J. Appl. Prob. **50,** 883–888 (2013)
Printed in England

EMERGENT STRUCTURES IN LARGE NETWORKS

DAVID ARISTOFF,* *University of Minnesota*
CHARLES RADIN,** *The University of Texas at Austin*

Abstract

We consider a large class of exponential random graph models and prove the existence of a region of parameter space corresponding to the emergent multipartite structure, separated by a phase transition from a region of disordered graphs. An essential feature is the formalism of graph limits as developed by Lovász *et al.* for dense random graphs.

Keywords: Exponential random graph model; complex network; phase transition

2010 Mathematics Subject Classification: Primary 60B99
Secondary 05C35

1. Introduction and statement of results

Complex networks, including the Internet, World Wide Web, social networks, biological networks, etc., are often modeled by probabilistic ensembles with one or more adjustable parameters; see, for instance, [4], [5], [9], [12], and the many references therein. We will use one of these standard families, the exponential random graph models (see the references in [2], [9], [12], and [13]), to study how the multipartite structure can exist in such networks, stable against random fluctuations, in imitation of the modeling of the crystalline structure of solids in thermal equilibrium.

Let H_1 be an edge, and let H_2 be any finite simple graph with $k \geq 2$ edges. We will be considering the two-parameter family of exponential random graph models, with probability mass function on graphs G_N with N nodes given by

$$\mathbb{P}_{\beta_1,\beta_2}(G_N) = \exp\{N^2[\beta_1 t_1(G_N) + \beta_2 t_2(G_N) - \psi_N(\beta_1, \beta_2)]\}, \tag{1}$$

where $t_i(G_N)$ is the density of graph homomorphisms $H_i \to G_N$:

$$t_i(G_N) = \frac{|\hom(H_i, G_N)|}{|V(G_N)|^{|V(H_i)|}}. \tag{2}$$

Here $V(\cdot)$ denotes a vertex set, and the term $\psi_N(\beta_1, \beta_2)$ in (1) gives the probability normalization.

We think of the parameters β_1 and β_2 as representing mechanisms for influencing the network, as pressure and temperature do in models of materials in thermal equilibrium. Indeed, it is easy to see by differentiation that if β_1 is fixed, varying β_2 will vary the mean value of the 'energy' density, $t_2(G_N)$; similarly, if β_2 is fixed, varying β_1 will vary the mean value of the edge density, $t_1(G_N)$. Furthermore, if the mean value $\mathbb{E}_{\beta_1,\beta_2}[t_1(G_N)]$ of $t_1(G_N)$ is fixed and $\beta_2 \ll 0$

Received 20 October 2011; revision received 17 July 2012.
* Postal address: Department of Mathematics, University of Minnesota, Minneapolis, MN 55455, USA.
Email address: daristof@umn.edu
** Postal address: Mathematics Department, The University of Texas at Austin, Austin, TX 78712-1202, USA.
Email address: radin@math.utexas.edu

then, as we will see below, the random graph will have a very low value for the mean value $\mathbb{E}_{\beta_1,\beta_2}[t_2(G_N)]$ of $t_2(G_N)$. However, if $\mathbb{E}_{\beta_1,\beta_2}[t_1(G_N)]$ is fixed, any variation of $\beta_2 > 0$ does not affect $\mathbb{E}_{\beta_1,\beta_2}[t_2(G_N)]$ (when N is large) [15]. It is natural to treat separately the cases $\beta_2 < 0$ and $\beta_2 > 0$. The former is called *repulsive*, the latter *attractive*; see [15]. The attractive case $\beta_2 > 0$ has been completely analyzed in [15], so we concentrate here on the case with repulsion, $\beta_2 < 0$.

It is useful to analyze the phenomenon in the last paragraph, as regards $\beta_2 \ll 0$, in two stages. First, consider the nonprobabilistic optimization problem in which one minimizes the density $t_2(G_N)$ among graphs G_N of N nodes, corresponding intuitively to $\beta_2 = -\infty$. Such problems have been widely studied following the pioneering work of Turán [16]. One can understand the exponential random graph models as a means of analyzing such 'extremal graph theory' problems using the language of statistical mechanics [14], [17]. The function $\psi_N(\beta_1, \beta_2)$ represents the free energy of a grand canonical ensemble, which is the Legendre transform of the entropy of a microcanonical ensemble. The latter is the usual setting for extremal graph theory problems.

Fundamental to our results are questions of analyticity of the normalization in (1), which we discuss next. (See [8] for elementary properties of real analytic functions of several real variables.) An explicit formulation of the normalization is

$$\psi_N(\beta_1, \beta_2) = \frac{1}{N^2} \ln\left(\sum_{G_N} \exp\{N^2[\beta_1 t_1(G_N) + \beta_2 t_2(G_N)]\}\right). \tag{3}$$

It is proven in [2] that

$$\psi_\infty(\beta_1, \beta_2) = \lim_{N\to\infty} \psi_N(\beta_1, \beta_2)$$

exists for all β_1, β_2. By Theorem 6.1 of [2], the method, using analyticity, of the proof of Theorem 3.10 of [15] can be immediately extended to prove that $\psi_\infty(\beta_1, \beta_2)$ is analytic in the real variables β_1 and β_2 when $|\beta_2| < 2/[k(k-1)]$, where k is the number of edges in H_2. It is also noted in [15] that at points where ψ_∞ is analytic,

$$\frac{\partial}{\partial\beta_j}\psi_\infty(\beta_1, \beta_2) = \lim_{N\to\infty} \frac{\partial}{\partial\beta_j}\psi_N(\beta_1, \beta_2), \tag{4}$$

that is, the partial derivatives commute with the limit $N \to \infty$. Partial derivatives of ψ_∞, when they exist, give information on the large-N mean and variance of the densities $t_1(G_N)$ and $t_2(G_N)$ (see [15]), and it is standard in the corresponding modeling of materials, in part for this reason, to define phases and phase transitions as follows (see [6]).

Definition. A phase is an open connected region of the parameter space $\{(\beta_1, \beta_2)\}$ which is maximal for the condition that $\psi_\infty(\beta_1, \beta_2)$ is analytic. The 'high temperature phase' is that domain of analyticity of $\psi_\infty(\beta_1, \beta_2)$ which contains the strip $-2/[k(k-1)] < \beta_2 < 0$. There is a phase transition at (β_1^*, β_2^*) if (β_1^*, β_2^*) is a boundary point of an open set on which ψ_∞ is analytic, but ψ_∞ is not analytic at (β_1^*, β_2^*).

In this notation our main result is as follows.

Theorem 1. *Assume that the chromatic number $\chi(H_2)$ of H_2 is at least 3. Then there is a function $s(\beta_1)$, $-\infty < \beta_1 < \infty$, with $s(\beta_1) \le -2/k(k-1)$, such that, for every β_1, the interval $\{(\beta_1, \beta_2) \mid \beta_2 \le s(\beta_1)\}$ does not intersect the high temperature phase.*

2. Proof of Theorem 1

We write $\mathbb{P}$ for the probability mass function $\mathbb{P}_{\beta_1,\beta_2}$ given by (1), and $\mathbb{E}$ for the expectation $\mathbb{E}_{\beta_1,\beta_2}$.

Before beginning we need some notation; see [1], [2], [3], [9], and [10] for discussions of the ideas behind these terms, which basically provide the framework for 'infinite volume limits' for graphs, in analogy with the infinite volume limit in statistical mechanics [14].

To each graph G on N nodes we associate the following function on $[0, 1]^2$:

$$f^G(x, y) = \begin{cases} 1 & \text{if } (\lceil Nx \rceil, \lceil Ny \rceil) \text{ is an edge of } G, \\ 0 & \text{otherwise.} \end{cases}$$

We define $\mathcal{W}$ to be the space of measurable functions $h\colon [0, 1]^2 \to [0, 1]$ which are symmetric, i.e. $h(x, y) = h(y, x)$ for all x, y. For $h \in \mathcal{W}$, we define

$$t(H, h) = \int_{[0,1]^\ell} \prod_{(i,j)\in E(H)} h(x_i, x_j)\, \mathrm{d}x_1 \cdots \mathrm{d}x_\ell,$$

where $E(H)$ is the edge set of H and $\ell = |V(H)|$ is the number of nodes in H, and note that, for a graph G, $t(H, G)$ defined in (2) has the same value as $t(H, f^G)$. For $g \in \mathcal{W}$, we write $t_i(g) = t(H_i, g)$ for $i = 1, 2$.

We define an equivalence relation on $\mathcal{W}$ as follows: $f \sim g$ if and only if $t(H, f) = t(H, g)$ for every simple graph H. Elements of the quotient space, $\tilde{\mathcal{W}}$, are called 'graphons', and the class containing $h \in \mathcal{W}$ is denoted $\tilde{h}$.

On $\tilde{\mathcal{W}}$ we define a metric in steps as follows. First, on $\mathcal{W}$ we define

$$d_\square(f, g) = \sup_{S,T\subseteq[0,1]} \left| \int_{S\times T} [f(x, y) - g(x, y)]\, \mathrm{d}x\, \mathrm{d}y \right|.$$

Let Σ be the space of measure preserving bijections σ of $[0, 1]$, and, for f in $\mathcal{W}$ and $\sigma \in \Sigma$, define $f_\sigma(x, y) = f(\sigma(x), \sigma(y))$. Using this, we define a metric on $\tilde{W}$ by

$$\delta_\square(\tilde{f}, \tilde{g}) = \inf_{\sigma_1,\sigma_2} d_\square(f_{\sigma_1}, g_{\sigma_2}).$$

In the topology induced by this metric, $\tilde{\mathcal{W}}$ is compact [11].

Next we need a few terms associated with ψ_∞. Define, on $[0, 1]$,

$$I(u) = \tfrac{1}{2}u \ln(u) + \tfrac{1}{2}(1 - u) \ln(1 - u),$$

and, on $\tilde{\mathcal{W}}$,

$$I(\tilde{h}) = \int_{[0,1]^2} I(h(x, y))\, \mathrm{d}x\, \mathrm{d}y.$$

Also, on $\tilde{\mathcal{W}}$ we define

$$T(\tilde{h}) = \beta_1 t_1(h) + \beta_2 t_2(h).$$

The above is relevant because it was proven in Theorem 3.1 of [2] that $\psi_\infty(\beta_1, \beta_2)$ is the solution of an optimization problem:

$$\psi_\infty(\beta_1, \beta_2) = \sup_{\tilde{h}\in\tilde{\mathcal{W}}} [T(\tilde{h}) - I(\tilde{h})]. \tag{5}$$

(Note that it follows immediately from (5) that $\psi_\infty(\beta_1, \beta_2)$ is convex.) From Theorem 3.2 of [2] one has some control on the asymptotic behavior as $N \to \infty$, i.e.

$$\delta_\square[\tilde{G}_N, \tilde{F}^*(\beta_1, \beta_2)] \to 0 \quad \text{in probability as } N \to \infty,$$

where $\tilde{F}^*(\beta_1, \beta_2)$ is the (nonempty) subset of $\tilde{W}$ on which $T - I$ is maximized, and $\tilde{G}_N = \tilde{f}^{G_N}$.

We now return to our proof. Our proof will be by contradiction, so we assume from here on that $\psi_\infty(\beta_1, \beta_2)$ is analytic in β_1 and β_2 on the *entire* half-line $L = \{(\beta_1^*, \beta_2) \colon \beta_2 < 0\}$, where β_1^* is arbitrary but fixed. We will find a contradiction, which will prove the existence of the function $s(\beta_1)$. Consider the function

$$C(\beta_1, \beta_2) := \left(\frac{\partial \psi_\infty}{\partial \beta_1}(\beta_1, \beta_2)\right)^k - \frac{\partial \psi_\infty}{\partial \beta_2}(\beta_1, \beta_2), \tag{6}$$

where k is the number of edges in H_2. Note that $C(\beta_1, \beta_2)$ is analytic on L, since $\psi_\infty(\beta_1, \beta_2)$ is.

Proposition 3.2 of [15] proves that, for all $\beta_2 < 0$, there is a unique solution $u^*(\beta_1, \beta_2)$ to the optimization of

$$\beta_1 u + \beta_2 u^k - \tfrac{1}{2} u \ln u - \tfrac{1}{2}(1-u)\ln(1-u)$$

for $u \in [0, 1]$. Then from Theorems 6.1 and 4.2 of [2] we can use the same argument as used to prove Equations (33) and (34) of [15] to prove that, for $-2/[k(k-1)] < \beta_2 < 0$,

$$\frac{\partial}{\partial \beta_1}\psi_\infty(\beta_1, \beta_2) = \lim_{N\to\infty} \mathbb{E}[t_1(G_N)] = t_1(u^*) = u^*(\beta_1, \beta_2),$$
$$\frac{\partial}{\partial \beta_2}\psi_\infty(\beta_1, \beta_2) = \lim_{N\to\infty} \mathbb{E}[t_2(G_N)] = t_2(u^*) = (u^*(\beta_1, \beta_2))^k.$$

It follows that $C(\beta_1^*, \beta_2) = t_1(u^*)^k - t_2(u^*) = 0$ for $-2/[k(k-1)] < \beta_2 < 0$. Since a function of one variable which is analytic on L and constant on a subinterval must be constant on L, it follows that

$$C(\beta_1^*, \beta_2) = 0 \quad \text{on } L, \tag{7}$$

and so C is identically 0 on the whole high temperature phase. (Any point in the phase can be connected to the β_1 axis by an analytic curve.)

Fix $\varepsilon > 0$ and $i \in \{1, 2\}$. Recall that $\beta_1 = \beta_1^*$ is fixed arbitrarily. Write $\tilde{F}^*(\beta_2)$ for the set $\tilde{F}^*(\beta_1, \beta_2) \subset \tilde{W}$ defined above. Using Theorem 7.1 of [2], choose β_2' sufficiently negative so that, for every $\beta_2 < \beta_2'$,

$$\sup_{\tilde{f} \in \tilde{F}^*(\beta_2)} \delta_\square(\tilde{f}, p\tilde{g}) < \frac{\varepsilon}{3k}, \tag{8}$$

where $p = \mathrm{e}^{2\beta_1}/(1 + \mathrm{e}^{2\beta_1})$ and $g(x, y) = 1$ unless $\lfloor(\chi(H_2) - 1)x\rfloor = \lfloor(\chi(H_2) - 1)y\rfloor$, in which case $g(x, y)$ has value 0.

Let $\beta_2 < \beta_2'$. Using Theorem 3.2 of [2], choose $N_0(\beta_2)$ such that $N > N_0(\beta_2)$ implies that

$$\mathbb{P}\left(\delta_\square(\tilde{G}_N, \tilde{F}^*(\beta_2)) \geq \frac{\varepsilon}{3k}\right) < \frac{\varepsilon}{3k}. \tag{9}$$

Let $N > N_0(\beta_2)$ and $A_{\varepsilon,N} = \{G_N \colon \delta_\square(\tilde{G}_N, \tilde{F}^*(\beta_2)) < \varepsilon/(3k)\}$. There exist $\tilde{h}_{G_N} \in \tilde{F}^*(\beta_2)$ corresponding to each $G_N \in A_{\varepsilon,N}$ such that

$$\delta_\square(\tilde{G}_N, \tilde{h}_{G_N}) < \frac{\varepsilon}{3k}. \tag{10}$$

Write $\mathbb{E}|_A$ for the restriction of the expectation to the set A. Using (8) and (10), we have

$$\begin{aligned}\mathbb{E}|_{A_{\varepsilon,N}}[\delta_\square(\tilde{G}_N, p\tilde{g})] &= \sum_{G_N \in A_{\varepsilon,N}} \delta_\square(\tilde{G}_N, p\tilde{g})\mathbb{P}(G_N) \\ &\le \sum_{G_N \in A_{\varepsilon,N}} [\delta_\square(\tilde{G}_N, \tilde{h}_{G_N}) + \delta_\square(\tilde{h}_{G_N}, p\tilde{g})]\mathbb{P}(G_N) \\ &< \sum_{G_N \in A_{\varepsilon,N}} \left[\frac{\varepsilon}{3k} + \frac{\varepsilon}{3k}\right]\mathbb{P}(G_N) \\ &\le \frac{2\varepsilon}{3k} \end{aligned} \tag{11}$$

for $N > N_0(\beta_2)$.

From Lemma 4.1 of [10], it is easy to see that

$$|t_i(G_N) - t_i(pg)| \le k\delta_\square(\tilde{G}_N, p\tilde{g}). \tag{12}$$

Write $\bar{A}_{\varepsilon,N} = \{G_N : \delta_\square(\tilde{G}_N, \tilde{F}^*(\beta_2)) \ge \varepsilon/(3k)\}$. From (9), (11), (12), and the fact that $\delta_\square(\cdot,\cdot) \le 1$,

$$\begin{aligned}|\mathbb{E}[t_i(G_N)] - t_i(pg)| &\le \mathbb{E}[|t_i(G_N) - t_i(pg)|] \\ &\le k\mathbb{E}[\delta_\square(\tilde{G}_N, p\tilde{g})] \\ &= k(\mathbb{E}|_{A_{\varepsilon,n}}[\delta_\square(\tilde{G}_N, p\tilde{g})] + \mathbb{E}|_{\bar{A}_{\varepsilon,N}}[\delta_\square(\tilde{G}_N, p\tilde{g})]) \\ &< k\left(\frac{2\varepsilon}{3k} + \frac{\varepsilon}{3k}\right) \\ &= \varepsilon \end{aligned} \tag{13}$$

for $N > N_0(\beta_2)$. Direct computation of (3) shows that

$$\frac{\partial \psi_N}{\partial \beta_i}(\beta_1^*, \beta_2) = \mathbb{E}[t_i(G_N)]. \tag{14}$$

Combining (14) with (4), we may take the limit $N \to \infty$ in (13) to obtain

$$\left| t_i(pg) - \frac{\partial \psi_\infty}{\partial \beta_i}(\beta_1^*, \beta_2) \right| < \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary,

$$\lim_{\beta_2 \to -\infty} \frac{\partial \psi_\infty}{\partial \beta_i}(\beta_1^*, \beta_2) = t_i(pg). \tag{15}$$

Direct computation using Equation (2.10) of [2] yields

$$t_2(pg) = 0 \quad \text{and} \quad t_1(pg) = \frac{\mathrm{e}^{2\beta_1}(\chi(H) - 2)}{(1 + \mathrm{e}^{2\beta_1})(\chi(H) - 1)} > 0. \tag{16}$$

Now, by combining (6) with (15)–(16), we find that $\lim_{\beta_2 \to -\infty} C(\beta_1^*, \beta_2) > 0$, in contradiction with (7), which proves the theorem.

3. Conclusion

Consider any of the two-parameter exponential random graph models with repulsion covered by our theorem. We have proven that the high temperature phase is separated from the low energy regime by a phase transition. Our proof is based on the traditional modeling of equilibrium statistical mechanics using analyticity and an order parameter [7], [14], [17]. We also emphasize that this method could not have been used to prove the transition found in [15] for attractive exponential random graph models since there is a critical point for that transition: indeed, there is only one phase for $\beta_2 > 0$.

There remain many open questions. Perhaps the most pressing is the character of the singularity of $\psi_\infty(\beta_1, \beta_2)$ at the boundary of the high energy phase. In the attractive case there is only one phase, but there are jump discontinuities, in the first derivatives of $\psi_\infty(\beta_1, \beta_2)$ (namely, the average edge and energy densities), across a curve where two regions of the phase abut, while the edges are independent in the probabilistic sense throughout the phase [15]. We do not know the nature of the singularity at the boundary of the high energy phase for the case of repulsion studied in this paper, though we expect the first derivatives of $\psi_\infty(\beta_1, \beta_2)$ to be discontinuous across the boundary. In analogy with equilibrium materials there may be multipartite phases with different numbers of parts at low energy, though this may require more complicated interactions [2].

Acknowledgements

It is a pleasure for CR to acknowledge useful discussions with Mei Yin, and support at a workshop of The American Institute of Mathematics in August 2011.

References

[1] Borgs, C. *et al.* (2008). Convergent graph sequences I: subgraph frequencies, metric properties, and testing. *Adv. Math.* **219,** 1801–1851.
[2] Chatterjee, S. and Diaconis, P. (2011). Estimating and understanding exponential random graph models. Preprint. Available at http://arxiv.org/abs/1102.2650v3.
[3] Chatterjee, S. and Varadhan, S. R. S. (2011). The large deviation principle for the Erdős-Rényi random graph. *Europ. J. Combinatorics* **32,** 1000–1017.
[4] Fienberg, S. E. (2010). Introduction to papers on the modeling and analysis of network data. *Ann. Appl. Statist.* **4,** 1–4.
[5] Fienberg, S. E. (2010). Introduction to papers on the modeling and analysis of network data—II. *Ann. Appl. Statist.* **4,** 533–534.
[6] Fisher, M. E. and Radin, C. (2006). Definitions of thermodynamic phases and phase transitions. 2006 Workshop Rep. Available at http://www.aimath.org/WWN/phasetransition/Defs16.pdf.
[7] Kadanoff, L. P. (2011). Theories of matter: infinities and renormalization. In *The Oxford Handbook of the Philosophy of Physics*, ed. R. Batterman, Oxford University Press.
[8] Krantz, S. G. and Parks, H. R. (2002). *A Primer of Real Analytic Functions*, 2nd edn. Birkhäuser, Boston, MA.
[9] Lovász, L. (2009). Very large graphs. *Current Develop. Math.* **2008,** 67–128.
[10] Lovász, L. and Szegedy, B. (2006). Limits of dense graph sequences. *J. Combinatorial Theory B* **96,** 933–957.
[11] Lovász, L. and Szegedy, B. (2007). Szemerédi's lemma for the analyst. *Geom. Funct. Anal.* **17,** 252–270.
[12] Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press, Oxford.
[13] Robins, G., Pattison, P., Kalish, Y. and Lushern, D. (2007). An introduction to exponential random graph (p^*) models for social networks. *Social Networks* **29,** 173–191.
[14] Ruelle, D. (1969). *Statistical Mechanics: Rigorous Results*. World Scientific, River Edge, NJ.
[15] Radin, C. and Yin, M. (2013). Phase transitions in exponential random graphs. To appear in *Ann. Appl. Prob.*
[16] Turán, P. (1941). On an extremal problem in graph theory. *Mat. Fiz. Lapok* **48,** 436–452 (in Hungarian).
[17] Yeoman, J. M. (1992). *Statistical Mechanics of Phase Transitions*. Clarendon Press, Oxford.