Removing Phase Transitions from Gibbs Measures

Ian E. Fellows
Fellows Statistics

Mark S. Handcock
University of California, Los Angeles

Abstract

Gibbs measures are a fundamental class of distributions for the analysis of high dimensional data. Phase transitions, which are also known as degeneracy in the network science literature, are an emergent property of these models that well describe many physical systems. However, the reach of the Gibbs measure is now far outside the realm of physical systems, and in many of these domains multiphase behavior is a nuisance. This nuisance often makes distribution fitting impossible due to failure of the MCMC sampler, and even when an MLE fit is possible, if the solution is near a phase transition point, the plausibility of the fit can be highly questionable. We introduce a modification to the Gibbs distribution that reduces the effects of phase transitions, and with properly chosen hyper-parameters, provably removes all multiphase behavior. We show that this new distribution is just as easy to fit via MCMC-MLE as the Gibbs measure, and provide examples in the Ising model from statistical physics and ERGMs from network science.

1 Introduction

The Gibbs measure is a class of probability distributions describing the configuration of a (usually high dimensional) set of random variables. Initially it was motivated primarily by statistical mechanics [Gibbs et al., 1902], though in more recent decades it has found widespread usage in many areas of statistics and machine learning, including computer vision [Li, 2012], neuroscience [Stephens et al., 2011] and the modeling of social networks [Frank and Strauss, 1986] to name a few.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

While the Gibbs measure has shown wide applicability as a general modeling framework, the presence of multiphase behavior has caused some practical difficulty. Nowhere is this more apparent than in the network science field, where the popular Exponential-family Random Graph Model (ERGM) has been repeatedly shown to suffer from model frailty due to the existence of multiphase parameter configurations. This is known as "model degeneracy" in the network science literature, and is perhaps the most fundamental challenge facing ERGMs both in theory [Schweinberger, 2011, Handcock, 2003, Bhamidi et al., 2008, Chatterjee et al., 2013] and practice [Robins et al., 2007, Goldenberg et al., 2010].

The contribution of this work is to introduce a new class of statistical models for high dimensional data that retains the flexibility of the Gibbs measure, without the undesirable multiphase properties. The form of this model class allows us to ensure that it is not multiphase at any parameter values, and that the likelihood puts significant probability mass around the configuration that is being modeled.

Let X be a set of random variables with realization x and $\mathcal{N}$ be the sample space in which X may exist. A Gibbs measure is defined by the probability mass function

$$p_{\text{gibbs}}(x|\theta) = \frac{1}{Z(\theta)} e^{\sum_k \theta_k g_k(x)} \qquad x \in \mathcal{N}$$

where θ is a vector of d parameters, g is a vector valued set of sufficient statistics, and $Z(\theta) = \sum_x e^{\sum_i \theta_i g_i(x)}$ is the partition function assuring that the probabilities sum to unity. For notational convenience we will be assuming throughout that x is discrete. The continuous case may be handled with ease by replacing sums over x by integrals.

Since the likelihood is only a function of the sufficient statistics, the probability of observing a set of statistics is

$$p_{\text{gibbs}}(g(x)|\theta) = N(g(x)) \frac{1}{Z(\theta)} e^{\sum_k \theta_k g_k(x)} \qquad x \in \mathcal{N},$$

where $N(g(x))$ is the count of configurations with

statistics equal to $g(x)$.

p_{gibbs} is an exponential family distribution [Barndorff-Nielsen, 1978] with natural parameters θ. An alternate way to define the model is in terms of the mean value parameters, which are defined as the expected values of the sufficient statistics $\mu(\theta) = E_\theta(g(X))$. There is a one-to-one relationship between the natural and mean value parameters, so a model may be expressed in either terms [Barndorff-Nielsen, 1978].

One of the most attractive features of p_{gibbs} as a class of general modeling distributions is that it satisfies the maximum entropy principle, in that among all possible probability distributions (p') for X with identical mean values (i.e. $E_{p'}(g(X)) = \mu(\theta)$), p_{gibbs} has the largest entropy [Jaynes, 1957]. Because entropy is a measure of the informativeness of the distribution, the Gibbs measure may be thought of as the least informative distribution consistent with the mean value parameters.

Performing likelihood-based inference on Gibbs measures is a non-trivial task because for even moderately multivariate X, the sum in the normalizing constant Z is typically computationally intractable, and thus it is impossible to evaluate the likelihood directly. However, indirect sampling methods may be used by expressing the likelihood and its derivatives in terms of expectations and covariances. The log likelihood of the distribution is

$$\ell_{\text{gibbs}}(\theta|x) = \sum_k \theta_k g_k(x) - \log(Z(\theta)),$$

and the first and second derivatives may then be found to be

$$\frac{\partial \ell_{\text{gibbs}}}{\partial \theta_i} = g_i(x) - \mu_i(\theta)$$

and

$$\frac{\partial^2 \ell_{\text{gibbs}}}{\partial \theta_i \partial \theta_j} = -\text{cov}(g_i(X), g_j(X)).$$

It is clear from setting the first derivative to zero that at the maximum likelihood estimate (MLE) of θ, the mean value parameters must match the observed sufficient statistics (i.e. $g(x) = \mu(\hat{\theta}_{\text{mle}})$).

Since p_{gibbs} is known up to a constant, we may use Markov Chain Monte Carlo (MCMC) methods to generate a sample from it. Using this sample, the above equations may be approximated using weighted sample means and covariances in place of the population expectations and covariances. The precise mechanics of using these approximations to find the maximum likelihood estimate of θ is known as Markov Chain Monte Carlo Maximum Likelihood Estimation (MCMCMLE), and has been used successfully across many disparate types of Gibbs measures [Geyer and Thompson, 1992, Hunter and Handcock, 2012, Descombes et al., 1997].

1.1 Example: The Ising Model

A canonical example of a Gibbs measure is the Ising [1925] model of ferromagnetism, which in this case we will define over an n-by-n toroidal lattice. Each x_{ij} may either have magnetic spin up or down ($x_{ij} \in \{-1, 1\} \ \forall \ i, j \in \{1, ..., n\}$), and is influenced by its four neighbors $x_{(i-1)j}$, $x_{(i+1)j}$, $x_{i(j-1)}$, and $x_{i(j+1)}$. Here we allow the subscripts to wrap, such that a subscript of 0 refers to index n and a subscript of $n+1$ refers to index 1. The Ising probability model has two sufficient statistics. The first models the general tendency for up or down spins and is defined as $g_1(x) = \sum_{i,j} x_{ij}$. The second is the number of neighbors sharing the same spin and is defined as $g_2(x) = \sum_{i,j} x_{ij}(x_{(i+1)j} + x_{i(j+1)})$.

While the Ising model was developed in the context of statistical physics, the tendency of like to be connected to like is a more universal concept, and thus the model has found extensive use in domains far afield from magnetism. To name a few examples, it is used in computer vision for image denoising [Cohen et al., 2015], in network modeling the classic two-star model can be viewed as an Ising model [Palla et al., 2004, Park and Newman, 2004], and in sociology it is useful in understanding group choice [Galam, 1997] and urban segregation [Stauffer, 2008].

2 Phases In Gibbs Measures

The Gibbs distribution captures many of the characteristics of typical physical systems, even those which in other domains may lead to computational difficulties and implausible models. In particular, Gibbs measures often exhibit the phenomena of a phase transition point. This has been extensively studied in the physics literature [Georgii, 2011]. In the context of social network modeling, multiphase behavior has resulted in significant difficulty in finding good models for large complex networks [Handcock, 2003, Schweinberger, 2011, Chatterjee et al., 2013, Horvát et al., 2015].

Consider an Ising model where $\theta_1 = 0$ and θ_2 is very large. If one were to start an MCMC chain at a configuration where most vertices are set to 1. At equilibrium we would find almost perfect alignment to 1. Similarly, if the starting point for the chain was mostly -1 then the equilibrium state would be a global -1 alignment.

Such behavior has important implications for prac-

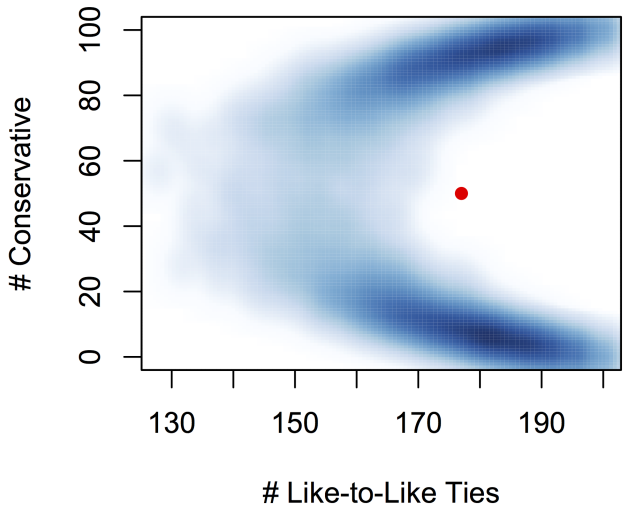

Figure 1: Simulations from the Ising model at the MLE with $\hat{\theta}_{\text{mle}} = (0, 0.45)$. The observed configuration is marked as a red point. Note that the configurations similar to the observed are extremely uncommon under the maximum likelihood model.

tice. The extreme multimodal nature of the likelihood dramatically reduced the computational efficiency of MCMC algorithms such as Metropolis-Hastings. This has restricted their routine usage in MCMC-based maximum likelihood estimation [Handcock, 2003]. Secondly, if the true value is close to a phase transition point and the MLE was found, the model fit would typically not be useful [Chatterjee et al., 2013].

Suppose instead of magnetic spins, the Ising model indicated "liberal" or "conservative" political affiliation and $n = 100$. Imagine that we observe a community where exactly half of people are conservative and there is a very strong tendency for like political affiliation to be connected to like, with 177 of the 200 connections being between the same affiliations. This can be considered a simple version of the social influence model introduced in Robins et al. [2001]. Recall that at the MLE, the mean value parameters equal their observed values, so average simulations from the fit will consist of 50% conservatives. However, as shown in Figure 1, the MLE has two phases, one where almost all are conservative and one where almost all are liberal. Even though the average of these two modes matches the observed value, configurations like the one observed are extremely rare. Thus the fit of the Ising model is at best highly questionable.

Let N^* be a smoothed twice differentiable approximation of N, and $p^*(g(x)) = \frac{N^*(g(x))}{N(g(x))} p(g(x))$ be an approximation of the exponential family probability distribution p. This smoothed version of N^*, introduced by Horvát et al. [2015], is a mathematical convenience removing the effects of small local fluctuations and assuring the existence of a Hessian. While the choice of N* could result in a probability distribution, this is not a hard requirement of its construction, and in many cases, N* could be made to be arbitrarily close to N while still maintaining 2^{nd} order differentiability. Following Horvát et al. [2015] we define uniphase and multiphase as:

Definition 1 *A distribution $p(g(x))$ is uniphase with respect to p^* if it contains no local minima or saddle points, and is multiphase otherwise.*

Horvát et al. [2015] developed an insightful connection between multiphase behavior and the configuration count function N. We restate two results from that work here.

Lemma 1 *Let f be a continuous twice differentiable function and $h(y) = f(y)e^{\theta \cdot y}$, where y is a vector. h has no minima or saddlepoints in y for all θ if and only if f is strictly log concave.*

Theorem 1 *A Gibbs measure $p_{gibbs}(g(x)|\theta)$ is uniphase with respect to smoothing p^*_{gibbs} for all θ, if and only if N^* is strictly log concave.*

Theorem 1 reduces the existence of multiphase parameter values to a statement about configuration density, and motivates using sufficient statistics in the model such that N is approximately log concave. However, as a practical insight it is difficult to use. In most cases, due to the high dimensional nature of x, N is intractable and is therefore difficult to compute.

3 A Robust Extension of the Gibbs Measure

In order to reduce the appearance of distributional modes that are far away from one another in terms of the sufficient statistics, we will use the same maximum entropy framework that underlies Gibb's measures, but add an additional constraint that the variance of the sufficient statistics should be less than or equal to a maximum value. The maximum entropy problem is

$$\begin{aligned} \underset{q}{\text{maximize}} \quad & \sum_x q(x) \log(q(x)) \\ \text{subject to} \quad & \sum_x q(x) = 1, \\ & E_q(g_i(X)) = \mu_i, \\ & E_q((\mu_i - g_i(X))^2) \leq \kappa_i, \\ & \forall\, i \in \{1, \ldots, d\}. \end{aligned}$$

where $q(x) \geq 0$ is a probability mass function. Reformulating the optimization problem using Karush-Kuhn-Tucker multipliers into an unconditional minimization yields an objective function of

$$J(q) = \sum_x \Bigg(- q(x)\log(q(x)) + \theta_0 q(x) + \sum_k^d \theta_k g_k(x) q(x) - \sum_k^d \beta_k^{-2}(\mu_k - g_k(x))^2 q(x) \Bigg)$$

where all $\beta_k^{-2} > 0$. We may then apply the calculus of variations on the functional J. Noting that the constraints are constants and thus $\partial\mu/\partial q = 0$, finding the root of the derivative of the term inside the sum results in

$$\begin{aligned} 0 &= \frac{\partial}{\partial q}\Bigg(- q(x)\log(q(x)) + \theta_0 q(x) + \sum_k^d \theta_k g_k(x) q(x) - \sum_k^d \beta_k^{-2}(\mu_k - g_k(x))^2 q(x) \Bigg) \\ &= -1 - \log(q(x)) + \theta_0 + \sum_k^d \theta_k g_k(x) - \sum_k^d \beta_k^{-2}(\mu_k - g_k(x))^2. \end{aligned}$$

Rearranging terms and absorbing θ_0 into the partition function, we arrive at the maximum entropy distribution subject to variation constraints

$$q(x|\theta,\beta) = \frac{1}{Z(\theta,\beta)} e^{\sum_k \theta_k g_k(x) - \sum_k \beta_k^{-2}(\mu_k(\theta,\beta) - g_k(x))^2}, \tag{1}$$

where $\mu(\theta,\beta) = E_q(g(x))$ makes the dependence of the mean value parameters of θ and β explicit. A related distributional family where the squared deviation term is centered around a general m is

$$q'(x|\theta,\beta,m) = \frac{1}{Z(\theta,\beta,m)} e^{\sum_k \theta_k g_k(x) - \sum_k \beta_k^{-2}(m_k - g_k(x))^2}. \tag{2}$$

We call these distributions *tapered Gibbs measures* and recognize that q and q' are similar to the Gibbs measure, but have a term attached tapering the likelihood of configurations far from the mode of the term (μ and m respectively). Each β_k may be interpreted as the strength of the attractive force to the tapering term's mode for statistic g_k, with a reduction of 1 being applied to the log likelihood when a configuration's sufficient statistic is β_k units away from the mode and increasing quadratically thereafter. We may also note that q is subclass of q' in that

$$q(x|\theta,\beta) = q'(x|\theta,\beta,m = E_{q'}(g(X))).$$

Both q and q' are exponential family, and because m does not depend on θ, q' can be seen as a Gibbs measure with an additional offset term reducing the entropy.

Theorem 2 *For any m, there exists a vector α such that for any β satisfying $0 < \beta_i < \alpha_i \ \ \forall\, i \in \{1,...,d\}$, q and q' are uniphase for all θ with respect to smoothing q^* and q'^*.*

Proof. For q' with any particular m, taking the tapering term and N^* together and applying Lemma 1 we see that the theorem is true if and only if $\log(N^*(g(x))) - \sum_k \beta_k^{-2}(m_k - g_k(x))^2$ is strictly concave, or equivalently that its Hessian is negative definite. We have no control over the Hessian of the first term, however the Hessian of the second is a diagonal matrix with elements β_i^{-2}. By choosing all α_i^{-2} to be greater than the largest eigenvalue of $\nabla^2 \log(N^*(g(x)))$ we achieve a negative definite result for the whole term. The result for q follows as it is a subclass of q'.

■

It is certainly possible to entertain other functional forms of the tapering term that down weight extreme values. However, the proof of Theorem 2 shows that our quadratic form is a natural one, as its Hessian depends only on β and not m or g.

Theorem 2 guarantees that for small enough β, q is uniphase; however, the computational complexity of N makes it difficult to check directly if a particular β yields a multiphase distribution. From a practical standpoint, one can find a good uniphase value for β by drawing MCMC samples from different values for β and inspecting the histograms of the sufficient statistics for multimodality.

4 Estimation and Inference

Because q' simply involved adding on an offset term to a Gibbs measure, maximum likelihood inference for θ is unchanged. The first and second derivatives are

$$\frac{\partial \ell_{q'}}{\partial \theta_i} = g_i(x) - \mu_i(\theta,\beta,m)$$

and

$$\frac{\partial^2 \ell_{q'}}{\partial \theta_i \partial \theta_j} = -\mathrm{cov}(g_i(X), g_j(X)).$$

As in the Gibbs measure case, we may sample from q' using MCMC and approximate the population expectations using sample expectations. Thus, MCMCMLE may be performed to find the MLE. q' may not be a particularly compelling class of distributions depending on the choice of m. Ideally we would like to counteract the multiphase tendencies of the Gibbs measure

and have a mode centered around the mean value parameters. So, q is a more appropriate distribution for general modeling.

One must know both θ and μ in order to be able to sample from q, which on its face make the prospect of MCMCMLE inference seem unlikely. We can side step this problem using q'. First, we establish some properties of q. The log likelihood is

$$\ell_q(\theta|x,\beta) = \sum_i \left(\theta_i g_i(x) - \beta_i^{-2}(\mu_i(\theta,\beta) - g_i(x))^2\right) - \log(Z(\theta,\beta)).$$

In order to simplify notation, we define the derivative with respect to θ_i of the log of the numerator of the likelihood as

$$t_i(x,\theta,\beta) = g_i(x) - \sum_k 2\beta_k^{-2}\frac{\partial \mu_k(\theta,\beta)}{\partial \theta_i}(\mu_k(\theta,\beta) - g_k(x)).$$

The derivatives of the mean value parameters are

$$\begin{aligned}\frac{\partial \mu_r(\theta,\beta)}{\partial \theta_i} &= \operatorname{cov}(g_r(X), t_i(X,\theta,\beta)) \\ &= \operatorname{cov}(g_r(X), g_i(X)) \\ &\quad - \sum_k 2\beta_k^{-2}\frac{\partial \mu_k(\theta,\beta)}{\partial \theta_i}\operatorname{cov}(g_r(X), g_k(X)).\end{aligned}$$

The derivative of μ is on both the right and left hand side of this equation, so to solve this linear system of equations, define a matrix B and vectors c^i with elements $B_{rk} = 2\beta_k^{-2}\operatorname{cov}(g_r(X), g_k(X))$ and $c^i_r = \operatorname{cov}(g_r(X), g_i(X))$. The derivatives of the mean value parameters are then

$$\frac{\partial \mu(\theta,\beta)}{\partial \theta_i} = (B + I)^{-1}c^i,$$

where I is the identity matrix.

The first derivative of the log likelihood is

$$\begin{aligned}\frac{\partial \ell_q}{\partial \theta_i} &= t_i(x,\theta,\beta) - E(t_i(X,\theta,\beta)) \\ &= g_i(x) - \mu_i(\theta,\beta) \\ &\quad - 2\sum_k \beta_k^{-2}\frac{\partial \mu_k(\theta,\beta)}{\partial \theta_i}\left(\mu_k(\theta,\beta) - g_k(x)\right) \quad (3)\end{aligned}$$

Setting Equation 3 to 0, we see a solution under the usual moment conditions of $g(x) = \mu(\hat{\theta}_{\text{mle}}, \beta)$. The second derivative of the log likelihood is

$$\begin{aligned}\frac{\partial \ell_q}{\partial \theta_i \partial \theta_j} &= -\frac{\partial \mu_i(\theta,\beta)}{\partial \theta_j} \\ &\quad - 2\sum_k \beta_k^{-2}\frac{\partial^2 \mu_k(\theta,\beta)}{\partial \theta_i \partial \theta_j}\left(\mu_k(\theta,\beta) - g_k(x)\right) \\ &\quad - 2\sum_k \beta_k^{-2}\frac{\partial \mu_k(\theta,\beta)}{\partial \theta_i}\frac{\partial \mu_k(\theta,\beta)}{\partial \theta_j}.\end{aligned}$$

At the MLE this simplifies to

$$\left.\frac{\partial \ell}{\partial \theta_i \partial \theta_j}\right|_{\hat{\theta}_{\text{mle}}} = -\frac{\partial \mu_i(\theta,\beta)}{\partial \theta_j} - 2\sum_k \beta_k^{-2}\frac{\partial \mu_k(\theta,\beta)}{\partial \theta_i}\frac{\partial \mu_k(\theta,\beta)}{\partial \theta_j}. \quad (4)$$

4.1 Inference for q

Using the fact that at the MLE of q, $g(x) = \mu(\theta,\beta)$, we can use MCMCMLE inference on q' to find the MLE of q. Noting that

$$\begin{aligned}q'(x|\hat{\theta}_{\text{mle}}, \beta, m = g(x)) &= q'(x|\hat{\theta}_{\text{mle}}, \beta, m = E_{q'}(g(X))) \\ &= q(x|\hat{\theta}_{\text{mle}}, \beta),\end{aligned}$$

finding the MLE of q reduces to finding the MLE of q' setting $m = g(x)$. At this point, both the natural ($\hat{\theta}_{\text{mle}}$) and mean value parameters ($\mu(\hat{\theta}_{\text{mle}}, \beta) = g(x)$) are known and thus q may be sampled from using MCMC.

Unfortunately, there is no information in the likelihood that can be used to estimate β as $\frac{\partial \ell_q}{\partial \beta} = 0$ at $\hat{\theta}_{\text{mle}}$, and choosing a β that yields a uniphase distribution and sufficiently realistic stochastic variation is likely domain dependent. That said, β_k causes only a small adjustment to the log likelihood (< 1) when $|\mu_k - g_k(x)| < \beta_k$, so if one has a target variance v_k for g_k then choosing $\beta = r\sqrt{(v_k)}$ will ensure that only observations more than r "standard deviations" away from the mean receive large tapering adjustments. This strategy is employed in the example of Section 4.3 with $r = 2$.

4.2 Example: The Ising Model Revisited

One plausible theory as to why the Ising model in Section 2 failed to represent the data well is that the model only considered local neighbor-to-neighbor interactions and ignored the possibility of more global cohesive effects. When making political affiliation decisions, actors are influenced by their neighbors, but also by larger scale media, which may be targeted at a more average individual, perhaps tempering more radical philosophies and pulling those actors toward the

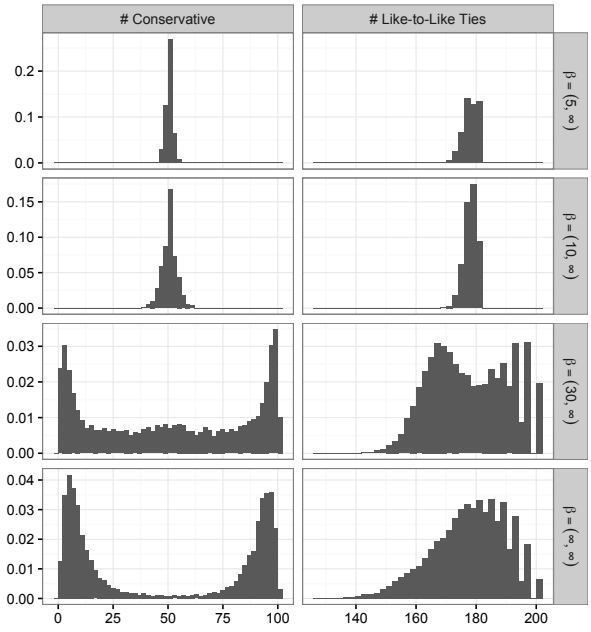

Figure 2: Histogram density estimates based on 10,000 simulations at each β value from the MLE fit to the configuration with 50 conservatives and 177 like affiliation ties. Smaller values of β remove the bimodality of the phase transition.

average position. This tendency to pull of actors to the average is precisely the effect of the tapering term in Equation 1.

Figure 2 shows simulations from the MLE model fit of q to our hypothetical community with varying degrees of tapering. No tapering was added to g_2 as $\beta_2 = \infty$ in all simulations. When both βs are infinite, the model reduces to a Gibbs measure. At $\beta_1 = 30$ we start to see increasing density around the mean value statistic, but the two modes are still prominent. At $\beta_1 = 10$, configurations with 55 or 45 conservatives receive a log probability reduction of 1, and the ones with 0 or 100 conservatives receive a log probability reduction of 100. The distribution of counts of conservatives is symmetric and looks nearly Gaussian. $\beta_1 = 5$ has a qualitatively similar look to that of $\beta_1 = 10$, except the variation in the counts has been reduced.

With no tapering term, we see a high variation in the counts of conservatives (g_1). A priori, an analyst might have expected a level of variation similar to a binomial distribution, which would be $v_1 = 100 * 0.5 * (1 - 0.5) = 25$. Using $r = 2$ yields $\beta_1 = 2\sqrt{25} = 10$ and at $\beta_1 = 10$, the MLE displays none of the bimodality present in the untapered model.

4.3 Example: A Network Model for Network Science Collaboration

Exponential-family random graph models (ERGM) represent a very general and flexible class of distributions for modeling relational ties. A quite extensive literature on these models exists, see for example Robins et al. [2007] and references therein. An ERGM is a Gibbs measure where elements of x indicate the present or absence of a relational tie between individuals. Node i is tied to node j if $x_{ij} = 1$ and $x_{ij} = 0$ otherwise.

What makes ERGMs different from any other Gibbs Measure is the choice of sufficient statistics to use in modeling. These statistics are chosen to match our theoretical understanding of the forces governing social interactions, six of which are described here.

edges The number of edges in the graph.

k-star The number of subgraphs where k edges are connected to a single vertex.

isolates The number of vertices with no neighbors.

triangles The number of triangles present in the graph.

k-esp The number of edges whose vertices share exactly k neighbors in common.

dcp The cross-product of the degrees of connected vertices divided by the number of total edges.

The first three terms ("edges," "k-star" and "isolates") model the distribution of the number of neighbors of the vertices in the network. This is known as the degree distribution. The next two terms "k-esp" and "triangles" model the transitivity of relations. If i is connected to j and k, then j and k may be more likely to be connected. The final term "dcp" models tendencies of high degree vertices to be connected to other high degree vertices.

The above terms have been repeatedly shown to have extreme difficulties in modeling real world social networks due to multiphase behavior [Schweinberger, 2011, Chatterjee et al., 2013, Handcock, 2003], especially in networks of any significant size. However, because we now have the tools to remove phase transitions, it is possible to model networks never before amenable to ERGM analysis.

Newman [2006] collected and introduced a network representing the co-authorship of scientists publishing on the topic of network science. The full network contains 1589 scientists, of which we removed one outlying biology paper with an extreme number of authors

Term	$g(x)$	$\hat{sd}(g(X))$	$\hat{\theta}$	$\hat{se}(\hat{\theta})$
edges	2555	21.73	-4.07	0.22
2-star	12816	152.52	0.11	0.02
3-star	38035	277.11	-0.0005	0.004
isolates	128	8.73	-4.00	0.17
0-esp	221	9.73	-8.32	0.21
1-esp	503	20.76	-3.49	0.13
2-esp	534	25.23	-1.79	0.10
3-esp	458	24.33	-0.94	0.09
4-esp	262	18.83	-0.50	0.08
5-esp	82	9.62	-0.66	0.12
triangles	2625	49.38	0.63	0.06
dcp	41.22	0.95	-21.60	2.90

Table 1: Parameter and summary statistics for an MCMCMLE fit of network science collaborations.

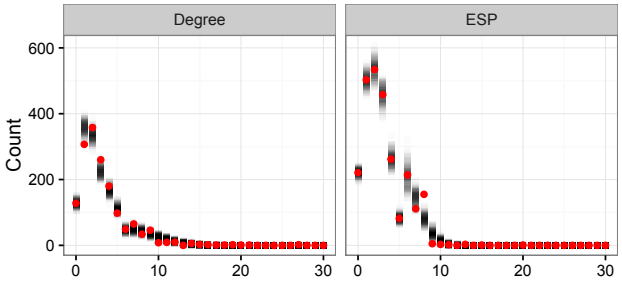

Figure 3: The degree and ESP distributions from 200 simulated networks (black) from the MLE compared to the observed network (red).

[Giot et al., 2003]. This resulted in an undirected network with 1572 vertices and 2555 edges.

Under a completely random graph, we would expect the distribution of sub-graph counts to be distributed approximately Poisson, and thus they have variance equal to their expectation. Using the logic of Section 4.1 we set a target variance equal to the expectation of the sufficient statistics at the MLE and $r = 2$, which yields $\beta = 2\sqrt{E_{\hat{q}}(g(X))} = 2\sqrt{g(x)}$.

Table 1 shows model summaries for the MLE fit. The standard deviations of the sufficient statistics ($\hat{sd}(g(X))$) were estimated by drawing MCMC samples from the fit distribution. $\hat{\theta}$ are the MLE parameter estimates and $\hat{se}(\hat{\theta})$ are the standard errors of the estimates. The standard errors were, as per standard practice in ERGMs, calculated by inverting the estimated observed Fisher information (see Equation 4).

The negative isolates term shows us that there are fewer authors who collaborate with no-one than we'd expect by chance. The positive 2-star term tells use that the degree distribution has higher spread than would be expected by chance, and the small (insignificant) 3-star term indicates that it is not overly skewed. All of the "esp" and "triangle" terms are very significant indicating that there is much more transitivity than if collaboration ties were formed randomly. Finally the highly negative "dcp" term suggests that high degree individuals tend to collaborate with low degree individuals. This makes intuitive sense, as senior researchers tend to pair with students and junior faculty in their writings.

Figure 3 shows a goodness-of-fit plot for the full degree and ESP distributions. We see good agreement between the networks simulated from the model and the observed graph. Figure 4 displays the observed graph alongside three simulated networks. The simulated networks share many similarities with the observed graph, including a large component, many triangles, and a hub-and-spoke pattern caused by high degree nodes being connected to lower degree nodes.

Fitting this model without using the tapering term proved fruitless. All attempts yielded an MCMCMLE process that diverged either to the empty or completely full graph, suggesting that a multiphase point had been reached. Utilizing the "degeneracy robust" statistics of Snijders et al. [2006] in place of the more intuitive statistics we used here similarly led to a divergent MCMC process.

5 Discussion

Phase transitions are an interesting emergent phenomena in Gibbs measures. In many physical systems, this closely matches the reality of distinct transitions between magnetic polarity or liquids and solids. However, in a great many cases, phase transitions are an unwanted anomaly. One that not only stymies computational efforts to find a maximum likelihood fit, but also renders that fit an unrealistic representation of the underlying phenomena.

Researchers have focused much effort on engineering sufficient statistics that are less prone to phase transitions, with some success [Snijders et al., 2006, Horvát et al., 2015]; however recent theoretical work suggests that as networks become larger, phase transition behavior may become more and more common [Chatterjee et al., 2013]. Further, while there is an art to engineering features, there is no theoretical result showing that statistics stable in one network will be stable in another.

The probability model developed here represents a new perspective on the problem. By augmenting the distribution and down-weighting extreme configurations, we achieve a model that not only shows good behavior in practice, but has theoretical guarantees regarding

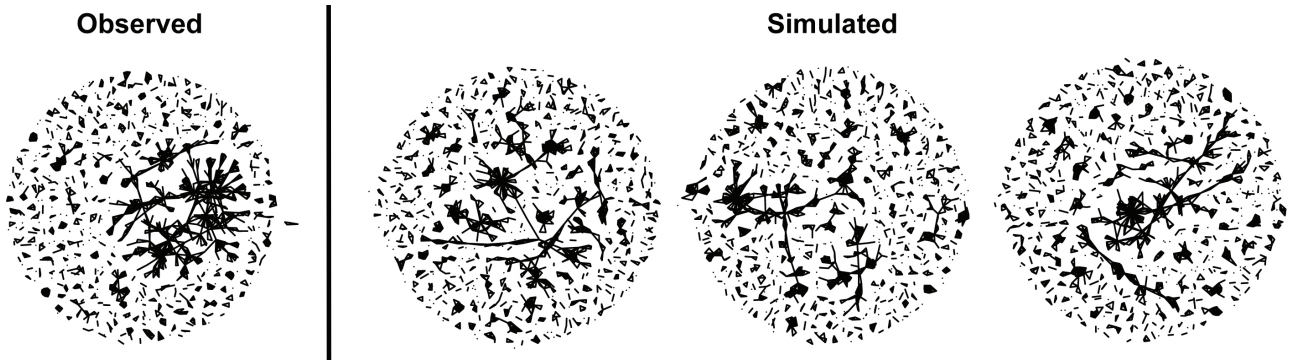

Figure 4: The observed network science collaboration network and three independent networks simulated from the MLE.

its stability. Using these methods frees the researcher to utilize sufficient statistics that are of theoretical interest, improving the interpretability of their results.

Using the rule $\beta_k = r\sqrt{v_k}$ with $r = 2$ provided non-degenerate fits in both the Ising (Section 4.2) and network science (Section 4.3) analyses. However, choice of β remains an important open area for future work.

Acknowledgement

This work was funded by NIH grant R21HD075714 and NSF grants MMS-0851555 and SES-1357619

References

J Willard Gibbs et al. *Elementary principles in statistical mechanics*. C. Scribner's sons;[etc., etc.], 1902.

Stan Z Li. *Markov random field modeling in computer vision*. Springer Science & Business Media, 2012.

Greg J Stephens, Leslie C Osborne, and William Bialek. Searching for simplicity in the analysis of neurons and behavior. *Proceedings of the National Academy of Sciences*, 108(Supplement 3): 15565–15571, 2011.

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

Michael Schweinberger. Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, 106(496): 1361–1370, 2011. ISSN 01621459.

Mark S. Handcock. Statistical models for social networks: Inference and degeneracy. In R. Breiger, K. Carley, and P. Pattison, editors, *Dynamic Social Network Modeling and Analysis*, volume 126, pages 302–322. Committee on Human Factors, Board on Behavioral, Cognitive, and Sensory Sciences, National Academy Press, Washington, DC., 2003.

Shankar Bhamidi, Guy Bresler, and Allan Sly. Mixing time of exponential random graphs. In *Foundations of Computer Science, 2008. FOCS'08. IEEE 49th Annual IEEE Symposium on*, pages 803–812. IEEE, 2008.

Sourav Chatterjee, Persi Diaconis, et al. Estimating and understanding exponential random graph models. *The Annals of Statistics*, 41(5):2428–2461, 2013.

Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2):173–191, 2007.

Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, and Edoardo M Airoldi. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

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

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Charles Geyer and Elizabeth Thompson. Constrained Monte Carlo Maximum Likelihood for Dependent Data. *Journal of the Royal Statistical Society. Series B*, 54(3):657–699, 1992.

David R Hunter and Mark S Handcock. Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 2012.

Xavier Descombes, Robin Morris, Josiane Zerubia, and Marc Berthod. Maximum likelihood estimation of markov random field parameters using markov chain monte carlo algorithms. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 133–148. Springer, 1997.

Ernst Ising. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift fur Physik*, 31:253–258, February 1925.

Eliahu Cohen, Ron Heiman, Maya Carmi, Ofer Hadar, and Asaf Cohen. When physics meets signal processing: Image and video denoising based on ising theory. *Signal Processing: Image Communication*, 34:14–21, 2015.

Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Statistical mechanics of topological phase transitions in networks. *Physical Review E*, 69(4): 046117, 2004.

Juyong Park and M. E. J. Newman. Solution of the two-star model of a network. *Physical Review E*, 70 (6):066146, 2004.

Serge Galam. Rational group decision making: A random field ising model at t= 0. *Physica A: Statistical Mechanics and its Applications*, 238(1):66–80, 1997.

Dietrich Stauffer. Social applications of two-dimensional ising models. *American Journal of Physics*, 76(4):470–473, 2008.

Hans-Otto Georgii. *Gibbs measures and phase transitions*, volume 9. Walter de Gruyter, 2011.

Szabolcs Horvát, Éva Czabarka, and Zoltán Toroczkai. Reducing degeneracy in maximum entropy models of networks. *Physical review letters*, 114(15):158701, 2015.

Garry Robins, Phillippa Pattison, and Peter Elliott. Network models for social influence processes. *Psychometrica*, 66(2):161–190, 2001.

M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.

Loic Giot, Joel S Bader, Cory Brouwer, Amitabha Chaudhuri, Bing Kuang, Y Li, YL Hao, CE Ooi, Brian Godwin, E Vitols, et al. A protein interaction map of drosophila melanogaster. *science*, 302(5651): 1727–1736, 2003.

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.