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

\[
P_{\beta_1, \beta_2}(G_N) = \exp\{N^2[\beta_1 t_1(G_N) + \beta_2 t_2(G_N) - \psi_N(\beta_1, \beta_2)]\},
\]

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

\[
t_i(G_N) = \frac{|\text{hom}(H_i, G_N)|}{|V(G_N)||V(H_i)|}.
\]

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 \( \beta_1 \) and \( \beta_2 \) 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 \( \beta_1 \) is fixed, varying \( \beta_2 \) will vary the mean value of the ‘energy’ density, \( t_2(G_N) \); similarly, if \( \beta_2 \) is fixed, varying \( \beta_1 \) 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 \( \beta_2 \ll 0 \),
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 $s(\beta_1) \leq -2/(k(k-1))$, such that, for every $\beta_1$, the interval $[s(\beta_1), \beta_2]$, $\beta_2 \leq s(\beta_1)$ does not intersect the high temperature phase.
2. Proof of Theorem 1

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

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 $[0, 1]^2$:

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

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

$$t(H, h) = \int_{[0, 1]^\ell} \prod_{(i,j) \in E(H)} h(x_i, x_j) \, dx_1 \cdots dx_\ell,$$

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 $g \in W$, we write $t_i(g) = t(H_i, g)$ for $i = 1, 2$.

We define an equivalence relation on $W$ 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, $\tilde{W}$, are called ‘graphons’, and the class containing $h \in W$ is denoted $\tilde{h}$.

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

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

Let $\Sigma$ be the space of measure preserving bijections $\sigma$ of $[0, 1]$, and, for $f$ in $W$ 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, $\tilde{W}$ is compact [11].

Next we need a few terms associated with $\psi_\infty$. Define, on $[0, 1]$,

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

and, on $\tilde{W}$,

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

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

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

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:

$$\psi_\infty(\beta_1, \beta_2) = \sup_{\tilde{h} \in \tilde{W}} [T(\tilde{h}) - I(\tilde{h})].$$  (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 \( \tilde{G}_N = \tilde{f}^{\tilde{G}_N} \).

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 \( \beta_1 \) and \( \beta_2 \) on the entire half-line \( L = \{ (\beta_1^*, \beta_2) : \beta_2 < 0 \} \), where \( \beta_1^* \) is arbitrary but fixed. We will find a contradiction, which will prove the existence of the function \( s(\beta_1) \). Consider the function

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

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

\[
\beta_1 u + \beta_2 u^k - \frac{1}{2} u \ln u - \frac{1}{2} (1 - u) \ln(1 - u)
\]

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 \( -2/k(k-1) < \beta_2 < 0 \),

\[
\frac{\partial}{\partial \beta_1} \psi_\infty(\beta_1, \beta_2) = \lim_{N \to \infty} \mathbb{E}[t_1(\tilde{G}_N)] = t_1(u^*) = u^*(\beta_1, \beta_2),
\]

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

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

\[
C(\beta_1^*, \beta_2) = 0 \quad \text{on } L,
\]

and so \( C \) is identically 0 on the whole high temperature phase. (Any point in the phase can be connected to the \( \beta_1 \) axis by an analytic curve.)

Fix \( \epsilon > 0 \) 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 \( \beta_2^* \) sufficiently negative so that, for every \( \beta_2 < \beta_2^* \),

\[
\sup_{j \in \tilde{F}^*(\beta_2)} \delta_{\square}(\tilde{f}, \tilde{p}^j) < \frac{\epsilon}{3k}
\]

(8)

where \( p = e^{2H_1}/(1 + e^{2H_1}) \) 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

\[
P \left( \delta_{\square}(\tilde{G}_N, \tilde{F}^*(\beta_2)) \geq \frac{\epsilon}{3k} \right) < \frac{\epsilon}{3k}.
\]

(9)

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

\[
\delta_{\square}(\tilde{G}_N, \tilde{h}_{G_N}) < \frac{\epsilon}{3k}.
\]

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

$$E|_{A_{ε,N}}[δ□(G_N, \tilde{p}g)] = \sum_{G_N \in A_{ε,N}} δ□(\tilde{G}_N, \tilde{p}g)P(G_N)$$

$$\leq \sum_{G_N \in A_{ε,N}} [δ□(\tilde{G}_N, \tilde{h}_G) + δ□(\tilde{h}_G, \tilde{p}g)]P(G_N)$$

$$< \sum_{G_N \in A_{ε,N}} \left[ \frac{ε}{3k} + \frac{ε}{3k} \right]P(G_N)$$

$$\leq \frac{2ε}{3k}$$

(11)

for $N > N_0(β_2)$.

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

$$|t_i(G_N) - t_i(pg)| \leq kδ□(\tilde{G}_N, \tilde{p}g).$$

(12)

Write $\tilde{A}_{ε,N} = \{G_N : δ□(\tilde{G}_N, \tilde{F}^*_2(β_2)) ≥ \varepsilon/(3k)\}$. From (9), (11), (12), and the fact that $δ□(\cdot, \cdot) ≤ 1$,

$$|E[t_i(G_N)] - t_i(pg)| ≤ E[|t_i(G_N) - t_i(pg)|]$$

$$≤ kE[δ□(\tilde{G}_N, \tilde{p}g)]$$

$$= k(E[\tilde{A}_{ε,N}][δ□(\tilde{G}_N, \tilde{p}g)] + E[\tilde{A}_{ε,N}][δ□(\tilde{G}_N, \tilde{p}g)])$$

$$< k \left( \frac{2ε}{3k} + \frac{ε}{3k} \right)$$

$$= ε$$

(13)

for $N > N_0(β_2)$. Direct computation of (3) shows that

$$\frac{∂ψ_N}{∂β_i}(β_1^*, β_2) = E[t_i(G_N)].$$

(14)

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

$$|t_i(pg) - \frac{∂ψ^∞}{∂β_i}(β_1^*, β_2)| < ε.$$

Since $ε > 0$ was arbitrary,

$$\lim_{β_2 \to -∞} \frac{∂ψ^∞}{∂β_i}(β_1^*, β_2) = t_i(pg).$$

(15)

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

$$t_2(pg) = 0 \quad \text{and} \quad t_1(pg) = \frac{e^{2β_1}(χ(H) - 2)}{(1 + e^{2β_1})(χ(H) - 1)} > 0.$$

(16)

Now, by combining (6) with (15)–(16), we find that $\lim_{β_2 \to -∞} C(β_1^*, β_2) > 0$, 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] Kadonoff, 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 Lusher, 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.