Reducing Degeneracy in Maximum Entropy Models of Networks

Szabolcs Horvát,[1] Éva Czabarka,[2] and Zoltán Toroczkai[1,3,*]

[1]*Department of Physics and Interdisciplinary Center for Network Science & Applications, University of Notre Dame, Notre Dame, Indiana 46556, USA*

[2]*Department of Mathematics, University of South Carolina, Columbia, South Carolina 29208, USA*

[3]*Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, Indiana 46556 USA*

(Received 2 July 2014; published 14 April 2015)

Based on Jaynes's maximum entropy principle, exponential random graphs provide a family of principled models that allow the prediction of network properties as constrained by empirical data (observables). However, their use is often hindered by the degeneracy problem characterized by spontaneous symmetry breaking, where predictions fail. Here we show that degeneracy appears when the corresponding density of states function is not log-concave, which is typically the consequence of nonlinear relationships between the constraining observables. Exploiting these nonlinear relationships here we propose a solution to the degeneracy problem for a large class of systems via transformations that render the density of states function log-concave. The effectiveness of the method is demonstrated on examples.

DOI: 10.1103/PhysRevLett.114.158701 PACS numbers: 89.70.Cf, 05.20.-y, 87.23.Ge, 89.75.Hc

Our understanding and modeling of complex systems is always based on partial information, limited data and knowledge. The only principled method of predicting properties of a complex system subject to what is known (data and knowledge) is based on the Maximum Entropy Principle of Jaynes [1,2]. Using this principle, he re-derived the formalism of statistical mechanics, both classical [1] and the time-dependent quantum density-matrix formalism [2], using Shannon's information entropy [3]. The method generates a probability distribution $P(\mu)$ over all the possible (micro)states μ of the system by maximizing the entropy $S[P] = -\sum_{\mu} P(\mu) \ln P(\mu)$ subject to what is known, the latter expressed as ensemble averages over $P(\mu)$. In this context, the given data and the available knowledge act as *constraints*, restricting the set of candidate states describing the system. $P(\mu)$ is then used via the usual partition function formalism to make unbiased predictions about other observables.

The applicability of Jaynes's method extends well beyond physics [4], and in particular, it has been applied in biology [5–12], neuroscience [13–21], ecology [22,23], sociology [24,25], economics [26,27], engineering [28,29], computer science [30], etc. It also received attention within network science [31–38], leading to a class of models known as exponential random graphs (ERGs). Despite its popularity, however, this method often presents a fundamental problem, the *degeneracy problem*, that seriously hinders its applicability [34,35]. When this problem occurs, $P(\mu)$ lacks concentration around the averages of the constrained quantities and the typical microstates do not obey the constraints. In the case of ERGs, the generated graphs, for example, may either be very sparse, or very dense, but hardly any will have a density close to that of the data network. Predictions based on such distributions can be significantly off. Two basic questions arise related to the degeneracy problem: (1) Under what conditions does it occur? (2) How can we eliminate or minimize this problem?

In this Letter we answer both questions and present a solution that significantly reduces degeneracy, then illustrate its effectiveness on concrete examples. We will present our analysis and results using the language of networks and ERG models; however, our findings are generally applicable. Let us consider the set $\mathcal{G}_N$ of all labeled simple graphs $G \in \mathcal{G}_N$ (no parallel edges, or self-loops) on N nodes, corresponding here to microstates μ, and an arbitrary set of graph measures or observables $\mathbf{m}(G) = m_1(G), \ldots, m_K(G)$, e.g., the number of edges $m_|$, 2-stars $m_\vee$, triangles m_Δ, the degree of the 9th node, etc. These measures represent the constraints and we assume that we are given specific values $\mathbf{m}^0$, for them (input data). They may come from an empirical network G^0, or could represent averages from several empirical data sets. A key assumption in Jaynes's method is to impose these data at the level of ensemble averages:

$$\mathbf{m}^0 = \langle \mathbf{m}(G) \rangle = \sum_{G \in \mathcal{G}_N} \mathbf{m}(G) P(G), \tag{1}$$

and the goal is to determine the ensemble itself, i.e., the probabilities $P(G)$ for all G, as constrained by Eq. (1) and normalization: $\sum_{G \in \mathcal{G}_N} P(G) = 1$. Since the number of constraints K is usually small, the system of Eq. (1) is strongly underdetermined, the number of unknowns being $|\mathcal{G}_N| = 2^{\mathcal{O}(N^2)}$. Following Jaynes, the least biased distribution $P(G)$ obeying the constraints is the one that maximizes the entropy $S[P] = -\sum_{G \in \mathcal{G}_N} P(G) \ln P(G)$ subject to Eq. (1) and normalization. The method of Lagrange multipliers then yields the family of Gibbs distributions:

$$P(G) = P(G;\boldsymbol{\beta}) = \frac{e^{-\sum_{k=1}^{K}\beta_k m_k(G)}}{Z(\boldsymbol{\beta})} = \frac{e^{-\boldsymbol{\beta}\cdot\mathbf{m}(G)}}{Z(\boldsymbol{\beta})}, \qquad (2)$$

where $Z(\boldsymbol{\beta}) = \sum_{G\in\mathcal{G}_N} e^{-\boldsymbol{\beta}\cdot\mathbf{m}(G)}$ is the partition function. The $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ are Lagrange multipliers associated with the constraints $\mathbf{m} = (m_1, \ldots, m_K)$, determined from solving system (1) with Eq. (2), i.e.,

$$\langle m_k\rangle = \frac{\partial F(\boldsymbol{\beta})}{\partial\beta_k} \qquad (3)$$

where $F(\boldsymbol{\beta}) = -\ln Z(\boldsymbol{\beta})$ denotes the free energy. The average of some other graph measure $q(G)$ in this ensemble will be $\langle q\rangle = \sum_{G\in\mathcal{G}_N} q(G)P(G;\boldsymbol{\beta})$. The distribution $P(G;\boldsymbol{\beta})$ defines the corresponding exponential random graph model, hereinafter referred to as the ERG($\mathbf{m}$) model. Equation (3) admits a maximum likelihood interpretation: its solution is the set of parameters $\boldsymbol{\beta}$ that maximize the probability $P(G^0;\boldsymbol{\beta}) = Z^{-1}(\boldsymbol{\beta})e^{-\boldsymbol{\beta}\cdot\mathbf{m}^0}$ of the graph G^0 for which $\mathbf{m}(G^0) = \mathbf{m}^0$. Note that all graphs having the same properties, $\mathbf{m}$ will have the same probability in the ERG($\mathbf{m}$) model.

Since the partition function is determined by the graph measures only, we may write $Z(\boldsymbol{\beta}) = \sum_{\mathbf{m}}\mathcal{N}(\mathbf{m})e^{-\boldsymbol{\beta}\cdot\mathbf{m}}$, where $\mathcal{N}(\mathbf{m})$ is a *counting function*, representing the number of graphs that have the same values for these measures, equivalent to the density of states function in physics. For example, $\mathcal{N}(m_|, m_\Delta)$ is the number of graphs with $m_|$ edges and m_Δ triangles. To simplify the notations, in the following we will work with adimensional and rescaled quantities $m_i \in [0, 1]$ [39]. Let us denote the domain of $\mathcal{N}$ by $\mathcal{D} = \{\mathbf{m}\in[0,1]^K | \mathcal{N}(\mathbf{m}) \geq 1\}$. Therefore, the probability that a graph sampled by the ERG($\mathbf{m}$) model will have the given $\mathbf{m}$ is

$$p(\mathbf{m};\boldsymbol{\beta}) = \frac{\mathcal{N}(\mathbf{m})}{Z(\boldsymbol{\beta})}e^{-\boldsymbol{\beta}\cdot\mathbf{m}}, \qquad (4)$$

and thus we can write Eq. (3) as the mean of $p(\mathbf{m};\boldsymbol{\beta})$:

$$\langle\mathbf{m}\rangle = \sum_{\mathbf{m}}\mathbf{m}\,p(\mathbf{m};\boldsymbol{\beta}). \qquad (5)$$

Sharp constraints.—In the discussion above the constraints were imposed at the level of averages. It may happen, however, that some of the data hold for all states of the system, akin to integrals of motion in physics. In this case, we restrict ourselves to the largest set of graphs $\mathcal{G}_N(\mathbf{m}^0)\subseteq\mathcal{G}_N$, all having the same value $\mathbf{m}^0$ for those particular measures. We refer to these types of constraints as *sharp constraints*. Examples include the set of all graphs with a given number of edges [the $G(N, M)$ model], introduced by Erdős and Rényi [40], or those with a given degree sequence [41,42], or with a given joint-degree matrix [43]. While sharp constraint problems are mathematically hard in general, counting problems, i.e., computing $\mathcal{N}(\mathbf{m})$, were shown to be the hardest [44,45].

The degeneracy problem.—When solving Eq. (3) [or Eq. (5)] for $\boldsymbol{\beta}$ with given $\langle\mathbf{m}\rangle = \mathbf{m}^0$ we are fixing the parameters $\boldsymbol{\beta}(\mathbf{m}^0) \equiv \boldsymbol{\beta}^0$. It may happen that $p(\mathbf{m};\boldsymbol{\beta}^0)$ is multimodal, with the probability mass concentrated around two or more disjoint and well separated (by $\mathcal{O}(1)$ distances) domains in the observables $\mathbf{m}$, in which case the ERG($\mathbf{m}$) is called degenerate. As examples, let us consider the two ERG models, ERG($m_|, m_\vee$) and ERG($m_|, m_\Delta$), shown in Fig. 1. Figures 1(b) and 1(d) show $p(\mathbf{m};\boldsymbol{\beta})$ at parameter values corresponding to averages ($\langle m_1\rangle, \langle m_2\rangle$) indicated by the black dots. We see that both models are degenerate: for these input values (or corresponding parameters), the sampled graphs will be either very dense or very sparse, practically none with observable values similar to the input data. This is true even in the case when the averages are realizable by specific graphs [seen more clearly in Fig. 1(d)]. Observe that the $\langle\mathbf{m}\rangle$ averages can come from any point in the convex hull of $\mathcal{D}$ (and only from there). Also note that in both cases $\mathcal{N}(\mathbf{m})$ itself is unimodal; however, $p(\mathbf{m};\boldsymbol{\beta}^0)$ is multimodal [46]. It is important to emphasize that when degeneracy occurs the graphs sampled by $p(\mathbf{m};\boldsymbol{\beta})$ are coming from regions with significant probability mass whose separation is large, *comparable to unity.* Strictly speaking, $\mathcal{N}(\mathbf{m})$ is a combinatorial function and it may be jagged locally (integer effects). However, samples from nearby peaks are similar, which is fine for modeling purposes, it is not considered degenerate.

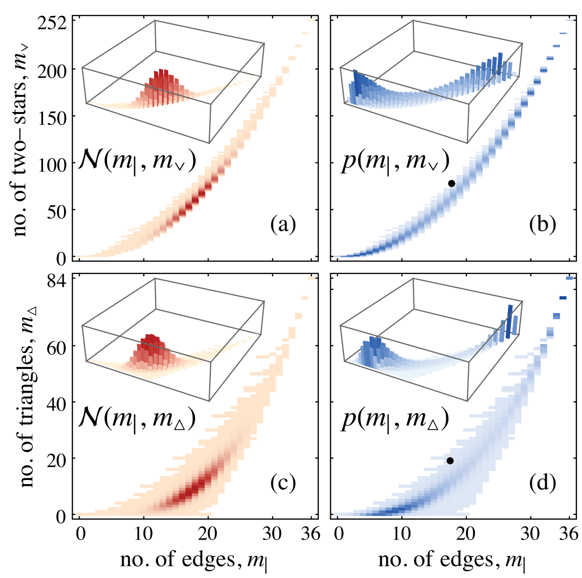

FIG. 1 (color online). Degenerate ERG models. Plots are from exact enumeration of all labeled graphs on $N = 9$ nodes. (a) The counting function $\mathcal{N}(m_|, m_\vee)$. Color intensity is proportional to the value of $\mathcal{N}$, white means $\mathcal{N} = 0$ there. (b) Distribution $p(\mathbf{m};\boldsymbol{\beta})$ from ERG($m_|, m_\vee$) at $\beta_|^0 = 2.20$ and $\beta_\vee^0 = -0.313$, corresponding to the black dot. (c) $\mathcal{N}(m_|, m_\Delta)$. (d) $p(\mathbf{m};\boldsymbol{\beta})$ from ERG($m_|, m_\Delta$) with $\beta_|^0 = 1.24$ and $\beta_\Delta^0 = -0.610$ from the black dot. Insets show 3D versions of the intensity plots. Note from Eq. (4) that the domains of $\mathcal{N}$ and p always coincide.

For that reason, (keeping the notation) in the remainder we will refer to the *smoothened, continuous* version of $\mathcal{N}(\mathbf{m})$, preserving only its long-wavelength properties. For another, non-network example of a degenerate maximum entropy model see Supplemental Material [46]. Degeneracy can be best understood in 1D, $K=1$. Let $f:[a,b]\to\mathbb{R}^+$ be a twice differentiable positive function, and let $g(x)=f(x)e^{-\beta x}$. Since $g(x)>0$, the condition for $g(x)$ not to be multimodal for any β is that it should not have any minima in (a,b) for any β. This is true if in any stationary point x_0, i.e., with $g'(x_0)=0$, the function g is concave, $g''(x_0)<0$. For a stationary point x_0 we have $\beta=f'(x_0)/f(x_0)$. Computing $g''(x_0)$ and eliminating β from it using the above, we get $f''(x_0)f(x_0)<f'(x_0)^2$. Any $x_0\in(a,b)$ can be stationary, since $f(x_0)>0$ and thus the corresponding $\beta=f'(x_0)/f(x_0)$ always exists to make x_0 stationary. Therefore, $g(x)$ will be nondegenerate if and only if $f''(x)f(x)-f'(x)^2<0$ for all $x\in(a,b)$. This is, however, equivalent to saying that $f(x)$ is strictly *log-concave*; i.e., $\ln f(x)$ is (strictly) concave: $d^2\big(\ln f(x)\big)/dx^2<0$ for any $x\in(a,b)$. For example, Gaussians are log-concave. Generalizing this for arbitrary dimensions (for proof see the Supplemental Material [46]), we can announce the following:

Theorem.—The ERG$(\mathbf{m})$ is nondegenerate if and only if the density of states $\mathcal{N}(\mathbf{m})$ is strictly log-concave.

The necessary and sufficient conditions for function $\mathcal{N}(\mathbf{m})$ to be log-concave [50] are that (*i*) its domain $\mathcal{D}$ is convex and (*ii*) if (*i*) holds, to satisfy the Prékopa-Leindler–type inequality $\mathcal{N}(\lambda\mathbf{m}+(1-\lambda)\mathbf{n})>\mathcal{N}(\mathbf{m})^{\lambda}\mathcal{N}(\mathbf{n})^{1-\lambda}$ for any $1<\lambda<1$ and $\mathbf{m},\mathbf{n}\in\mathcal{D}$ [51]. It is important to note that the theorem above reduces degeneracy to purely graph theoretical properties. In two or higher dimensions degeneracy occurs frequently, and the typical approach has been simply to switch to an entirely different set of measures [52]. Realistically, however, we might not have other data, or its collection would not be an option; we want to extract the maximum possible information from the available data. Additionally, one would prefer to let the domain of expertise dictate the natural variables. For example, the triangle count is an important variable in sociology, as it expresses the level of transitivity, an important measure for social networks; yet the corresponding ERG model is degenerate [32].

Solution.—Here we propose to work still with the same variables $\mathbf{m}$ (same data) as in the degenerate ERG model, however, to consider a *one-to-one* transformation $\mathbf{m}\leftrightarrow\boldsymbol{\xi}=\mathbf{F}(\mathbf{m})$ such that the corresponding counting function

$$\bar{\mathcal{N}}(\boldsymbol{\xi})=\mathcal{N}(\mathbf{F}^{-1}(\boldsymbol{\xi})) \tag{6}$$

is log-concave [53]. Because of the one-to-one nature, one can still work with or plot the distributions in the same coordinate system $\mathbf{m}$ [see Figs. 2(b) and 2(c)], but the graphs are sampled by the nondegenerate model ERG$(\boldsymbol{\xi})=$ ERG$(\mathbf{F}(\mathbf{m}))$, with constraints $\boldsymbol{\xi}^0=\mathbf{F}(\mathbf{m}^0)=\langle\boldsymbol{\xi}\rangle$. There is no recipe for obtaining such a transformation in general (it might even not exist, e.g., when $\mathcal{D}$ is not singly connected); however, there is a large class of problems where this can be achieved, to which the degenerate models in the literature belong. This is the case when the convexity condition (*i*) is violated. To better understand the nature of the $\mathbf{F}$ function in this situation, let us focus on the 2D case. If $m_1(G)$ and $m_2(G)$ were independent, $\mathcal{D}$ would be rectangular and therefore convex. Instead, the shapes of the domains in Fig. 1 indicate that there is a nonlinear confining relationship between the variables, on average. For the $(m_|,m_\vee)$ case it holds that $m_\vee\sim m_|^2$ on average [Fig. 2(a), thick orange line]. Similarly, for $(m_|,m_\Delta)$ we have $m_\Delta\sim m_|^3$ (not shown). Focusing on the $(m_|,m_\vee)$ case we can pinpoint why such nonlinear dependencies cause degeneracy. Since $m_\vee\sim m_|^2$, choosing the constraints arbitrarily we are *independently* setting both the average of $m_|$ and its spread $\sigma=(\langle m_|^2\rangle-\langle m_|\rangle^2)^{\frac{1}{2}}$. This is shown most directly by looking at an ERG$(m_|,m_|^2)$ model (see Fig. 3). Since the network is finite, the spread σ can be tuned from a small value corresponding to a unimodal distribution for $m_|$, Figs. 3(a)–3(c), to its maximum Figs. 3(d)–3(f), where the probability mass is bimodal, hence causing degeneracy. Note, a linear relation between the variables will not cause degeneracy. This suggests to choose $\mathbf{F}$ such as to convexify the domain via linearization, i.e., to have $\xi_1\sim\xi_2$. For example, for the $(m_|,m_\vee)$ case this could be done via $\xi_|=m_|^{2\theta}$, $\xi_\vee=m_\vee^{\theta}$, with $\theta>0$ arbitrary, as shown in Fig. 2(b) for $\theta=1$, or for $\theta=1/2$ in the model of Fig. 4.

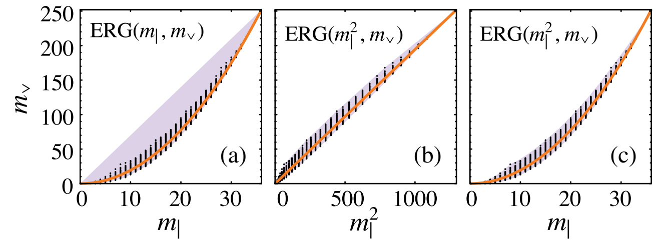

FIG. 2 (color online). Domain $\mathcal{D}$ (black dots) and its convex hull (purple shading); the averages $\langle\mathbf{m}\rangle$ take their values from the convex hull. The orange line is $m_\vee\sim m_|{}^2$. (a) ERG$(m_|,m_\vee)$, in $(m_|,m_\vee)$ space. Note the large region of possible $\langle\mathbf{m}\rangle$ values with no realizable graphs (no black dots). (b) ERG$(m_|^2,m_\vee)$, in $(m_|^2,m_\vee)$ space. Now $\mathcal{D}$ and its convex hull almost coincide. (c) ERG$(m_|^2,m_\vee)$ in $(m_|,m_\vee)$, compare with (a).

Recall that in the original (degenerate) ERG$(\mathbf{m})$ we had $\langle\mathbf{m}\rangle=\mathbf{m}^0$ precisely, by definition. However, the new model ERG$(\boldsymbol{\xi})$ is constrained by $\langle\boldsymbol{\xi}\rangle_\xi=\mathbf{F}(\mathbf{m}^0)\equiv\boldsymbol{\xi}^0$, where the subscript ξ indicates averages in ERG$(\boldsymbol{\xi})$. Here $\langle\mathbf{m}\rangle_\xi\neq\mathbf{m}^0$, yet $\langle\mathbf{m}\rangle_\xi\approx\mathbf{m}^0$ will hold. Let $\boldsymbol{\kappa}^0$ denote the Lagrange parameters in the ERG$(\boldsymbol{\xi})$ model. For the ith component, the difference is on the order of $\frac{1}{2}|\sum_{\boldsymbol{\xi}}(\boldsymbol{\xi}-\boldsymbol{\xi}^0)^T H[F_i^{-1}](\boldsymbol{\xi}^0)(\boldsymbol{\xi}-\boldsymbol{\xi}^0)p(\boldsymbol{\xi};\boldsymbol{\kappa}^0)|\leq(K/2)\|H[F_i^{-1}](\boldsymbol{\xi}^0)\|_2\|\mathrm{Cov}(\boldsymbol{\xi},\boldsymbol{\xi})\|_2$, where $H[F_i^{-1}](\boldsymbol{\xi}^0)$ is

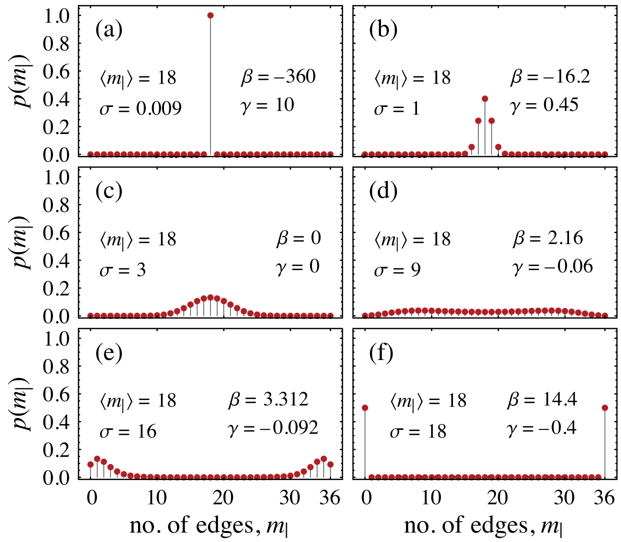

FIG. 3 (color online). Distribution of the edge count $m_|$ of the sampled graphs ($N = 9$ nodes) in the ERG($m_|, m_|^2$) model at various parameter values, where $p(m_|) \propto \mathcal{N}(m_|)\exp(-\beta m_| - \gamma m_|^2)$.

the Hessian of $F_i^{-1}(\boldsymbol{\xi})$ computed in $\boldsymbol{\xi}^0$ and $\|\cdot\|_2$ is the spectral norm. Since ERG($\boldsymbol{\xi}$) is nondegenerate, $p(\boldsymbol{\xi};\boldsymbol{\kappa}^0)$ will be concentrated around $\boldsymbol{\xi}^0$, in a region that is small compared to unity, and additionally, over this region the variability of $\mathbf{F}$ is small [$\mathbf{F}$ straightens the whole domain $\mathcal{D}$, varying significantly only over $\mathcal{O}(1)$ distances]. Thus, while this transformation leads to minor differences, it resolves the degeneracy problem and the samples are with high probability from the neighborhood of graphs for which the given constraints are typical.

Validation.—In the following we test the method on Zachary's well-known karate club (ZKC) data set [54], which describes a network G^0 of club friendships (the Supplemental Material [46] shows the test for another network [49]). ZKC has $N = 34$, $m_|^0 = 78$, $m_\vee^0 = 528$ and $m_\Delta^0 = 45$. Using Markov chain Monte Carlo (MCMC) sampling and a stochastic root finding method, we fitted the ERG($m_|, m_\vee$) model to G^0 obtaining $\beta_|^0 = 2.610$, $\beta_\vee^0 = -0.08125$, and a degenerate $p(m_|, m_\vee; \beta_|^0, \beta_\vee^0)$, shown in Fig. 4(a).

Next we fitted the model ERG($\xi_| = m_|, \xi_\vee = \sqrt{m_\vee}$), obtaining $\kappa_|^0 = 3.625$ and $\kappa_\vee^0 = -7.998$ and a nondegenerate distribution $p(m_|, m_\vee; \kappa_|^0, \kappa_\vee^0)$, shown in Fig. 4(b). The averages are summarized in Table I. Even though here we solve for $\langle\sqrt{m_\vee}\rangle_\xi = \sqrt{m_\vee^0}$, we expect that $\langle m_\vee\rangle_\xi \approx m_\vee^0$. This is confirmed in the $\langle m_\vee\rangle$ column of Table I. Note that due to the degeneracy of ERG($m_|, m_\vee$), its prediction for $\langle\sqrt{m_\vee}\rangle^2$ is 370, far from 528, whereas ERG($m_|, \sqrt{m_\vee}$) predicts all quantities well.

Let us now consider the number of triangles m_Δ. To the extent in which $m_|^0$ and $m_\vee^0$ determine m_Δ, the corresponding ERG model should predict m_Δ as well. Unsurprisingly, ERG($m_|, m_\vee$) produces a bimodal distribution $p(m_\Delta)$, as

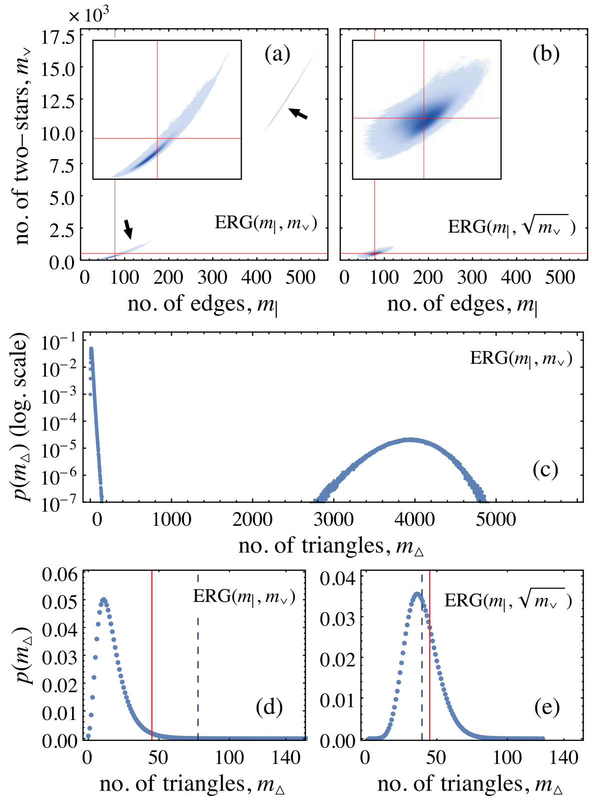

FIG. 4 (color online). Modeling the Zachary Karate Club data. Distributions for ERG($m_|, m_\vee$) [(a), (c), (d)] and ERG($m_|, \sqrt{m_\vee}$) [(b), (e)] after fitting. (a) $p(m_|, m_\vee)$ in ERG($m_|, m_\vee$) and (b) $p(m_|, m_\vee)$ in ERG($m_|, \sqrt{m_\vee}$). The crosshairs are at $(m_|^0, m_\vee^0)$. Insets are magnifications around $(m_|^0, m_\vee^0)$. Arrows (a) indicate the two modes of the degenerate distribution. (c)–(e) show $p(m_\Delta)$ in the two models. The red vertical lines are at m_Δ^0 and the dashed ones are model averages.

shown in Figs. 4(c) and 4(d), and predicts $\langle m_\Delta\rangle = 78$, far from 45. Additionally, 45 and 78 are produced with low probability in the ERG($m_|, m_\vee$) model [see Fig. 4(d)]. The ERG($m_|, \sqrt{m_\vee}$) convexified model, however, predicts $\langle m_\Delta\rangle_\xi = 40$, and both 40 and 45 are produced with high probability in this model, see Fig. 4(e).

It is important to note that the degeneracy problem, the reason for its occurrence, and the solution proposed here are general, applicable beyond network modeling. We have shown that degeneracy will typically appear when the constraining observables (input data) are nonlinearly constraining one another so that the density of states function

TABLE I. Averages of measures in the fitted ERG models. G^0 denotes the ZKC network. For the averages we also indicate the standard error of the MCMC estimates.

	$\langle m_\| \rangle$	$\langle m_\vee \rangle$	$\langle\sqrt{m_\vee}\rangle^2$	$\langle m_\Delta \rangle$
G^0 (ZKC)	78	528	528	45
ERG($m_\|, m_\vee$)	77.8 ± 0.5	530 ± 9	370 ± 4	77.7 ± 2.3
ERG($m_\|, \sqrt{m_\vee}$)	77.9 ± 0.5	530.7 ± 2.7	527.3 ± 2.5	39.5 ± 0.3

is not log-concave. To avoid degeneracy, but still be able to use the same input data, here we proposed one-to-one mappings of the observables (so that no information is lost) in ways that render the density of states function log-concave.

We thank L. Székely, K. Bassler, M. Varga, and D. C. Vural for discussions. We acknowledge financial support in part from Grant No. FA9550-12-1-0405 from the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA) and, additionally, in part from the Defense Threat Reduction Agency (DTRA) under Grant No. HDTRA 1-09-1-0039.

*toro@nd.edu

[1] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[2] E. T. Jaynes, Phys. Rev. **108**, 171 (1957).
[3] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[4] S. Pressé, K. Ghosh, J. Lee, and K. A. Dill, Rev. Mod. Phys. **85**, 1115 (2013).
[5] A. K. Livesey and J. C. Brochon, Biophys. J. **52**, 693 (1987).
[6] P. J. Steinbach, R. Ionescu, and C. R. Matthews, Biophys. J. **82**, 2244 (2002).
[7] G. Yeo and C. B. Burge, J. Comput. Biol. **11**, 377 (2004).
[8] L. P. Watkins, H. Chang, and H. Yang, J. Phys. Chem. A **110**, 5191 (2006).
[9] M. Vergassola, E. Villermaux, and B. I. Shraiman, Nature (London) **445**, 406 (2007).
[10] Z. M. Saul and V. Filkov, Bioinformatics **23**, 2604 (2007).
[11] A. M. Walczak, G. Tkačik, and W. Bialek, Phys. Rev. E **81**, 041905 (2010).
[12] F. Morcos, N. P. Schafer, R. R. Cheng, J. N. Onuchic, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **111**, 12408 (2014).
[13] J. Shlens, G. D. Field, J. L. Gauthier, M. I. Grivich, D. Petrusca, A. Sher, A. M. Litke, and E. J. Chichilnisky, J. Neurosci. **26**, 8254 (2006).
[14] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Nature (London) **440**, 1007 (2006).
[15] A. Tang, D. Jackson, J. Hobbs, W. Chen, J. L. Smith, H. Patel, A. Prieto, D. Petrusca, M. I. Grivich, A. Sher, P. Hottowy, W. Dabrowski, A. M. Litke, and J. M. Beggs, J. Neurosci. **28**, 505 (2008).
[16] O. Marre, S. El Boustani, Y. Frégnac, and A. Destexhe, Phys. Rev. Lett. **102**, 138101 (2009).
[17] F.-C. Yeh, A. Tang, J. P. Hobbs, P. Hottowy, W. Dabrowski, A. Sher, A. Litke, and J. M. Beggs, Entropy **12**, 89 (2010).
[18] G. J. Stephens, L. C. Osborne, and W. Bialek, Proc. Natl. Acad. Sci. U.S.A. **108**, 15565 (2011).
[19] E. Ganmor, R. Segev, and E. Schneidman, Proc. Natl. Acad. Sci. U.S.A. **108**, 9679 (2011).
[20] T. Watanabe, S. Hirose, H. Wada, Y. Imai, T. Machida, I. Shirouzu, S. Konishi, Y. Miyashita, and N. Masuda, Nat. Commun. **4**, 1370 (2013).
[21] M. Ercsey-Ravasz, N. T. Markov, C. Lamy, D. C. Van Essen, K. Knoblauch, Z. Toroczkai, and H. Kennedy, Neuron **80**, 184 (2013).
[22] S. J. Phillips, R. P. Anderson, and R. E. Schapire, Ecol. Model. **190**, 231 (2006).
[23] J. Harte, *Maximum Entropy and Ecology* (Oxford University Press, New York, 2011).
[24] P. Fronczak, A. Fronczak, and J. A. Hołyst, Phys. Rev. E **75**, 026103 (2007).
[25] A. Wimmer and K. Lewis, Am. J. Sociology **116**, 583 (2010).
[26] F. M. Bass, J. Market. Res. **11**, 1 (1974).
[27] S. Agrawal, Z. Wang, and Y. Ye, in *Internet and Network Economics*, edited by C. Papadimitriou and S. Zhang, Lecture Notes in Computer Science Vol. 5385 (Springer-Verlag, Berlin, 2008), pp. 126–137.
[28] S. F. Gull and G. J. Daniell, Nature (London) **272**, 686 (1978).
[29] U. Skoglund, L. G. Öfverstedt, R. M. Burnett, and G. Bricogne, J. Struct. Biol. **117**, 173 (1996).
[30] R. Rosenfeld, Comput. Speech Lang. **10**, 187 (1996).
[31] P. W. Holland and S. Leinhardt, J. Am. Stat. Assoc. **76**, 33 (1981).
[32] D. Strauss, SIAM Rev. **28**, 513 (1986).
[33] J. Park and M. E. J. Newman, Phys. Rev. E **70**, 066146 (2004).
[34] J. Park and M. E. J. Newman, Phys. Rev. E **70**, 066117 (2004).
[35] J. Park and M. E. J. Newman, Phys. Rev. E **72**, 026136 (2005).
[36] G. L. Robins, P. Pattison, Y. Kalish, and D. Lusher, Soc. Networks **29**, 173 (2007).
[37] P. Fronczak, A. Fronczak, and M. Bujok, Phys. Rev. E **88**, 032810 (2013).
[38] T. House, EPL **105**, 68006 (2014).
[39] In the figures, however, we indicate the full range of the values.
[40] B. Bollobás, *Random Graphs*, 2nd ed. (Cambridge University Press, Cambridge, England, 2001).
[41] H. Kim, Z. Toroczkai, I. Miklós, P. Erdős, and L. Székely, J. Phys. A **42**, 392001 (2009).
[42] C. I. Del Genio, H. Kim, Z. Toroczkai, and K. E. Bassler, PLoS One **5**, e10012 (2010).
[43] É. Czabarka, A. Dutle, P. L. Erdős, and I. Miklós, Discrete Appl. Math. **181**, 283 (2015).
[44] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani, Theor. Comput. Sci. **43**, 169 (1986).
[45] V. Vazirani, *Approximation Algorithms* (Springer, New York, 2003).
[46] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.114.158701 for animated versions of Fig. 1, proof of the theorem in arbitrary dimensions, non-network examples of degenerate maximum entropy models, and for validation of the method using a larger network data set. It includes Refs. [27,47–49].
[47] P. Ramond, *Group Theory: A Physicist's Survey* (Cambridge University Press, Cambridge, England, 2010).
[48] M. Bóna, *Combinatorics of Permutations (Discrete Mathematics and its Applications)* (Chapman and Hall/CRC Press, London, 2004).
[49] P. M. Gleiser and L. Danon, Adv. Compl. Syst. **06**, 565 (2003).
[50] S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, England, 2004).
[51] Equivalent to the usual definition for strict concavity $g(\lambda\mathbf{x}_1+(1-\lambda)\mathbf{x}_2)<tg(\mathbf{x}_1)+(1-t)g(\mathbf{x}_1)$, $\forall\mathbf{x}_i\in\mathcal{D},\lambda\in(0,1)$ with $g=\ln\mathcal{N}$.
[52] T. A. B. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock, Sociol. Methodol. **36**, 99 (2006).
[53] Since $\mathbf{F}$ is bijective, the counts of states with $\boldsymbol{\xi}$ is the same as the counts of states with $\mathbf{m}=\mathbf{F}^{-1}(\boldsymbol{\xi})$.
[54] W. W. Zachary, Journal of Anthropological Research **33**, 452 (1977).