie (TNT) sampler that bypasses the computational overhead of the `ergm` package. Tuning follows the recommendations of (SBF19), where the step size ϵ and number of leapfrog steps L are calibrated via dual averaging over the first $1,000$ iterations to target the HMC-optimal acceptance probability $\delta = 0.651$ derived in (BPR13) also adopted by (SBF19), with fixed integration time $t = \epsilon L = 2.38/\sqrt{d}$. This choice of t matches the proposal scaling of the exchange algorithm at optimal tuning. The sensitivity of the approximation to δ is explored in Section 3.6.1.

The mass matrix M is set to an estimate of the inverse posterior covariance $\widehat{\Sigma}^{-1}$, also following the procedure in (SBF19). The posterior mode is first determined via a Robbins-Monro scheme (RM51),

$$\theta^{(i+1)} = \theta^{(i)} + \frac{\alpha}{i}\,\widehat{\nabla}_\theta \log \pi(\theta^{(i)} \mid y), \tag{3.22}$$

where $\widehat{\nabla}_\theta \log \pi(\theta^{(i)} \mid y)$ is the noisy gradient estimate from Equation 3.5 evaluated at $\theta^{(i)}$ using $N = 10$ auxiliary draws, and $\alpha = 1$ is the step size. The scheme is run for 200 iterations, with the mode estimate taken as $\theta^* = \theta^{(201)}$. An additional 500 auxiliary draws $\{u^{(1,\theta^*)}, \ldots, u^{(500,\theta^*)}\}$ are then simulated from $p(\cdot \mid \theta^*)$, and the Hessian of the log posterior at θ^* is approximated as

$$\widehat{\Sigma}^{-1} = \operatorname{cov}\big[g(u^{(1,\theta^*)}), \ldots, g(u^{(500,\theta^*)})\big] + \nabla^2_\theta \log p(\theta^*), \tag{3.23}$$

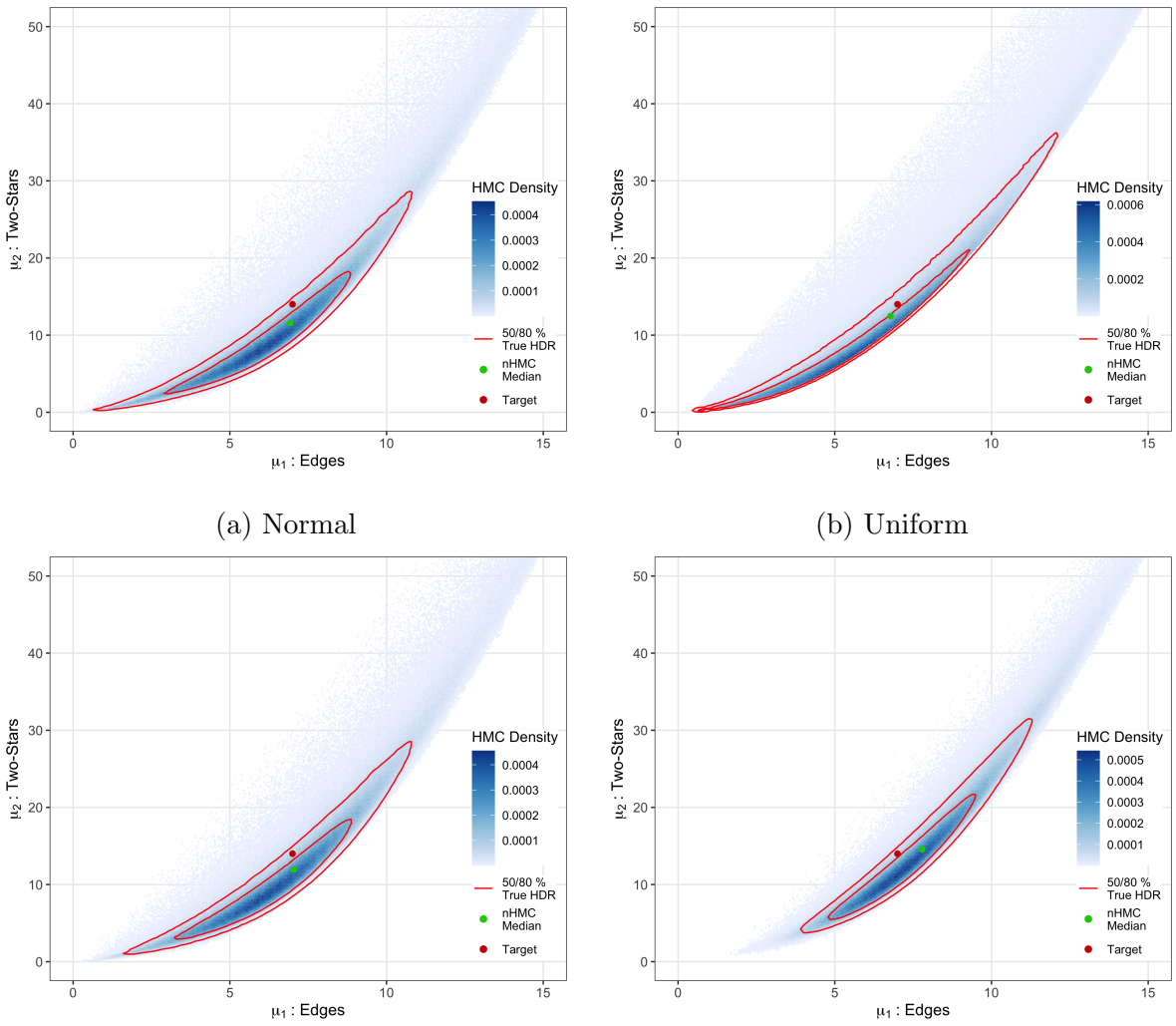
which uses the identity $\nabla^2_\theta \log Z(\theta) = \operatorname{cov}_\theta[g(Y)]$ specified to the ERGM, where $\theta^\top g(y)$ is linear in θ and thus contributes no data-dependent term to the Hessian. The mass matrix is then set to $M = \widehat{\Sigma}^{-1}$.

To start the sampler in a reasonable region of θ-space, the chain is initialized at the frequentist maximum pseudolikelihood estimate (DHG09). Across the experiments reported in this section, each chain is run for $100,000$ iterations after burn-in (which is also the dual averaging iterations) with $N = 10$ auxiliary draws per leapfrog step. Given the small network size, we can afford a much larger chain compared to the $5,000$ iterations reported in (SBF19) for a network with 34 nodes. As in that study, 20 separate chains were run, each initialized at the MPLE plus random Gaussian noise.

With the sampler in place, the main validation concern is whether the approximate non-degeneracy prior used within noisy HMC reproduces the posterior obtained under the exactly computed prior in Section 3.5.2. Table 3.4 reports posterior summaries for the approximate setup across the same prior specifications examined in Table 3.2.

The approximate posterior medians follow the exact medians closely across all prior choices, with the overlay of exact and approximate contours shown in Figure 3.10 for the normal, uniform, and non-degeneracy priors at $p \in \{1, 10\}$. At $p = 1$, the approximate median is $(7.031, 11.923)$ compared to the exact median $(6.98, 11.55)$ in Table 3.2. The discrepancy of 0.05 on the edge coordinate and 0.37 on the two-star coordinate is small relative to the scale of the posterior. The same behavior holds at stronger prior strengths, with the approximate median at $p = 10$ of $(7.784, 14.584)$ matching the exact median $(7.75, 14.55)$ to within 0.04 on each statistic. Posterior means show similar agreement, and PMSE values track the exact results with the approximate PMSE decreasing monotonically from $p = 1$ to $p = 10$ as in the exact case.

The $D_{\mathrm{KL}}(p\|p_{nHMC})$ column in Table 3.4 provides a direct measure of how closely the noisy HMC posterior approximates the exact posterior computed in Section 3.5.2. This quantity is different in purpose from the $D_{\mathrm{KL}}(\pi\|p)$ reported in Tables 3.2 and 3.3, which measures how far the posterior departs from the prior and therefore captures the influence of the likelihood relative to the prior. The $D_{\mathrm{KL}}(p\|p_{nHMC})$ reported here instead measures the accuracy of the posterior approximation, that is, the cost of replacing exact computation



(a) Normal

(b) Uniform

(c) p=1

(d) p=10

Figure 3.10: Overlay of the true posterior distributions (50% and 80% highest density region; red contours) over noisy HMC estimated posterior distributions in the μ-space for a network with observed statistics $g(y_{obs}) = (7, 14)$, under the normal, uniform, and non-degeneracy priors with varying strength hyperparameters. Points identified are the target $g(y_{obs})$ and the noisy HMC posterior medians.

with the importance-reweighting estimator in Equation 3.19 and the noisy HMC sampler. These values are small and stable across all prior choices, ranging from 0.333 at $p = 20$ to 0.442 for the uniform prior. The narrow range indicates that the quality of the approximation does not depend on which prior is used, and that the noisy HMC sampler recovers each exact posterior to a similar degree. Since the exact posteriors in Section 3.5.2 use the true hull from full enumeration while the approximate posteriors here use the adaptively estimated hull, this agreement also confirms that the adaptive hull converges adequately for the $n = 7$ setting. The $p \in [5, 10]$ range identified in Section 3.5.2 therefore continues to hold under the approximate setup.

Prior Choice	$\boldsymbol{P(g(Y) \in \text{int}(\mathcal{C}))}$	**PMSE**	$\mathbf{D_{KL}}(p\|p_{nHMC})$	**Posterior Median**		**Posterior Mean**	
				μ_1	μ_2	μ_1	μ_2
Normal	0.850	(10.347, 267.379)	0.424	6.929	11.560	7.398	16.310
Uniform	0.833	(12.153, 289.772)	0.442	6.779	12.479	7.287	17.489
NDG, $p = 1$	0.866	(8.441, 197.274)	0.416	7.031	11.923	7.410	15.913
NDG, $p = 1.5$	0.874	(8.048, 185.320)	0.421	7.124	12.231	7.486	16.035
NDG, $p = 2$	0.881	(7.626, 171.998)	0.409	7.193	12.449	7.538	16.068
NDG, $p = 3$	0.889	(7.296, 161.658)	0.409	7.321	12.903	7.654	16.337
NDG, $p = 5$	0.905	(6.468, 134.188)	0.372	7.535	13.670	7.826	16.621
NDG, $p = 10$	0.921	(5.508, 99.748)	0.351	7.784	14.584	8.005	16.850
NDG, $p = 20$	0.936	(5.713, 98.409)	0.333	8.190	16.391	8.388	18.295

Table 3.4: Comparison of noisy HMC approximated posterior distributions in the μ-space for a network with observed statistics $g(y_{obs}) = (7, 14)$ using the normal prior, uniform prior, and the non-degeneracy prior under varying strength hyperparameters.

Sampler behavior is comparable across priors in this validation. The approximate posteriors match the exact posteriors of Section 3.5.2 for the non-degeneracy case, and identical sampler configuration is used for the normal and uniform priors. Differences between approximate posteriors across the different priors therefore reflect prior effects rather than sampler artifacts.

Having established that the approximate non-degeneracy posterior matches the exact posterior, we now repeat the between-prior comparison of Section 3.5.2 using noisy HMC throughout. The pattern seen under exact computation carries through under the approximate sampler. The non-degeneracy prior improves on the normal and uniform priors on PMSE at every strength setting through $p = 10$. The approximate PMSE for the normal prior is $(10.347, 267.379)$ and for the uniform prior is $(12.153, 289.772)$, both substantially worse than any non-degeneracy row in Table 3.4. At $p = 20$, the edge component of PMSE moves up from 5.508 to 5.713, consistent with the same increase observed in the exact case in Table 3.2 and reflecting the prior beginning to override the likelihood. Posterior medians under the non-degeneracy prior approach $g(y_{\text{obs}}) = (7, 14)$ as p increases from 1 to 10, whereas the normal and uniform priors produce medians $(6.929, 11.560)$ and $(6.779, 12.479)$, respectively, that are pulled below the observation on both coordinates in the same asymmetric manner identified in Section 3.5.2. The predictive non-degeneracy $P(g(Y) \in \text{int}(\mathcal{C}))$ increases monotonically across the non-degeneracy rows, from 0.866 at $p = 1$ to 0.936 at $p = 20$, and exceeds both the normal prior's 0.850 and the uniform prior's 0.833 at every strength setting. This matches the ordering observed under exact computation, where the non-degeneracy prior consistently placed more density in the non-degenerate region than either standard choice. The $p \in [5, 10]$ range continues to provide the best balance between excluding degenerate regions and allowing the likelihood to shape the posterior. The findings of Section 3.5.2 therefore do not rely on enumeration of all 2^{21} graph configurations. The approximate non-degeneracy prior, computed from auxiliary draws already generated by noisy HMC, is sufficient in arriving at the same comparative conclusions.

While the qualitative conclusions remain, the approximate posteriors are systematically more dispersed than their exact counterparts. PMSE is higher in every row of Table 3.4 compared to Table 3.2, and posterior means are shifted further from $g(y_{\text{obs}})$ on both coordinates. This spread is consistent with the general behavior of noisy MCMC methods, which target an approximate stationary distribution rather than the true posterior (AFE16). Al-

though the $D_{\mathrm{KL}}(p\|p_{nHMC})$ values in Table 3.4 show comparable overall distributions being approximate across priors, the effect on PMSE is uneven. For the normal and uniform priors, the exact posterior concentrates asymmetrically near the degenerate regions of the μ-space, and the approximation noise dilutes this concentration. The exact PMSE of $(6.857, 127.825)$ for the normal prior and $(10.094, 195.984)$ for the uniform both increase substantially under the approximate sampler. For the non-degeneracy prior, the exact posterior already avoids the degenerate regions, leaving less asymmetric concentration for the noise to disrupt. The PMSE increase is correspondingly smaller, from $(5.187, 91.717)$ to $(5.508, 99.748)$ at $p = 10$. That is, the non-degeneracy posterior is more robust to the practical approximation because it does not place density in the near-degenerate tail that causes the broadening under the normal and uniform priors.

For all prior choices examined in this section, dual averaging produces L at or near 1, meaning the sampler operates as a Metropolis-adjusted Langevin algorithm (MALA) rather than a multi-step HMC. This is a consequence of the $n = 7$ posterior being simple enough that a single leapfrog step, even at large ϵ, traverses the posterior without overshooting. For comparison, (SBF19) report $\epsilon \approx 0.003$ and $L = 487$ for the edges and two-stars ERGM on the Zachary karate club with $n = 34$ nodes, where the posterior curvature is far more severe and forces small steps over a long trajectory. The difference is network size rather than model specification, as both settings use the same edges and two-stars parameterization. That is, the 2^{21} graph configurations and 144 distinct statistic combinations of the $n = 7$ case produce a posterior landscape that is mild enough for $L = 1$ to be enough under any reasonable tuning. The sensitivity of the approximation to this tuning is examined in the following.

3.6.1 Tuning Noisy HMC

The results in this section were computed using the tuning recommendations of (SBF19) with the acceptance probability set to the HMC-optimal $\delta = 0.651$ from (BPR13). This choice

controls the tradeoff between the step size ϵ and the number of leapfrog steps L produced by dual averaging, and was adopted as a default without further calibration. Here we examine whether the posterior approximation quality depends on this choice. Table 3.5 reports L, ϵ, and $D_{KL}(p\|p_{nHMC})$ across three acceptance targets: the MALA-optimal rate of 0.574 from (RR98), the HMC-optimal 0.651 from (BPR13) used in the main results, and a higher target of 0.75. Each quantity is averaged over 20 replications of $100,000$ post-burn-in iterations for the normal and uniform priors and for the non-degeneracy prior at $p \in \{1, 5\}$.

Acceptance Probability	**Normal**	**Uniform**	**NDG p = 1**	**NDG p = 5**
$\delta = 0.574$	$L : 1.000$; $\epsilon : 1.961$	$L : 1.350$; $\epsilon : 1.166$	$L : 1.000$; $\epsilon : 1.874$	$L : 1.000$; $\epsilon : 1.351$
	$D_{KL} : 0.462$ (0.023)	$D_{KL} : 0.484$ (0.040)	$D_{KL} : 0.437$ (0.104)	$D_{KL} : 0.435$ (0.017)
$\delta = 0.651$	$L : 1.000$; $\epsilon : 1.729$	$L : 1.350$; $\epsilon : 0.902$	$L : 1.000$; $\epsilon : 1.599$	$L : 1.100$; $\epsilon : 1.000$
	$D_{KL} : 0.424$ (0.020)	$D_{KL} : 0.442$ (0.023)	$D_{KL} : 0.416$ (0.019)	$D_{KL} : 0.372$ (0.016)
$\delta = 0.750$	$L : 1.000$; $\epsilon : 1.351$	$L : 10.900$; $\epsilon : 0.239$	$L : 1.000$; $\epsilon : 1.187$	$L : 4.150$; $\epsilon : 0.386$
	$D_{KL} : 0.369$ (0.017)	$D_{KL} : 0.390$ (0.013)	$D_{KL} : 0.343$ (0.082)	$D_{KL} : 0.331$ (0.008)

Table 3.5: Sensitivity of noisy HMC tuning to the target acceptance probability δ for the $n = 7$ edges and two-stars ERGM with $g(y_{\text{obs}}) = (7, 14)$. Each cell reports the leapfrog steps L and step size ϵ from dual averaging, alongside $D_{KL}(p\|p_{nHMC})$ between the exact and approximate posteriors. Acceptance targets are the MALA-optimal 0.574 (RR98), HMC-optimal 0.651 (BPR13), and 0.75.

The reported L and ϵ values are averages over the 20 replications, which accounts for the fractional L values in the table. Across $\delta = 0.574$ and $\delta = 0.651$, dual averaging produces L at or near 1 for every prior, confirming the MALA regime described in the preceding paragraph. The step size ϵ decreases as δ increases, which is the expected behavior that a stricter acceptance target requires a more conservative step to maintain the target rate. At $\delta = 0.75$, the normal and NDG $p = 1$ priors remain at $L = 1$, but the uniform prior and NDG $p = 5$ shift to $L = 10.9$ and $L = 4.15$, respectively, with correspondingly small

ϵ. For the uniform prior, this is likely caused by the boundedness of the prior in θ-space ($[-4, 2] \times [-0.05, 1]$), which creates hard boundaries in the posterior that force the leapfrog integrator into shorter, more careful steps as the acceptance target tightens. This behavior only appearing at $\delta = 0.75$ and not at $\delta = 0.651$ indicates that the $\delta = 0.651$ target used in the main results of this section does not push the sampler into this territory.

The main finding of this table is the stability of $D_{KL}(p\|p_{nHMC})$ across all configurations. Values range from 0.331 (NDG $p = 5$ at $\delta = 0.75$) to 0.484 (uniform at $\delta = 0.574$), with standard deviations generally below 0.04. Within each prior, D_{KL} decreases modestly as δ increases, consistent with the smaller step sizes producing a more accurate leapfrog integrator and therefore tighter gradient estimates. However, the magnitude of this improvement is small with respect to the baseline D_{KL} values. Even the uniform prior at $\delta = 0.75$, where L jumps from 1.35 to 10.9 and ϵ drops from 1.166 to 0.239, sees D_{KL} decrease only from 0.484 to 0.390. That is, a tenfold increase in L and a fivefold decrease in ϵ produces a D_{KL} reduction of less than 0.1. The approximation quality is not sensitive to the acceptance target for this network size.

These results support the use of $\delta = 0.651$ adopted in the main results of this section. The D_{KL} values at that target are close to those obtained at $\delta = 0.75$ without the sampler configuration instabilities observed for the uniform prior at the higher target. However, the robustness seen here is a property of the $n = 7$ posterior, which is mild enough for even a single large leapfrog step to produce an adequate approximation. The following Section 3.7 applies the same dual averaging approximation setup to the Kapferer tailor shop network. For that network, where $n = 39$ and the parameter space is three-dimensional, the posterior geometry requires the sampler to take more and smaller steps.

3.7 Kapferer Tailor Shop Empirical Study

Having validated the approximate non-degeneracy prior against exact ground truth in Section 3.6, we now apply it to an empirical network where exact computation is unavailable. Kapferer's tailor shop network (Kap72) records friendship and socioemotional interactions among 39 workers in a tailor shop in Zambia, observed over ten months in 1965. The network is undirected with 158 edges, where each edge indicates a reciprocal interactional relationship. This dataset has been analyzed in previous ERGM studies (HHH12; ZL24) and provides a test case at a network size ($n = 39$) where the $2^{\binom{39}{2}} = 2^{741}$ graph configurations cannot be enumerated.

We fit two models to this network. The first uses the geometrically weighted parameterization of (Hun07), whose diminishing-return structure was designed to prevent the model degeneracy that is often problematic with traditional k-star and triangle statistics (Han03; SPR06; Sch11). The model specification already controls degeneracy under this parameterization, so the non-degeneracy prior's additional effect should be small. The second model replaces the geometrically weighted dyadwise shared partner term with the unweighted two-star count, reintroducing a degeneracy-prone statistic. If the non-degeneracy prior improves inference, the improvement should be concentrated in this second model. Together, the two models test whether the non-degeneracy prior helps where the model is vulnerable to degeneracy and whether it interferes where the model is not, in comparison to normal and uniform priors.

For the first model, we adopt the specification of (HHH12), who found that the geometrically weighted degree effect was not significant for this network. The reduced model becomes

$$p_\theta(Y = y) \propto \exp\big\{\theta_1\, e(y) + \theta_2\, v(y;\tau) + \theta_3\, w(y;\tau)\big\}, \tag{3.24}$$

with fixed decay $\tau = 0.25$, where $e(y)$, $v(y;\tau)$, and $w(y;\tau)$ denote the edge count, GWESP, and GWDSP statistics, respectively. The second model replaces $w(y;\tau)$ with the two-star

count $s_2(y) = \sum_i \binom{d_i}{2}$:

$$p_\theta(Y = y) \propto \exp\{\theta_1 \, e(y) + \theta_2 \, s_2(y) + \theta_3 \, v(y; \tau)\}. \tag{3.25}$$

The noisy HMC sampler is configured in the same approach as done in Section 3.6, with each chain run for $100,000$ iterations after burn-in, $N = 10$ auxiliary draws per leapfrog step, and 20 independent replications per prior. The convex hull $\mathcal{C}$ required by the non-degeneracy prior is estimated adaptively in three dimensions using the same warm-start procedure as before, seeded from $100,000$ networks simulated at the MPLE. Dual averaging produces acceptance rates between 0.65 and 0.67, step sizes ϵ between 0.036 and 0.041, and trajectory lengths L between 41 and 46 across both models and all priors. Unlike the $n = 7$ setting in Section 3.6 where $L = 1$ required a sampler using simple MALA proposals, the Kapferer posterior requires long trajectories to capture its parameter space.

For networks of this size, the convex hull cannot be reliably determined from sampling alone, and the metric $P(g(Y) \in \text{int}(\mathcal{C}))$ used in Sections 3.5 and 3.6 is unavailable. We instead evaluate the posteriors using standard goodness-of-fit diagnostics on network features not included in the model (MHH08) and posterior predictive interquartile ranges of the sufficient statistics, reported in Tables 3.6 and 3.7. These diagnostics are computed from networks simulated under posterior draws of θ, which differs from the classical goodness-of-fit framework where networks are simulated from a single point estimate such as the MLE. The two answer different questions about model fit. The posterior predictive form is used throughout, since the goal is to evaluate the posterior rather than a point estimate.

3.7.1 Geometrically Weighted Model

Figure 3.11 and Table 3.6 present the results for the geometrically weighted model. The four goodness-of-fit panels are visually indistinguishable. The degree, edgewise shared partner, and geodesic distance distributions all overlap across priors. Table 3.6 confirms this across all five prior configurations. The predictive median for edges ranges from 168 to 173 across pri-

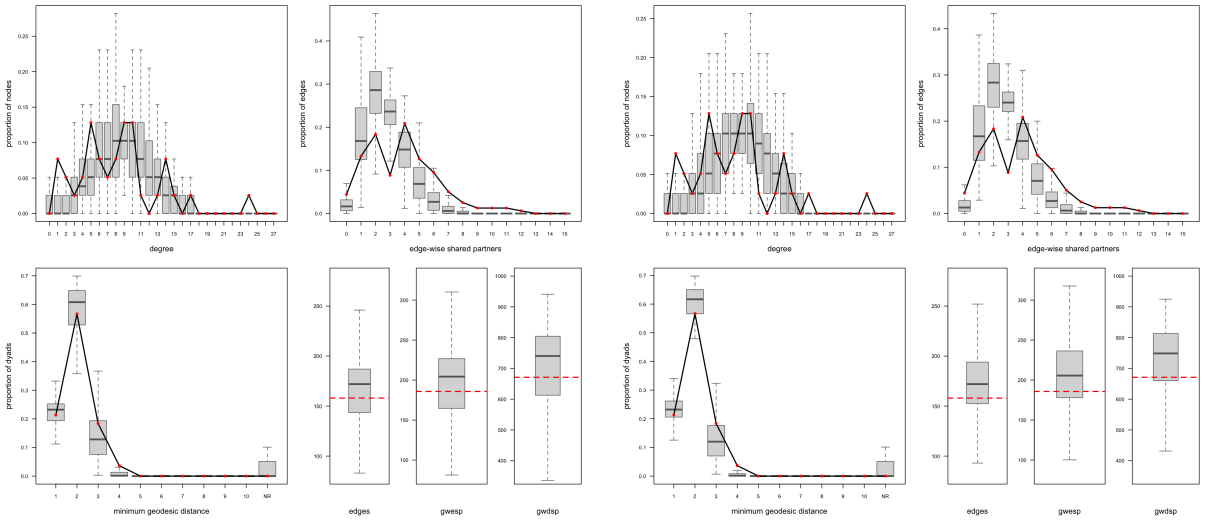
ors, with the observed value of 158 falling within each interquartile range. The GWESP and GWDSP intervals show the same agreement. The geometrically weighted parameterization controls degeneracy through the sufficient statistics, and the prior has little to add. That the non-degeneracy prior produces the same predictive quality as the normal and uniform priors confirms it does not interfere when the model already handles degeneracy.

Prior Choice	μ_1 : **Edges**	μ_2 : **GWESP**	μ_3 : **GWDSP**
Observed	158	185.8	671.1
Normal	172 [144, 187]	204 [165, 227]	740 [615, 804]
Uniform	172 [153, 194]	206 [178, 236]	749 [662, 813]
NDG, $p = 1$	168 [144, 184]	199 [165, 222]	725 [620, 796]
NDG, $p = 5$	169 [147, 188]	199 [170, 227]	727 [646, 802]
NDG, $p = 10$	173 [155, 189]	206 [180, 231]	748 [667, 805]

Table 3.6: Posterior predictive sufficient statistics for the Kapferer tailor shop network under the edges, GWESP(0.25), and GWDSP(0.25) model. Each cell reports the posterior predictive median with interquartile range $[25\%, 75\%]$ computed from 200 networks simulated from the pooled posterior. The observed row gives $g(y_{\text{obs}})$.

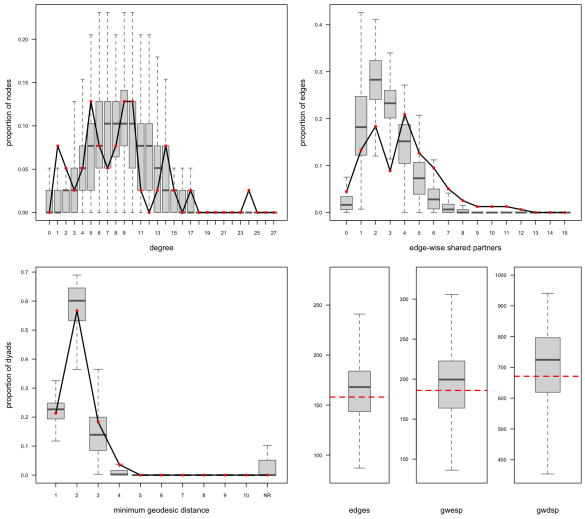
3.7.2 Two-Star Model

Replacing the GWDSP term with the unweighted two-star count introduces a degeneracy-prone statistic, and the differences across priors become noticeable. Under the normal prior, the predictive interquartile range for edges spans $[17, 688]$. That is, the middle 50% of networks simulated from the posterior range from 17 to 688 edges out of a possible 741. The two-star IQR of $[18, 23644]$ and GWESP IQR of $[3, 883]$ are similarly wide. The goodness-of-fit diagnostics in Figure 3.12(a) also reflect that the degree distribution is dominated by isolated nodes and the edgewise shared partner distribution concentrates at zero, consistent

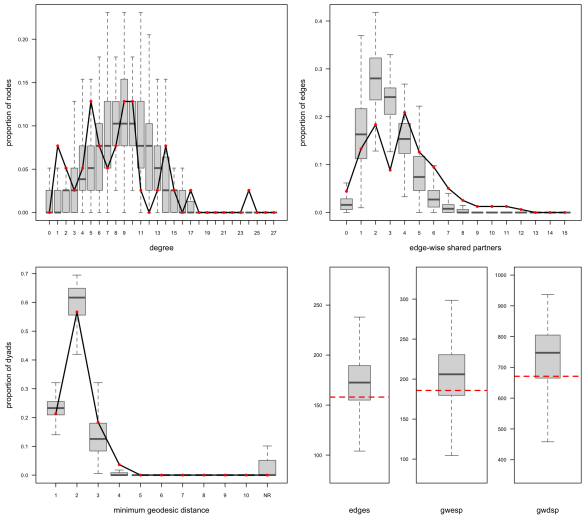


(a) Normal

(b) Uniform



(c) p=1



(d) p=10

Figure 3.11: Goodness-of-fit diagnostics for the Kapferer tailor shop network under the edges, GWESP(0.25), and GWDSP(0.25) model. Each panel shows degree distribution, edgewise shared partners, minimum geodesic distance, and posterior predictive boxplots for the three sufficient statistics, computed from 200 networks simulated from the pooled posterior across 20 replications. The observed network statistics are shown as the black line with red points (distributional diagnostics) and the red dashed line (sufficient statistics). Priors are the normal, uniform, and non-degeneracy prior at $p \in \{1, 10\}$.

with the posterior generating near-empty networks. The uniform prior in panel (b) narrows the upper quartile (edges Q75 drops from 688 to 334, two-stars from $23,644$ to $5,676$) due to its boundedness in θ-space, but the predictive spread remains much wider than the data would warrant.

The non-degeneracy prior at $p = 1$ tightens the predictive distribution around the observed statistics. The edges IQR narrows to $[14, 236]$ with a median of 176, compared to the observed 158. The two-star IQR falls from the normal prior's $[18, 23{,}644]$ to $[15, 3{,}036]$, and the GWESP IQR from $[3, 883]$ to $[3, 296]$. The contraction is concentrated in the upper quartile: the normal prior's edges Q75 of 688 out of a possible 741 places a quarter of its predictive networks near the complete graph, while the non-degeneracy prior brings this down to 236. The posterior predictive boxplots in Figure 3.12(c) show this directly, with the observed statistics sitting much closer to the center of the $p = 1$ distribution than under any other prior choice. The minimum geodesic distance plot corroborates this. Geodesic distance one is the fraction of node pairs sharing a direct edge, and panel (c) captures the observed proportion at distance one far more accurately than panel (a), with the uniform prior in panel (b) closer but still noticeably off. Because the non-degeneracy prior suppresses the full-graph tail, the majority of simulated networks have roughly the right edge density rather than the inflated density that near-complete configurations produce. The degree and edgewise shared partner distributions in panel (c), by contrast, remain visually similar to panels (a) and (b), with both still concentrated at zero. This asymmetry reflects which tail each diagnostic is sensitive to. The lower quartiles are comparable across all priors (edges Q25 ranges from 14 to 17), so all of them still generate a substantial fraction of near-empty networks. A near-empty network places almost all nodes at degree zero and almost all edges at ESP zero, inflating those bars in the averaged distributional plot regardless of the prior used. The geodesic distance one bar is not affected by these outliers, since a network with few edges contributes almost nothing to that bar. In summary, the non-degeneracy prior at $p = 1$ effectively eliminates the tendency of the posterior to produce near-complete graphs,

which is an overbearing source of degeneracy under the normal and uniform priors. The uniform prior's boundedness in θ-space works in the same direction, cutting the edges Q75 from 688 to 334, but the non-degeneracy prior at $p = 1$ does so much more aggressively, bringing it down to 236. The near-empty tail persists at comparable levels across all prior choices, and this is what continues to drive the degree and ESP distributions in the goodness-of-fit diagnostics.

Increasing the prior strength beyond $p = 1$ does not improve the posterior on this model. At $p = 5$, the predictive IQRs widen to levels worse than the uniform prior. The two-star Q75 rises from $3,036$ at $p = 1$ to $17,597$, exceeding the uniform prior's $5,676$. At $p = 10$, the predictive distribution is almost indistinguishable from the normal prior, with edges IQR $[16, 696]$ and two-star IQR $[20, 24190]$ matching the normal's $[17, 688]$ and $[18, 23644]$. The goodness-of-fit diagnostics in panel (d) confirm this behavior, such that the degree distribution under $p = 10$ shows the same isolated-node dominance as the normal prior in panel (a). This reverses the pattern from Section 3.6, where increasing p from 1 to 10 monotonically improved the posterior on the $n = 7$ network.

The difference between this section and Section 3.6 is the convex hull. On the $n = 7$ network, the hull was computed exactly from the 144 distinct statistic combinations. On the Kapferer network, the hull in three-dimensional statistic space is estimated adaptively from networks encountered during sampling. This approximation is adequate at $p = 1$, where the prior applies a mild penalty and the posterior concentrates on parameter values that generate non-degenerate networks. At higher p, the prior depends more heavily on the estimated hull, and the approximation is no longer sufficient. The combination of hull estimation error, the amplification of importance-sampling variance in the prior gradient by a factor of p, and the sensitivity of the noisy HMC sampler to gradient noise likely all contribute to the degradation, though their individual effects cannot be separated without further experimentation. The practical conclusion is that $p = 1$ is the appropriate choice when the hull must be approximated, in contrast to the $p \in [5, 10]$ range identified under

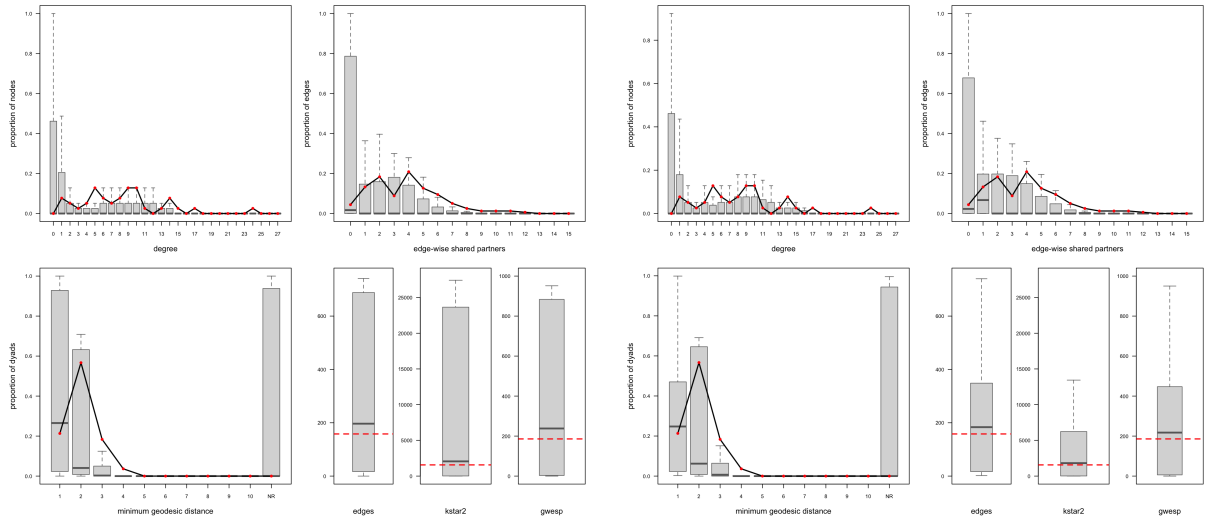
exact computation in Section 3.6.

Prior Choice	μ_1 : **Edges**	μ_2 : **Two-Stars**	μ_3 : **GWESP**
Observed	158	1566	185.8
Normal	197 [17, 688]	2062 [18, 23644]	238 [3, 883]
Uniform	184 [17, 334]	1817 [19, 5676]	217 [6, 428]
NDG, $p = 1$	176 [14, 236]	1624 [15, 3036]	202 [3, 296]
NDG, $p = 5$	180 [16, 593]	1714 [17, 17597]	213 [3, 761]
NDG, $p = 10$	191 [16, 696]	1918 [20, 24190]	229 [6, 894]

Table 3.7: Posterior predictive sufficient statistics for the Kapferer tailor shop network under the edges, two-stars, and GWESP(0.25) model. Each cell reports the posterior predictive median with interquartile range $[25\%, 75\%]$ computed from 200 networks simulated from the pooled posterior. The observed row gives $g(y_{\text{obs}})$.

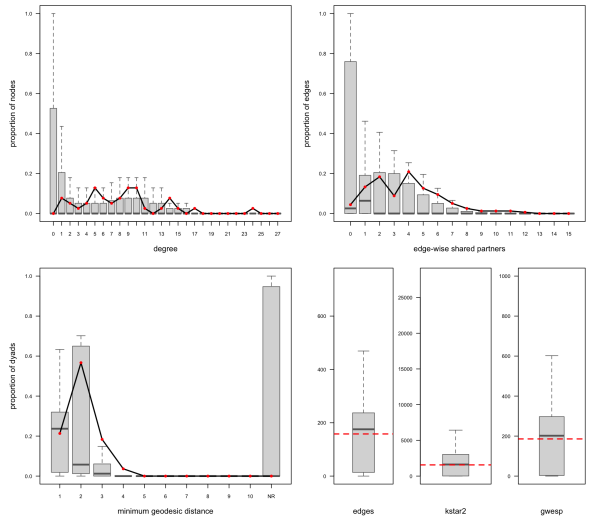
Despite these large predictive differences, the posterior parameter estimates in θ-space are nearly identical across priors. The posterior mean for the edges coefficient is -4.16 under the normal prior and -4.17 under all other priors. The two-star coefficient is 0.080 and the GWESP coefficient is between 1.02 and 1.03 under every configuration, with posterior standard deviations also comparable (0.59 to 0.63 for edges, 0.026 to 0.028 for two-stars). The non-degeneracy prior does not move the center or spread of the θ-space posterior, which is shaped by the likelihood. Its effect operates through the tails. The prior at $p = 1$ reduces density on θ values that generate degenerate networks, and the nonlinear mapping from θ to the sufficient statistic space amplifies these tail differences into the predictive differences visible in Table 3.7. Standard ERGM practice, which reports θ-space point estimates, would show no difference across priors and would not detect the degeneracy problem that the prior addresses.

The two models on the Kapferer network provide support for the non-degeneracy prior

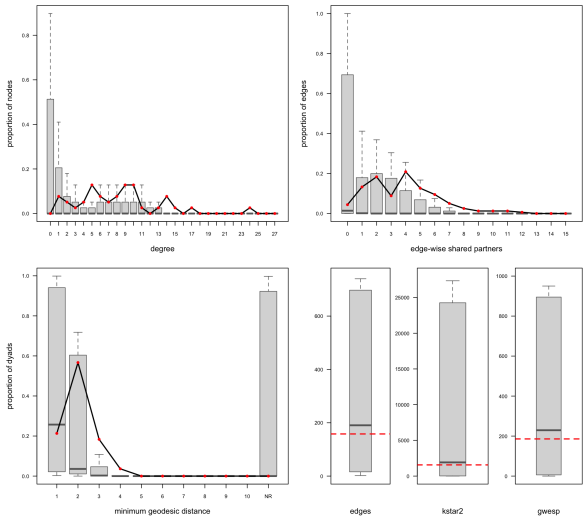


(a) Normal

(b) Uniform



(c) p=1



(d) p=10

Figure 3.12: Goodness-of-fit diagnostics for the Kapferer tailor shop network under the edges, two-stars, and GWESP(0.25) model. Each panel shows degree distribution, edgewise shared partners, minimum geodesic distance, and posterior predictive boxplots for the three sufficient statistics, computed from 200 networks simulated from the pooled posterior across 20 replications. The observed network statistics are shown as the black line with red points (distributional diagnostics) and the red dashed line (sufficient statistics). Priors are the normal, uniform, and non-degeneracy prior at $p \in \{1, 10\}$.

to be considered as a default choice in Bayesian ERGM inference. On the geometrically weighted model, the prior has no measurable effect, confirming it does not interfere when the sufficient statistics chosen for the model already keep degeneracy in check. On the two-star model, the prior at $p = 1$ produces substantially tighter predictive distributions and eliminates the near-complete graph tendency that persists under the normal and uniform priors. The degradation at higher p indicates that the prior strength should be matched to the quality of the convex hull estimate. When the hull is computed exactly, as in Section 3.6, stronger settings improve the posterior. When the hull must be approximated in higher dimensions, $p = 1$ is sufficient and higher values are counterproductive.

3.8 Discussion

This chapter proposed a prior distribution for Bayesian ERGMs that concentrates density on parameter values generating non-degenerate networks. The prior is defined through the probability that a network drawn from the ERGM at parameter θ has sufficient statistics in the interior of the convex hull of possible statistics. A strength parameter p controls how strongly the prior excludes degenerate regions, with $p = 1$ corresponding to the information of a single observed non-degenerate network. Standard prior choices for Bayesian ERGMs place substantial density on parameter values that generate near-empty or near-complete networks, an artifact of specifying priors in the natural parameter space without accounting for the sufficient statistic geometry. The non-degeneracy prior avoids this by defining density directly in terms of the non-degeneracy probability in the parameter space. The importance-reweighting approximation developed in Section 3.4.4 integrates into the noisy HMC framework at negligible additional cost by reusing auxiliary draws already generated for gradient estimation and normalizing constant ratio estimation. The approach requires no new sampling infrastructure beyond what noisy HMC already provides.

On the $n = 7$ network where exact computation is available, the non-degeneracy prior improves on the normal and uniform priors on PMSE, posterior median accuracy, and predictive non-degeneracy across all values of p tested. The normal prior places 43% of its density on the near-full graph region and 29% on near-empty graphs for this model, and these concentrations pull the posterior median away from the observed sufficient statistics. The non-degeneracy prior removes this pull, and the posterior median approaches the observation as p increases from 1 to 10. The reduction is far more complete on the full-graph side, where the prior density drops from 0.432 under the normal to 0.00009 at $p = 1$, compared to the near-empty side where it drops from 0.286 to 0.0522. The approximate posteriors computed via noisy HMC reproduce these findings, validating the importance-reweighting estimator and the adaptive hull for the small-network setting. The non-degeneracy posterior

is also less affected by the noisy HMC approximation error. Approximate samplers tend to broaden the posterior, and this broadening is more damaging when the posterior extends into the degenerate tails, as the normal and uniform posteriors do.

The Kapferer tailor shop network provides complementary evidence from two distinction model specifications. On the geometrically weighted model, where the sufficient statistics already prevent degeneracy, all priors produce almost identical predictive distributions and goodness-of-fit diagnostics. The non-degeneracy prior does not interfere when the model itself already handles degeneracy. On the two-star model, a model that is more prone to degeneracy, the non-degeneracy prior at $p = 1$ narrows the predictive interquartile range for edges from $[17, 688]$ under the normal prior to $[14, 236]$, with similar improvements on the two-star and GWESP statistics. In contrast, the normal and uniform priors generate posterior predictive networks ranging from near-empty to near-complete. The posterior parameter estimates in θ-space are nearly identical across all priors, and the non-degeneracy prior's effect operates through the tails of the θ-space posterior. The nonlinear mapping from θ to the sufficient statistic space amplifies these tail differences into large predictive differences very apparent in the posterior predictive distributions. Standard ERGM practice, which reports θ-space point estimates, would not easily detect the degeneracy problem that the prior addresses.

The most striking effect of the non-degeneracy prior, visible in both the exact $n = 7$ computation and the Kapferer approximation, is the near-total elimination of full-graph degeneracy. On $n = 7$, the prior at $p = 1$ reduces the density on near-complete graphs from 0.432 under the normal prior to 0.00009. On the Kapferer two-star model, the upper quartile for edges contracts from 688 under the normal prior to 236, removing the tendency for the posterior to generate near-complete networks. The near-empty tail is also reduced in both settings, though less dramatically. On $n = 7$ the near-empty density drops from 0.286 to 0.0522, and on Kapferer the lower quartile is comparable across all priors at 14 to 17 edges. Two factors help explain why the prior is more effective on the full-graph side. ERGMs

tend to snap to the complete graph more abruptly than they fade to the empty graph. The boundary between non-degenerate behavior and full-graph degeneracy in the parameter space is narrow, so the prior can push across it cleanly. The empty-graph boundary is more gradual, with a wider band of parameter values that produce somewhat sparse networks without being fully degenerate, and the prior tapers through this region rather than cutting it off. Both observations studied here also sit in the sparser portion of the feasible range, with 7 out of 21 possible edges on $n = 7$ and 158 out of 741 on Kapferer, so the likelihood naturally concentrates away from the full-graph region and reinforces the prior's effect on that side. The full-graph tail is also the more damaging form of degeneracy in practice, as it produces the enormous predictive IQRs and the unrealistic near-complete networks seen under the normal and uniform priors. That the non-degeneracy prior at $p = 1$ virtually eliminates this tail, which is its strongest practical effect.

The recommended strength p depends on how reliably the convex hull can be estimated. On the $n = 7$ network, where the hull is computed exactly from the 144 distinct statistic combinations, $p \in [5, 10]$ provides the best balance between excluding degenerate regions and allowing the likelihood to shape the posterior. On the Kapferer network, where the hull is approximated adaptively in three dimensions, $p = 1$ is sufficient and higher values are counterproductive. At $p = 5$ the predictive distributions widen back toward the uniform prior's, and at $p = 10$ they are indistinguishable from the normal prior's. The combination of hull estimation error and the amplification of importance-sampling variance in the prior gradient at higher p likely drives this degradation, though the individual contributions cannot be separated from the experiments presented here. For practitioners who do not know in advance whether their model is degeneracy-prone or how accurately the hull can be estimated, $p = 1$ is the safe default. The countering hyperparameter q, which would accommodate previously observed degenerate networks, was held at zero throughout this chapter because such observations are almost nonexistent in practice.

Theoretical questions are left open by this chapter. One of which is the convergence rate

of the importance-reweighting estimator in Equation 3.19 as the number of auxiliary draws grows. The estimator is consistent under standard self-normalized importance sampling conditions, but its rate has not been characterized. Another theoretical result is the posterior consistency at the data-generating parameter under the non-degeneracy prior as the network size increases.

The main limitation of the current approach is the quality of the convex hull approximation in higher-dimensional sufficient statistic spaces. The adaptive hull is built from networks encountered during sampling and does not cover the full boundary of the achievable region. One direction for improvement is to precompute sufficient statistics for known extreme network configurations, such as the empty and complete graphs, to anchor the hull vertices before adaptive sampling begins. This would provide a better starting approximation without requiring full enumeration. The non-degeneracy prior is defined generically for any ERGM and is not specific to the noisy HMC sampler used in this chapter. It could be incorporated into exchange algorithms, stochastic gradient Langevin dynamics, or other methods for doubly intractable posteriors by substituting the prior evaluation and its gradient into the relevant update steps. The prior's definition extends to directed networks and models with nodal covariates, since it depends only on the sufficient statistics and their convex hull. Discreteness of the sufficient statistics is required for the hull to be well-defined.

The non-degeneracy prior encodes the minimal substantive belief that the observed network is not degenerate, a belief that is underlying in virtually every applied ERGM study but that standard priors fail to express.

CHAPTER 4

Stochastic Gradient MCMC under the Non-Degeneracy Prior

4.1 Introduction

Chapter 3 proposed the non-degeneracy prior for Bayesian ERGMs and used noisy Hamiltonian Monte Carlo (SBF19) for posterior sampling. Noisy HMC requires a leapfrog trajectory at each iteration, where each of the L leapfrog steps draws auxiliary networks from the ERGM for gradient estimation and normalizing constant ratio estimation. The accept-reject step at the end of each trajectory requires the leapfrog estimator of $Z(\theta)/Z(\theta')$, which could become unreliable when the leapfrog trajectory is long or the auxiliary sample size is small.

In this chapter we replace noisy HMC with stochastic gradient Langevin dynamics (SGLD), following the framework of Zhang and Liang (ZL24) for Bayesian inference in ERGMs with intractable normalizing constants. SGLD updates the parameter vector with a single gradient step plus normal noise at each iteration, with no accept-reject step and no normalizing constant ratio to estimate. The gradient of the log posterior requires $\mathbb{E}_\theta[g(Y)]$, which is approximated by averaging sufficient statistics from auxiliary networks simulated via an inner Markov chain. The non-degeneracy prior gradient integrates into this update at zero additional simulation cost, because it is computed from the same auxiliary draws used for the likelihood gradient. This reuse of auxiliary samples parallels the strategy developed for noisy HMC in Chapter 3, but the single-step structure of SGLD simplifies the full accounting for computation.

Zhang and Liang (ZL24) prove that SGLD converges to the true posterior in 2-Wasserstein distance as the network size n and iteration count t grow, even with the number of auxiliary draws m held constant. This is a stronger guarantee than the approximate stationary distribution targeted by noisy HMC. The convergence result assumes that the posterior is strongly log-concave, which holds under normal or uniform priors but is not guaranteed under the non-degeneracy prior. However, strong log-concavity of the posterior only requires the likelihood Hessian to dominate the prior Hessian. The ERGM likelihood Hessian is $\mathrm{Cov}_\theta[g(Y)]$, whose eigenvalues grow as n^κ. The non-degeneracy prior Hessian stays bounded for fixed model dimension d and moderate prior strength p. At moderate network sizes the likelihood already dominates, so condition (A.1) holds in practice. Section 3.7 provides empirical support on the Kapferer network.

The implementation uses a hybrid resampling strategy in which auxiliary networks are drawn fresh every K Langevin steps and reused with importance-sampling reweighting between draws. This spreads the cost of network simulation while keeping the importance weights bounded over the K-step interval.

Throughout this chapter we adopt the notation of Chapter 3. The ERGM has natural parameter $\theta \in \mathbb{R}^d$, sufficient statistic vector $g(y) \in \mathbb{R}^d$, and probability mass function $p_\theta(y) = \exp\{\theta^\top g(y)\}/Z(\theta)$. In the notation of Zhang and Liang (ZL24), $g(y)$ corresponds to $S(y)$, the parameter dimension d corresponds to their p, and the network size n corresponds to their N. These previous notations are abandoned.

The remainder of this chapter is organized as follows. Section 4.2 reviews the SGLD algorithm for ERGMs and states the convergence result of Zhang and Liang (ZL24). Section 4.3 derives the non-degeneracy prior gradient and its importance-sampling estimator within the SGLD framework. Section 4.4 details the full algorithm, including the hybrid resampling strategy and convex hull computation. Section 4.5 serves as the discussions.

4.2 Stochastic Gradient Langevin Dynamics for ERGMs

4.2.1 The SGLD Update

For an ERGM with prior $\pi(\theta)$, the log-posterior gradient has the form

$$\nabla_\theta \log \pi(\theta \mid y_{\text{obs}}) = g(y_{\text{obs}}) - \mathbb{E}_\theta[g(Y)] + \nabla_\theta \log \pi(\theta), \tag{4.1}$$

where $\mathbb{E}_\theta[g(Y)] = \nabla_\theta \log Z(\theta)$ is the expected sufficient statistic vector under the ERGM at parameter θ. This identity follows from the exponential family form of the ERGM and was used in Chapter 3 for the noisy HMC gradient (Equation 3.4). The prior $\pi(\theta)$ can be normal, uniform, or the non-degeneracy prior developed in Chapter 3.

Since $\mathbb{E}_\theta[g(Y)]$ is intractable, it is estimated by averaging sufficient statistics over m auxiliary networks $\tilde{y}_1, \ldots, \tilde{y}_m$ drawn from an inner Markov chain targeting $p(\cdot \mid \theta)$,

$$\widehat{\mathbb{E}}_\theta[g(Y)] = \frac{1}{m} \sum_{\nu=1}^{m} g(\tilde{y}_\nu). \tag{4.2}$$

This estimator is biased when the inner chain has not fully mixed, and Zhang and Liang (ZL24) account for this bias in their convergence analysis.

The SGLD update (WT11) evolves θ by

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\varepsilon_t}{2} D \, \widehat{\nabla}_\theta \log \pi(\theta^{(t)} \mid y_{\text{obs}}) + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, \varepsilon_t D), \tag{4.3}$$

where $\varepsilon_t > 0$ is the step size and $D \in \mathbb{R}^{d \times d}$ is a symmetric positive definite preconditioning matrix. At initialization, D is set to the diagonal scaling $D = \text{diag}(1/g_1(y_{\text{obs}}), \ldots, 1/g_d(y_{\text{obs}}))$ suggested by Zhang and Liang (ZL24). During sampling, D is updated to the inverse sample covariance of auxiliary statistics restricted to the interior of the convex hull,

$$D = \widehat{\text{Cov}}_\theta[g(Y) \mid g(Y) \in \text{int}(\mathcal{C})]^{-1}, \tag{4.4}$$

computed via eigen-decomposition for numerical stability. This D is the inverse Fisher information of the ERGM restricted to the hull interior, used as a preconditioner. It rescales

each parameter direction so that all directions converge at similar rates rather than letting the most curved direction set the overall step size. The noise covariance $\varepsilon_t D$ in Equation (4.3) is matched to the drift coefficient through the same matrix D, as required by the Langevin diffusion.

4.2.2 Auxiliary Network Simulation

The auxiliary networks $\tilde{y}_1, \ldots, \tilde{y}_m$ are generated by running an inner Markov chain that targets $p(\cdot \mid \theta)$ and is initialized at the observed network y_{obs}, mirroring the convention of the double Metropolis-Hastings sampler (Lia10). For small networks, a Gibbs sampler is used in which each dyad of the network undergoes an update according to its conditional distribution in a full sweep. For large networks, the tie-no-tie (TNT) sampler is used instead. The TNT sampler updates only a subset of dyads at each step, selecting half from the current edges and half from the current non-edges, and accepting or rejecting each proposed change via a Metropolis-Hastings step. Because most real-world networks are sparse, the TNT sampler avoids the cost of updating dyads that have low probability of being edges.

The quality of the auxiliary sample depends on the inner chain's burn-in length, the thinning interval between collected samples, and the total number of samples m. These are free parameters of the algorithm and are discussed in Section 4.4.4.

4.2.3 Convergence Theory

Zhang and Liang (ZL24) establish convergence of the SGLD algorithm for ERGMs under the following conditions, stated here in our notation. Let $\pi_n^* = \pi(\theta \mid y_n)$ denote the posterior for a network of size n, and let $\pi_n^{(t)}$ denote the distribution of $\theta^{(t)}$ produced by SGLD at iteration t.

(A.1) The posterior π_n^* is strongly log-concave with parameter λ_n and has gradient-Lipschitz negative log-density with parameter Λ_n, where $c_0 n^\kappa \leq \lambda_n \leq \Lambda_n \leq c_1 n^\kappa$ for positive

constants c_0, c_1, and κ.

(A.2) The posterior has bounded second moment, $\int \|\theta\|^2 \, \pi(\theta \mid y_n) \, d\theta = O(d)$.

(A.3) The expected squared gradient norm satisfies $\mathbb{E}_\theta \|\nabla_\theta U(\theta)\|^2 = O(n^\kappa(\|\theta\|^2 + d))$, where $U(\theta) = -\log \pi(\theta \mid y_n)$.

Under these conditions, the convergence bound in Theorem 3.1 of Zhang and Liang (ZL24) holds. That is, assume conditions (A.1)–(A.3), then

$$W_2(\pi_n^{(t)}, \pi_n^*) = (1-\omega)^t \, W_2(\pi_n^{(0)}, \pi_n^*) + O\left(\frac{\sqrt{d\varepsilon}}{m \, n^{\kappa/2}}\right) + O(\sqrt{d\varepsilon}) + O\left(\frac{\sqrt{d}\,\varepsilon}{m}\right), \tag{4.5}$$

for some $\omega \in (0, 1)$, where W_2 denotes the 2-Wasserstein distance.

The first term captures geometric convergence of the Markov chain, and the remaining terms capture the combined bias from the Langevin discretization and the finite auxiliary sample. As noted by Zhang and Liang (ZL24) in their Remark 3.1, if the learning rate ε decreases with n and d such that $\varepsilon = o(\min\{n^{-\kappa}, d^{-1}\})$, and if $d = o(n^\kappa)$, then m can be held constant and $W_2(\pi_n^{(t)}, \pi_n^*) \to 0$ as $n \to \infty$ and $t \to \infty$. That is, a short inner Markov chain with fixed length would be satisfactory when the network is large enough, because the posterior concentrates as n grows.

The strong log-concavity condition (A.1) is satisfied when the prior $\pi(\theta)$ is itself log-concave, since the ERGM negative log-likelihood has Hessian $\mathrm{Cov}_\theta[g(Y)]$, which is positive semidefinite with eigenvalues scaling as n^κ (FH17). A normal or uniform prior preserves this property. The non-degeneracy prior does not, because $\log P_\theta(A) = \log Z_A(\theta) - \log Z(\theta)$ is a difference of two convex functions and its Hessian can be indefinite. As argued in Section 4.1, the likelihood Hessian dominates the prior Hessian at moderate prior strength and network size, so condition (A.1) holds in practice.

4.2.4 Comparison with Noisy HMC

The SGLD sampler differs from the noisy HMC sampler used in Chapter 3 in several ways. Noisy HMC evolves an L-step leapfrog trajectory at each outer iteration, drawing auxiliary networks at each of the L intermediate parameter values. At the end of each trajectory, an accept-reject step requires the leapfrog estimator of $Z(\theta)/Z(\theta')$, computed by chaining local importance-sampling ratios along the trajectory. SGLD instead takes a single gradient step with injected noise and has no accept-reject step. The per-iteration cost of SGLD is therefore m auxiliary network simulations, compared to $L \cdot m$ for noisy HMC, where L is typically in the range 40–50 for the Kapferer network studied in Chapter 3.

For the non-degeneracy prior, both samplers reuse auxiliary draws for the prior gradient at no additional simulation cost. The difference is that noisy HMC evaluates the prior at each of L leapfrog steps within a trajectory, while SGLD evaluates it once per Langevin step. The theoretical guarantees also differ. Noisy HMC targets an approximate stationary distribution whose accuracy depends on the leapfrog estimator's noise, while SGLD under the conditions of Equation 4.5 converges to the true posterior in 2-Wasserstein distance.

Our implementation departs from the standard SGLD algorithm of Zhang and Liang (ZL24) in four ways. First, the non-degeneracy prior gradient $\nabla_\theta \ell_{\mathrm{ND}}(\theta)$ is added to the Langevin update, computed from the same auxiliary draws used for the likelihood gradient. Second, auxiliary networks are reused across K consecutive Langevin steps with importance-sampling reweighting rather than drawn fresh at every iteration. The non-degeneracy prior needs larger auxiliary samples ($m = 500$–1000 versus $m = 10$ in Zhang and Liang's experiments) to get enough interior draws for the conditional mean $\widehat{\mathbb{E}}_\theta[g(Y) \mid A]$, and reusing each batch for K steps spreads this cost. Third, the preconditioner D is updated during sampling from the inverse covariance of auxiliary statistics in the hull interior, rather than held fixed at the initial diagonal scaling. Without this, degenerate draws inflate the covariance and shrink the step size. Fourth, the convex hull of achievable sufficient statistics is

estimated adaptively and used both for the prior gradient's hull-membership indicator and for the covariance restriction in the preconditioner. The first of these is the methodological contribution of this chapter. The other three are practical changes motivated by the prior.

4.3 Non-Degeneracy Prior Gradient in the SGLD Framework

4.3.1 The Non-Degeneracy Prior

The non-degeneracy prior developed in Chapter 3 is defined through the probability that a network drawn from the ERGM at parameter θ has sufficient statistics in the interior of the convex hull $\mathcal{C} = \text{conv}\{g(y) : y \in \mathcal{Y}_n\}$. Let $A = \{g(Y) \in \text{int}(\mathcal{C})\}$ denote the event that the sufficient statistics fall in the interior. The non-degeneracy prior with strength hyperparameters $p \geq 1$ and $q \geq 0$ is

$$\pi(\theta) \propto P_\theta(A)^p \{1 - P_\theta(A)\}^q, \tag{4.6}$$

where

$$P_\theta(A) = \frac{\sum_{y:\, g(y) \in \text{int}(\mathcal{C})} \exp\{\theta^\top g(y)\}}{Z(\theta)}. \tag{4.7}$$

The baseline case $p = 1$, $q = 0$ corresponds to the information of a single observed non-degenerate network, and q defaults to zero in most settings. Chapter 3 found that $p \in [5, 10]$ gave the best posterior performance when the convex hull was computed exactly on the simple $n = 7$ network, while $p = 1$ was sufficient and higher values were counterproductive when the hull had to be approximated on the larger Kapferer network.

The log-density of the prior, up to an additive constant, is then

$$\ell_{\text{ND}}(\theta) = p \log P_\theta(A) + q \log\{1 - P_\theta(A)\}. \tag{4.8}$$

4.3.2 Prior Gradient Derivation

The gradient of $\log P_\theta(A)$ with respect to θ follows from the exponential family form. Writing $P_\theta(A) = Z_A(\theta)/Z(\theta)$, where $Z_A(\theta) = \sum_{y:\, g(y)\in \mathrm{int}(\mathcal{C})} \exp\{\theta^\top g(y)\}$, and differentiating gives

$$\nabla_\theta \log P_\theta(A) = \mathbb{E}_\theta[g(Y) \mid A] - \mathbb{E}_\theta[g(Y)]. \tag{4.9}$$

The gradient of $P_\theta(A)$ itself follows by the chain rule,

$$\nabla_\theta P_\theta(A) = P_\theta(A)\left[\mathbb{E}_\theta[g(Y) \mid A] - \mathbb{E}_\theta[g(Y)]\right]. \tag{4.10}$$

Differentiating Equation (4.8) and substituting Equation (4.10) yields the gradient of the log-prior,

$$\nabla_\theta \ell_{\mathrm{ND}}(\theta) = \frac{p - (p+q)\,P_\theta(A)}{1 - P_\theta(A)}\left[\mathbb{E}_\theta[g(Y) \mid A] - \mathbb{E}_\theta[g(Y)]\right]. \tag{4.11}$$

An equivalent symmetric expression uses $A^c = \{g(Y) \notin \mathrm{int}(\mathcal{C})\}$ and the identity $\nabla_\theta \log\{1 - P_\theta(A)\} = \mathbb{E}_\theta[g(Y) \mid A^c] - \mathbb{E}_\theta[g(Y)]$, giving

$$\nabla_\theta \ell_{\mathrm{ND}}(\theta) = p\left[\mathbb{E}_\theta[g(Y) \mid A] - \mathbb{E}_\theta[g(Y)]\right] + q\left[\mathbb{E}_\theta[g(Y) \mid A^c] - \mathbb{E}_\theta[g(Y)]\right]. \tag{4.12}$$

When $q = 0$, the symmetric form reduces to $p\left[\mathbb{E}_\theta[g(Y) \mid A] - \mathbb{E}_\theta[g(Y)]\right]$, which involves no denominator $1 - P_\theta(A)$ and is therefore well-defined even when $P_\theta(A) = 1$. The fractional form (4.11) has denominator $1 - P_\theta(A)$ and is undefined at that boundary. For this reason the implementation uses the symmetric form whenever $q = 0$ and falls back to the symmetric form with q set to zero whenever the empirical estimate $\widehat{P}_\theta(A) = 1$, which can occur when all auxiliary draws have statistics in the interior of $\mathcal{C}$.

The complete posterior gradient under the non-degeneracy prior is therefore

$$\nabla_\theta \log \pi(\theta \mid y_{\mathrm{obs}}) = g(y_{\mathrm{obs}}) - \mathbb{E}_\theta[g(Y)] + \nabla_\theta \ell_{\mathrm{ND}}(\theta), \tag{4.13}$$

which decomposes into the likelihood gradient $g(y_{\mathrm{obs}}) - \mathbb{E}_\theta[g(Y)]$ and the prior gradient $\nabla_\theta \ell_{\mathrm{ND}}(\theta)$.

4.3.3 Importance-Reweighting Estimator

The auxiliary draws used to estimate $\mathbb{E}_\theta[g(Y)]$ in the likelihood gradient can be reused to estimate the conditional expectations in the prior gradient. Given m auxiliary networks $\tilde{y}_1, \ldots, \tilde{y}_m$ drawn from $p(\cdot \mid \theta_0)$ at a reference parameter θ_0, the self-normalized importance weights for evaluating expectations under $p(\cdot \mid \theta)$ are

$$w_\nu(\theta) = \exp\{(\theta - \theta_0)^\top [g(\tilde{y}_\nu) - g(y_{\text{obs}})]\}, \tag{4.14}$$

where the centering at $g(y_{\text{obs}})$ in the exponent improves numerical stability by keeping the exponents small when θ is near the posterior mode. The intractable normalizing constant ratio $Z(\theta_0)/Z(\theta)$ cancels in the self-normalized ratios, as in the importance-reweighting approximation for the non-degeneracy prior developed in Chapter 3 (Equation 3.19).

The conditional expectation over the interior, the unconditional expectation, and the non-degeneracy probability are estimated as

$$\widehat{\mathbb{E}}_\theta[g(Y) \mid A] = \frac{\sum_{\nu=1}^m w_\nu(\theta)\, g(\tilde{y}_\nu)\, \mathbf{1}\{g(\tilde{y}_\nu) \in \text{int}(\mathcal{C})\}}{\sum_{\nu=1}^m w_\nu(\theta)\, \mathbf{1}\{g(\tilde{y}_\nu) \in \text{int}(\mathcal{C})\}}, \tag{4.15}$$

$$\widehat{\mathbb{E}}_\theta[g(Y)] = \frac{\sum_{\nu=1}^m w_\nu(\theta)\, g(\tilde{y}_\nu)}{\sum_{\nu=1}^m w_\nu(\theta)}, \tag{4.16}$$

$$\widehat{P}_\theta(A) = \frac{\sum_{\nu=1}^m w_\nu(\theta)\, \mathbf{1}\{g(\tilde{y}_\nu) \in \text{int}(\mathcal{C})\}}{\sum_{\nu=1}^m w_\nu(\theta)}. \tag{4.17}$$

The plug-in estimator of the prior gradient is then

$$\widehat{\nabla}_\theta \ell_{\text{ND}}(\theta) = \frac{p - (p+q)\,\widehat{P}_\theta(A)}{1 - \widehat{P}_\theta(A)} \left[\widehat{\mathbb{E}}_\theta[g(Y) \mid A] - \widehat{\mathbb{E}}_\theta[g(Y)]\right], \tag{4.18}$$

with the symmetric form from Equation (4.12) used in place of the fractional form when $q = 0$.

When $\theta = \theta_0$, which is the case at the first Langevin step after a fresh auxiliary draw, all weights reduce to $w_\nu(\theta_0) = 1$ and the estimators simplify to unweighted sample averages,

$$\widehat{P}_{\theta_0}(A) = \frac{1}{m}\sum_{\nu=1}^m \mathbf{1}\{g(\tilde{y}_\nu) \in \text{int}(\mathcal{C})\}, \qquad \widehat{\mathbb{E}}_{\theta_0}[g(Y) \mid A] = \frac{\sum_\nu g(\tilde{y}_\nu)\, \mathbf{1}\{g(\tilde{y}_\nu) \in \text{int}(\mathcal{C})\}}{\sum_\nu \mathbf{1}\{g(\tilde{y}_\nu) \in \text{int}(\mathcal{C})\}}. \tag{4.19}$$

The full importance-weighting machinery in Equations (4.15)–(4.17) is needed on subsequent Langevin steps within a resample interval, where $\theta^{(t)}$ has drifted from θ_0. This is detailed in the following section.

The prior gradient evaluation uses the same auxiliary draws and the same importance weights as the likelihood gradient, and adds only the hull-membership indicator computation $\mathbf{1}\{g(\tilde{y}_\nu) \in \text{int}(\mathcal{C})\}$ at each auxiliary network. Since network simulation is by far the dominant computational cost per iteration, the non-degeneracy prior is evaluated at negligible cost once the auxiliary sample exists.

4.4 Algorithm

The full algorithm is given below, with each step labeled by its role in the SGLD update. When too few auxiliary draws fall in the hull interior or the importance weights degenerate, the algorithm falls back to the likelihood gradient alone.

4.4.1 Hybrid Resampling Strategy

The algorithm uses a three-level loop structure. At the outermost level, the sampler runs for T iterations after burn-in, collecting posterior draws. At the middle level, fresh auxiliary networks are drawn from $p(\cdot \mid \theta^{(t)})$ at the current parameter value, the reference θ_0 is set to $\theta^{(t)}$, and the convex hull and preconditioner D are updated. At the innermost level, K Langevin steps reuse the same auxiliary sample with importance-sampling reweighting to account for the drift of $\theta^{(t)}$ from the fixed θ_0.

The parameter K controls the tradeoff between simulation cost and importance-weight quality. Each fresh auxiliary draw requires running the inner Markov chain from the observed network, which is the most expensive operation per iteration. Reusing each batch for K Langevin steps rather than one spreads this cost over K updates. However, the importance weights $w_\nu(\theta^{(t)})$ in Equation (4.14) become less accurate as $\|\theta^{(t)} - \theta_0\|$ grows over the K

steps, because the effective sample size decreases when the weights concentrate on a few auxiliary networks. Small K (such as $K = 3$ or 5) keeps the weights fresh, while large K reduces cost at the expense of gradient accuracy. Within the inner loop, the importance weights are recomputed at each Langevin step from the current $\theta^{(t)}$ and the fixed auxiliary statistics Δ_ν.

This hybrid strategy lies between two extremes from the literature. Zhang and Liang (ZL24) draw fresh auxiliary networks at every iteration (corresponding to $K = 1$), and their convergence theory in Equation 4.5 assumes this. At the other extreme, a fixed reference sample drawn once at the MPLE would avoid all resampling cost but accumulate unbounded importance-weight degeneracy over the full chain. The hybrid approach with moderate K stays close to the theoretical guarantee of fresh-per-step while reducing simulation cost by a factor of K.

At the first middle-loop iteration, $10m$ auxiliary networks are drawn rather than m, in order to initiate the convex hull estimate with a larger and ideally more diverse set of sufficient statistic vectors. This warm-start is analogous to the 100,000-network MPLE-seeded hull used in Chapter 3.

4.4.2 Full Algorithm

The full SGLD algorithm under the non-degeneracy prior is given below.

1. **Initialize.** Set $\theta^{(0)}$ to the maximum pseudolikelihood estimate (MPLE). Set $D = \text{diag}(1/g_1(y_{\text{obs}}), \ldots, 1/g_d(y_{\text{obs}}))$. Initialize the convex hull $\mathcal{C} \leftarrow \emptyset$.

2. **For each outer iteration** $t = 1, \ldots, T + B$ (where B is the burn-in).

   (a) **Draw auxiliary networks.** Simulate $\tilde{y}_1, \ldots, \tilde{y}_m$ from $p(\cdot \mid \theta^{(t)})$ via the inner Markov chain (with $10m$ at $t = 1$). Set $\theta_0 \leftarrow \theta^{(t)}$.

(b) **Compute centered statistics.** For each $\nu = 1, \ldots, m$, compute $\Delta_\nu = g(\tilde{y}_\nu) - g(y_{\text{obs}})$.

(c) **Add to hull candidate set.** Append the current batch of statistics to $\mathcal{C}$.

(d) **Determine hull membership.** For each ν, compute $\mathbf{1}\{g(\tilde{y}_\nu) \in \text{int}(\mathcal{C})\}$ via the linear programming test described in Section 4.4.3.

(e) **Update D.** If the number of interior auxiliary networks exceeds 20, set $D = \widehat{\text{Cov}}[g(\tilde{y}) \mid A]^{-1}$ via eigendecomposition. Only interior samples are used because degenerate auxiliary draws have extreme statistics that inflate the covariance and shrink the Langevin step toward zero. Otherwise retain the previous D.

(f) **For each Langevin step** $k = 1, \ldots, K$.

i. Compute importance weights $w_\nu = \exp\{(\theta^{(t)} - \theta_0)^\top \Delta_\nu\}$.

ii. Compute the likelihood gradient $\widehat{\nabla}_\theta \log p(y_{\text{obs}} \mid \theta^{(t)})$ via the importance-weighted mean of Δ_ν.

iii. If the number of interior draws exceeds 5, compute the prior gradient $\widehat{\nabla}_\theta \ell_{\text{ND}}(\theta^{(t)})$ via Equation (4.18) using the same weights. Otherwise set $\widehat{\nabla}_\theta \ell_{\text{ND}}(\theta^{(t)}) = -\widehat{\nabla}_\theta \log p(y_{\text{obs}} \mid \theta^{(t)})$, which zeros the total drift and reduces the update to a pure noise step. This fallback avoids unstable updates when the conditional mean $\widehat{\mathbb{E}}_\theta[g(Y) \mid A]$ is estimated from too few interior draws.

iv. If $\|\widehat{\nabla}_\theta \ell_{\text{ND}}(\theta^{(t)})\|_1 > G_{\max}$, replace both the prior and likelihood gradients with zero for this step, reducing the update to a pure noise step. Large prior gradients arise when the importance weights concentrate on a small number of auxiliary networks near the hull boundary, making the conditional mean $\widehat{\mathbb{E}}_\theta[g(Y) \mid A]$ unstable. We use $G_{\max} = 10,000$.

v. Update $\theta^{(t)} \leftarrow \theta^{(t)} + \frac{\varepsilon}{2} D \left(\widehat{\nabla}_\theta \log p(y_{\text{obs}} \mid \theta^{(t)}) + \widehat{\nabla}_\theta \ell_{\text{ND}}(\theta^{(t)})\right) + \eta$, where $\eta \sim \mathcal{N}(0, \varepsilon D)$.

(g) **Clean up hull.** Remove points classified as interior from $\mathcal{C}$, deduplicate the

remaining boundary candidates, and cap $\mathcal{C}$ at 5,000 unique statistic vectors.

(h) If $t > B$, store $\theta^{(t)}$.

4.4.3 Convex Hull Computation

The indicator $\mathbf{1}\{g(\tilde{y}_\nu) \in \text{int}(\mathcal{C})\}$ requires testing whether each auxiliary network's sufficient statistics lie inside the convex hull of achievable statistic vectors. The hull $\mathcal{C}$ is estimated adaptively and grows during sampling.

At the first outer iteration, the warm-start auxiliary sample of $10m$ networks provides the initial hull candidate set. At each subsequent iteration, the current batch of auxiliary statistics is tested against the hull, and interior points are discarded. Only boundary and exterior points are retained in the hull representation, because interior points do not contribute to defining the boundary. The retained set is deduplicated and capped at 5,000 unique statistic vectors to bound the cost of subsequent tests. Because points are only added and never removed from the boundary set, the hull can only expand over the course of sampling.

Hull membership is tested by linear programming (LP). Let M denote the centered hull matrix, whose rows are the candidate boundary statistic vectors minus the centroid, and let $\delta_\nu = g(\tilde{y}_\nu) - \bar{g}$ denote the centered test point. For each auxiliary network ν, the LP

$$o_\nu^* = \min_{z \in \mathbb{R}^d} \delta_\nu^\top z \qquad \text{subject to} \quad Mz \geq -\mathbf{1}, \tag{4.20}$$

is solved, where $\mathbf{1}$ is a vector of ones. If the test point lies inside the convex hull of the rows of M, no separating direction z exists that can push $\delta_\nu^\top z$ below -1 while satisfying the constraints, so the optimum satisfies $o_\nu^* \geq -1$. If the test point lies outside the hull, a separating direction drives $o_\nu^* < -1$. A point is classified as interior when $o_\nu^* > -0.99$, which builds in a small tolerance near the boundary. The LP solver operates under a time limit per point, and points that cannot be classified within this budget are conservatively assigned to the exterior.

For the $n = 7$ network studied in Chapter 3, the hull has 10 boundary vertices among 144

distinct statistic combinations and converges to the true hull within the early iterations of the chain. For larger networks in higher-dimensional statistic spaces, the adaptively estimated hull is necessarily an approximation. Chapter 3 found that the quality of this approximation is the binding constraint on the prior strength p, with $p = 1$ performing well even when the hull is approximate and higher values of p degrading when the hull is not accurately estimated.

4.4.4 Tuning

The algorithm has several tuning parameters.

The step size ε controls the magnitude of each Langevin update. A constant ε targets an approximate posterior with $O(\sqrt{d\varepsilon})$ discretization bias (Equation 4.5). A decaying step size $\varepsilon_t = O(1/t^\gamma)$ with $0 < \gamma < 1$ eliminates this bias asymptotically (TTV16), at the cost of slower mixing in later iterations. The default is $\varepsilon = 0.1$.

The auxiliary sample size m determines the number of networks drawn at each resampling point. Zhang and Liang (ZL24) show theoretically that m can be held constant when $d = o(n^\kappa)$, and their experiments use $m = 10$ for moderate networks. However, the non-degeneracy prior gradient requires enough interior draws to estimate the conditional mean $\widehat{\mathbb{E}}_\theta[g(Y) \mid A]$ reliably, and larger m (in the range 500 to 1000) is used in practice.

The resample interval K sets the number of Langevin steps per fresh auxiliary draw. Values of K in the range 3 to 5 provide a reasonable tradeoff. The inner chain parameters (burn-in, thinning interval, sample size) control the quality of each auxiliary sample and follow standard ERGM MCMC practice.

The chain is initialized at the MPLE, and burn-in is set to B/K outer iterations. We default to $B = 1000$ and $T = 5000$ outer iterations with parallel chains.

4.5 Discussion

This chapter introduces a stochastic gradient Langevin sampler for Bayesian ERGMs under the non-degeneracy prior. The main methodological piece is the importance-reweighted estimator of the prior gradient, built from the same auxiliary networks that the likelihood gradient already requires. The prior is therefore evaluated at no additional simulation cost, as in the noisy HMC implementation of Chapter 3.

SGLD trades per-iteration cost for sample quality. A noisy HMC trajectory does $L \cdot m$ simulations and an accept-reject step. SGLD does m simulations per outer iteration and no acceptance check. The simulation work per outer iteration is lower under SGLD, but each SGLD step is also a smaller move in parameter space, and SGLD carries discretization bias that noisy HMC's accept-reject step partly corrects.

The non-degeneracy prior gradient fits naturally into any Bayesian ERGM framework that already draws auxiliary networks for likelihood evaluation. Noisy HMC, SGLD, double Metropolis-Hastings, and exchange algorithms all share the same need, that a sample from $p(\cdot \mid \theta)$ at the current parameter value to estimate $\mathbb{E}_\theta[g(Y)]$. The prior gradient is also a function of expectations under $p(\cdot \mid \theta)$, restricted to the interior of the convex hull. Adding it to any of these samplers requires the same hull-membership indicator and the same importance-reweighted estimators developed here. The methodological contribution of this chapter is the SGLD integration, but the prior itself is not tied to one sampler.

A few limitations should be acknowledged. The strong log-concavity condition required by Zhang and Liang's convergence theorem is not formally satisfied under the non-degeneracy prior. The dominance argument in Section 4.1 suggests the condition holds in practice for moderate prior strength and network size, but this is heuristic rather than proven. The hybrid resampling strategy with $K > 1$ Langevin steps per auxiliary draw also breaks the theorem's assumption of fresh draws at every step. Convex hull is estimated adaptively. Chapter 3 found that the quality of this approximation is the binding constraint on the prior

strength p. When the hull is poorly estimated, only $p = 1$ tends to perform well. Higher prior strengths require an accurate hull.

Some theoretical questions are left open by this chapter. We are interested in whether the 2-Wasserstein convergence guarantee in Equation 4.5 extends to the practical modifications detailed in Section 4.2.4. An extension would put the practical sampler on the same theoretical footing as the standard form. An additional question is whether the strong log-concavity condition (A.1) can be proven for the non-degeneracy posterior. A formal proof, or a counterexample identifying when (A.1) fails, would replace the dominance argument with a direct statement of where Zhang and Liang's theorem applies.

The chapter does not include empirical validation, which is the most pressing follow-up work. Running the sampler against the exact posteriors on the $n = 7$ network from Chapter 3 would measure how closely the approximate posterior matches the truth, and at what cost relative to noisy HMC. Larger networks where exact computation is infeasible would test the practical limits of the hull approximation. Several other directions are worth pursuing: better hull estimation methods that scale to higher-dimensional statistic spaces, variance reduction for the importance-reweighted estimator (which would let larger K run safely), and adaptation of the same prior gradient to other auxiliary-sampling frameworks along the lines just described.

CHAPTER 5

Conclusions

Networks are a natural representation of relationships between actors throughout the social and economic sciences. Friendships and partnerships can all be expressed as relational ties between pairs of actors. Exponential-family random graph models are one of the few frameworks that let researchers represent the dependence between these ties in a principled and interpretable way, and they have been used widely across these disciplines for several decades. The dependence between ties is what makes ERGMs valuable, and it is also what makes inference for them difficult.

The difficulty is the geometry of network sufficient statistics. The mapping between natural parameters and the network statistics they generate is highly nonlinear, and the curvature of the log-partition function is shaped by the same tie dependence that motivates the model class. Methods that ignore this geometry, whether in the likelihood, the prior, or the sampler, tend to fail in the regimes where ERGMs are most worth using. This dissertation develops Bayesian methodology that respects the geometry rather than averaging over it.

Chapter 2 closes off one tempting but unreliable solution. Variational mean-field approximation replaces the joint ERGM with a model of conditionally independent ties, which flattens exactly the curvature that the dependence structure of the model creates. The method's apparent success in prior work came from a methodological error in how the algorithm was initialized rather than from any genuine accuracy of the approximation. The regimes in which it does well are ones where the model is already close to tie-independent. When the model has the dependence structure that ERGMs are designed to capture, the

approximation breaks down. We additionally recommend tapered ERGMs as a useful model specification tool and MPLE-initialized MCMC-MLE as the reliable likelihood-level estimator, both of which carry forward into the rest of the dissertation.

Chapter 3 introduces the main Bayesian contribution. Default priors specified in the natural parameter space silently encode beliefs about graph structure that researchers would almost never endorse if asked directly, because the mapping from natural parameters to expected network statistics is highly nonlinear. The non-degeneracy prior corrects this by assigning density in proportion to the probability that the ERGM at a given parameter value generates networks with sufficient statistics in the interior of the convex hull of achievable statistics. The prior steers posterior mass away from parameter values that produce empty or full networks and onto values that produce realistic ones. The auxiliary network simulations that posterior samplers already require can be reused with importance reweighting to evaluate the prior and its gradient, so the prior integrates into existing samplers at essentially no additional computational cost.

Chapter 4 demonstrates that the non-degeneracy prior is not tied to one sampler. The gradient construction integrates naturally into stochastic gradient Langevin dynamics, a posterior sampling method that takes a single noisy gradient step per iteration rather than evolving a full leapfrog trajectory. Notably, Bayesian inference for ERGMs under the non-degeneracy prior viable for most auxiliary-sample-based frameworks, and the same auxiliary network simulations support both the likelihood gradient and the prior gradient at no extra cost under SGLD.

Together, these contributions point to a single methodological recipe for Bayesian ERGM practitioners. For model specification, choose geometrically weighted terms to avoid degenerate behavior, or use tapered ERGMs when constrained to a degeneracy-prone specification. For inference, initialize at the maximum pseudolikelihood estimate, place the non-degeneracy prior in place of the default normal or uniform priors, and pair the prior with whichever doubly intractable sampler is already in use. For diagnostics, examine the posterior predictive

distribution in the space of sufficient statistics, since posterior summaries in the natural parameter space are nearly invariant to the prior choice and will hide what the prior is doing. The non-degeneracy prior is computationally inexpensive and an effective default choice for Bayesian ERGM inference.

The Bayesian methods that respect ERGMs' geometry produce reliable inference even where the model class is most demanding.

Bibliography

[AFE16] Pierre Alquier, Nial Friel, Richard Everitt, and Aiden Boland. "Noisy Monte Carlo: Convergence of Markov Chains with Approximate Transition Kernels." *Statistics and Computing*, **26**(1–2):29–47, 2016.

[ALR13] Yves F. Atchadé, Nicolas Lartillot, and Christian P. Robert. "Bayesian Computation for Intractable Normalizing Constants." *Brazilian Journal of Probability and Statistics*, **27**(4):416–436, 2013.

[Bar78] Ole E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory.* Wiley, New York, 1978.

[BCL11] Peter J. Bickel, Aiyou Chen, and Elizaveta Levina. "The method of moments and degree distributions for network models." *The Annals of Statistics*, **39**(5), October 2011.

[BH22] Bart Blackburn and Mark S. Handcock. "Practical Network Modeling via Tapered Exponential-Family Random Graph Models." *Journal of Computational and Graphical Statistics*, pp. 1–14, 2022.

[BM17] Vincent Boucher and Ismael Mourifie. "My friend far, far away: a random field approach to exponential random graph models." *The Econometrics Journal*, **20**(3):S14–S46, 10 2017.

[BMP22] Gino Biondini, Antonio Moro, Barbara Prinari, and Oleg Senkevich. "p-star Models, Mean-field Random Networks, and the Heat Hierarchy." *Physical Review E*, **105**(1):014306, 2022.

[BPR13] Alexandros Beskos, Natesh Pillai, Gareth Roberts, Jesús-María Sanz-Serna, and Andrew Stuart. "Optimal Tuning of the Hybrid Monte Carlo Algorithm." *Bernoulli*, **19**(5A):1501–1534, 2013.

[CD13] Sourav Chatterjee and Persi Diaconis. "Estimating and Understanding Exponential Random Graph Models." *The Annals of Statistics*, **41**(5):2428–2461, 2013.

[De 17] Aureo De Paula. "Econometrics of Network Models." In Bo Honore, Ariel Pakes, Monika Piazzesi, and Larry Samuelson, editors, *Advances in Economics and Econometrics: Eleventh World Congress*, volume 1 of *Econometric Society Monographs*, chapter 8, p. 268–323. Cambridge University Press, 2017.

[DHG09] Marijtje A. J. van Duijn, Mark S. Handcock, and Krista J. Gile. "A Framework for the Comparison of Maximum Pseudo Likelihood and Maximum Likelihood Estimation of Exponential Family Random Graph Models." *Social Networks*, **31**:52–62, 2009.

[DI10] Steven N. Durlauf and Yannis M. Ioannides. "Social Interactions." *Annual Review of Economics*, **2**(1):451–478, 2010.

[DM10] Amir Dembo and Andrea Montanari. "Ising Models on Locally Tree-Like Graphs." *The Annals of Applied Probability*, **20**(2):565–592, 2010.

[EN78] Richard S. Ellis and Charles M. Newman. "Limit theorems for sums of dependent random variables occurring in statistical mechanics, II." *Probability Theory and Related Fields*, **44**(2):117–139, 1978.

[Eve12] Richard G. Everitt. "Bayesian Parameter Estimation for Latent Markov Random Fields and Social Networks." *Journal of Computational and Graphical Statistics*, **21**(4):940–960, 2012.

[FH17] Ian Fellows and Mark S. Handcock. "Removing phase transitions from Gibbs measures." In *Artificial Intelligence and Statistics*, pp. 289–297, 2017.

[FS86] Ove Frank and David Strauss. "Markov graphs." *Journal of the American Statistical Association*, **81**(395):832–842, 1986.

[GKM09] Steven M. Goodreau, James Kitts, and Martina Morris. "Birds of a Feather, or Friend of a Friend? Using Statistical Network Analysis to Investigate Adolescent Social Networks." *Demography*, **46**:103–125, 2009.

[GT92] Charles J. Geyer and Elizabeth A. Thompson. "Constrained Monte Carlo maximum likelihood calculations (with discussion)." *Journal of the Royal Statistical Society, Series B*, **54**:657–699, 1992.

[Han03] Mark S. Handcock. "Assessing Degeneracy in Statistical Models of Social Networks." Working paper #39, Center for Statistics and the Social Sciences, University of Washington, 2003.

[HHB03] Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris. *ergm: Fit, simulate and analyze exponential-family models for networks.* Statnet Project, `https://statnet.org`, 2003.

[HHB08a] Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris. "**ergm**: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks." *Journal of Statistical Software*, **24**(3):3, 2008.

[HHB08b] Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris. "statnet: Software tools for the representation, visualization, analysis and simulation of social network data." *Journal of Statistical Software*, **24**(1):1, 2008.

[HHH12] Ruth M. Hummel, David R. Hunter, and Mark S. Handcock. "Improving Simulation-Based Algorithms for Fitting ERGMs." *Journal of Computational and Graphical Statistics*, **21**(4):920–939, 2012.

[HKF21] Mark S. Handcock, Pavel N. Krivitsky, and Ian Fellows. **ergm.tapered***: Tapered Exponential-Family Models for Networks*. Los Angeles, CA, 2021. R package version 1.2.

[Hun07] David R. Hunter. "Curved Exponential Family Models for Social Networks." *Social Networks*, **29**(2):216–230, 2007.

[JKM18] Vishesh Jain, Frederic Koehler, and Elchanan Mossel. "The Mean-Field Approximation: Information Inequalities, Algorithms, and Complexity." In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1326–1347. PMLR, 06–09 Jul 2018.

[JL14] Ick Hoon Jin and Faming Liang. "Use of SAMC for Bayesian Analysis of Statistical Models with Intractable Normalizing Constants." *Computational Statistics & Data Analysis*, **71**:402–416, 2014.

[Kap72] Bruce Kapferer. *Strategy and Transaction in an African Factory*. Manchester University Press, Manchester, 1972.

[KH14] Pavel N. Krivitsky and Mark S. Handcock. "A separable model for dynamic networks." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**(1):29–46, 2014.

[KHM11] Pavel N. Krivitsky, Mark S. Handcock, and Martina Morris. "Adjusting for network size and composition effects in exponential-family random graph models." *Statistical Methodology*, **8**(4):319–339, jul 2011.

[KHM21] Pavel N. Krivitsky, David R. Hunter, Martina Morris, and Chad Klumb. "ergm 4: New features." *arXiv,* `https://arxiv.org/abs/2106.04997`, 2021.

[Lia07] Faming Liang. "Continuous Contour Monte Carlo for Marginal Density Estimation with an Application to a Spatial Statistical Model." *Journal of Computational and Graphical Statistics*, **16**(3):608–632, 2007.

[Lia10] Faming Liang. "A Double Metropolis–Hastings Sampler for Spatial Models with Intractable Normalizing Constants." *Journal of Statistical Computation and Simulation*, **80**(9):1007–1022, 2010.

[LJ13] Faming Liang and Ick Hoon Jin. "A Monte Carlo Metropolis–Hastings Algorithm for Sampling from Distributions with Intractable Normalizing Constants." *Neural Computation*, **25**(8):2199–2234, 2013.

[LJS16] Faming Liang, Ick Hoon Jin, Qifan Song, and Jun S. Liu. "An Adaptive Exchange Algorithm for Sampling from Distributions with Intractable Normalizing Constants." *Journal of the American Statistical Association*, **111**(513):377–393, 2016.

[LVM18] Andrey Y. Lokhov, Marc Vuffray, Sidhant Misra, and Michael Chertkov. "Optimal Structure and Parameter Learning of Ising Models." *Proceedings of the National Academy of Sciences*, **115**(7):E1495–E1503, 2018.

[Mel19] Angelo Mele. "mfergm: a package for variational approximations of (some) ergm models.", 2019.

[MGM06] Iain Murray, Zoubin Ghahramani, and David J. C. MacKay. "MCMC for Doubly-Intractable Distributions." In Rina Dechter and Thomas Richardson, editors, *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pp. 359–366, Cambridge, MA, USA, 2006. AUAI Press.

[MHH08] Martina Morris, Mark S. Handcock, and David R. Hunter. "Specification of Exponential-family Random Graph Models: Terms and Computational Aspects." *Journal of Statistical Software*, **24**(4), 2008.

[MPR06] Jesper Møller, Anthony N. Pettitt, Richard Reeves, and Kasper K. Berthelsen. "An Efficient Markov Chain Monte Carlo Method for Distributions with Intractable Normalising Constants." *Biometrika*, **93**(2):451–458, 2006.

[MZ23] Angelo Mele and Lingjiong Zhu. "Approximate Variational Estimation for a Model of Network Formation." *The Review of Economics and Statistics*, pp. 1–30, 03 2023.

[Pol19] Johannes van der Pol. "Introduction to Network Modeling Using Exponential Random Graph Models (ERGM): Theory and an Application Using R-Project." *Computational Economics*, **54**(3):845–875, 2019.

[RDevelopmentCoreTeam12] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[Res24] Joseph Resch. "Likelihood Inference Comparisons for Exponential Random Graph Models.", 2024.

[RH25] Joseph Resch and Mark S. Handcock. "Assessing Variational Likelihood Estimation in Network Formation Models: Code Repository.", 11 2025.

[Ric12] Federico Ricci-Tersenghi. "The Bethe Approximation for Solving the Inverse Ising Problem: A Comparison with Other Inversion Methods." *Journal of Statistical Mechanics: Theory and Experiment*, (08):P08015, 2012.

[RM51] Herbert Robbins and Sutton Monro. "A Stochastic Approximation Method." *The Annals of Mathematical Statistics*, **22**(3):400–407, 1951.

[RR98] Gareth O. Roberts and Jeffrey S. Rosenthal. "Optimal Scaling of Discrete Approximations to Langevin Diffusions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **60**(1):255–268, 1998.

[SBF19] Julien Stoehr, Alan Benson, and Nial Friel. "Noisy Hamiltonian Monte Carlo for doubly intractable distributions." *Journal of Computational and Graphical Statistics*, **28**(1):220–232, 2019.

[Sch11] Michael Schweinberger. "Instability, sensitivity, and degeneracy of discrete exponential families." *Journal of the American Statistical Association*, **106**(496):1361–1370, 2011.

[SKB20] Michael Schweinberger, Pavel N. Krivitsky, Carter T. Butts, and Jonathan R. Stewart. "Exponential-Family Models of Random Graphs: Inference in Finite, Super and Infinite Population Scenarios." *Statistical Science*, **35**(4):627 – 662, 2020.

[Sni02] Tom AB Snijders. "Markov chain Monte Carlo estimation of exponential random graph models." *Journal of Social Structure*, **3**(2):1–40, 2002.

[SPR06] Tom AB Snijders, Philippa E Pattison, Garry L Robins, and Mark S Handcock. "New specifications for exponential random graph models." *Sociological Methodology*, **36**(1):99–153, 2006.

[TF20] Linda S. L. Tan and Nial Friel. "Bayesian Variational Inference for Exponential Random Graph Models." *Journal of Computational and Graphical Statistics*, **29**(1):38–52, 2020.

[TTV16] Yee Whye Teh, Alexandre H. Thiery, and Sebastian J. Vollmer. "Consistency and Fluctuations for Stochastic Gradient Langevin Dynamics." *Journal of Machine Learning Research*, **17**(7):1–33, 2016.

[WT11] Max Welling and Yee Whye Teh. "Bayesian Learning via Stochastic Gradient Langevin Dynamics." In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 681–688. Omnipress, 2011.

[ZL24] Qian Zhang and Faming Liang. "Bayesian Analysis of Exponential Random Graph Models Using Stochastic Gradient Markov Chain Monte Carlo." *Bayesian Analysis*, **19**(2):595–621, 2024.