Applications of dense graph limits in probability and statistics

Sourav Chatterjee

(Courant Institute, NYU)

Based on joint works with
Persi Diaconis and S. R. S. Varadhan

Real world networks

- The last decade has seen an explosion in the study of real world networks, e.g. rail and road networks, biochemical networks, data communication networks such as the Internet, and social networks.
- Concerted interdisciplinary effort to develop new mathematical network models to explain characteristics of observed real world networks, such as power law degree behavior, small world properties, and a high degree of clustering.
- Clustering/transitivity/reciprocity refers to the prevalence of triangles in a graph.
- That is, a friend of a friend is more likely to be a friend than a random individual.
- Most of the popular network models, such as the preferential attachment and the configuration models, are locally tree-like and thus do not model the transitivity observed in real social networks.

Exponential Random Graphs

- In the social science literature, efforts to mathematically model transitivity have centered around the so-called Exponential Random Graph Models (ERGM), also called p^* models.
- Statistically, ERGM's are exponential families of distributions on the space of graphs on a given number of vertices.
- The sufficient statistics are usually simple graph parameters, such as the number of edges, number of triangles, etc.
- Notable early papers due to Holland and Leinhardt (1981), Frank and Strauss (1986). General development in the book of Wasserman and Faust. Recent progresses in Handcock (2003), Snijders et. al. (2006), Park and Newman (2004, 2005), etc.

Example

- Consider the model on simple graphs with n vertices,

$$p_{\beta_1,\beta_2}(G) = \exp\left(\beta_1 E + \frac{\beta_2}{n}\Delta - n^2\psi_n(\beta_1,\beta_2)\right)$$

  where E, Δ denote the number of edges and triangles in the graph G, and ψ_n is the normalizing constant.
- The normalization of the model ensures non-trivial large n limits. Without scaling, for large n, almost all graphs are empty or full.
- This model is studied by Strauss (1986), Park and Newman (2004, 2005), Häggstrom and Jonasson (1999), and many others.

Challenges

- For thirty years, nothing much could be done mathematically with these models. For example, no formula for ψ_n, no rigorously proven information about qualitative behavior.
- Approximation of the normalizing constant, necessary for evaluation of maximum likelihood estimates, is usually done with the aid of Markov Chain Monte Carlo.
- But: Bhamidi, Bresler and Sly (2008) have shown that, depending on β_1 and β_2,
  - either the model behaves like an Erdős-Rényi random graph (the uninteresting case),
  - or the usual MCMC algorithms take exponentially long time to converge.
- Alternative (widely used) approach via pseudo-likelihood methods. But: properties are poorly understood, and appreciably higher variability than MLE.

An asymptotic formula for the edge-triangle model

Recall the model:

$$p_{\beta_1,\beta_2}(G) = \exp\left(\beta_1 E + \frac{\beta_2}{n}\Delta - n^2\psi_n(\beta_1,\beta_2)\right),$$

where E and Δ are the number of edges and triangles in G and ψ_n is the normalizing constant.

Theorem (Chatterjee & Diaconis, 2011)

There is a negative constant $-c$, depending on β_1, such that when $-c < \beta_2 < \infty$,

$$\lim_{n\to\infty} \psi_n(\beta_1,\beta_2) = \sup_{0\le u\le 1}\left(\frac{\beta_1 u}{2} + \frac{\beta_2 u^3}{6} - \frac{1}{2}u\log u - \frac{1}{2}(1-u)\log(1-u)\right).$$

There is another negative constant $-d$, again depending on β_1, such that the formula is not valid if $\beta_2 < -d$.

The symmetric phase and symmetry breaking

- The region where the formula is valid is called the symmetric phase in our paper. Partially identified in an earlier work of Chatterjee & Dey (2009).
- In the symmetric phase, we prove that the model behaves essentially like an Erdős-Rényi graph (i.e. independent edges) with edge-probability u^*, where u^* is the value of u that solves the maximization problem in the theorem.
- When $\beta_2 < -d$, we prove that the model stops behaving like an Erdős-Rényi model and enters the region of broken symmetry.
- We do not understand this region very well. We can only prove that when β_2 is large and negative, the random graphs generated from the model look approximately like bipartite graphs.

Large negative β_2

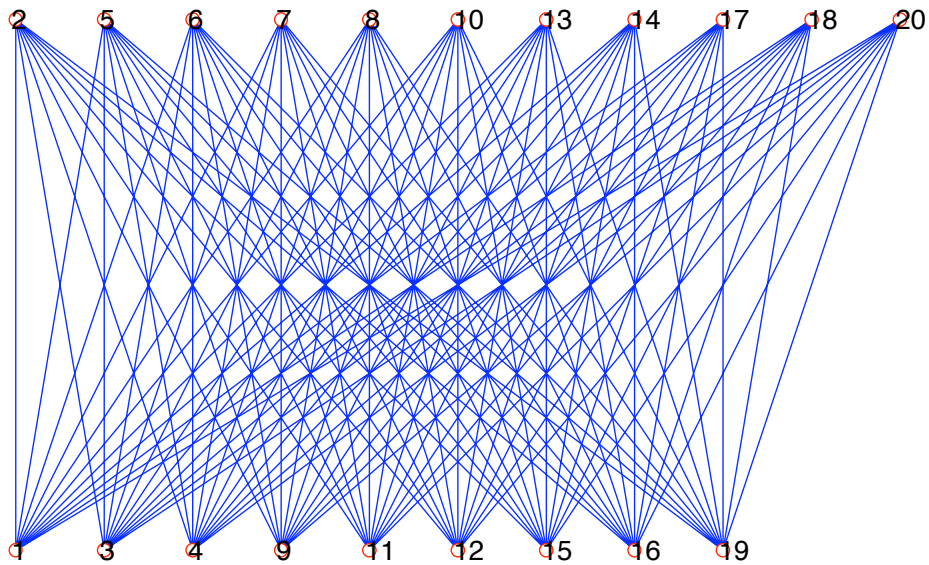


Figure: A simulated realization of the exponential random graph model on 20 nodes with edges and triangles as sufficient statistics, where $\beta_1 = 120$ and $\beta_2 = -400$. (Picture by Sukhada Fadnavis.)

Degeneracy

- Researchers, e.g. Handcock (2003), have observed that models like the edge-triangle model tend to exhibit a certain degeneracy: As the parameter values vary, the random graphs are either very sparse, or almost complete, skipping all intermediate structures.
- We have a theorem that gives a proof of this phenomenon.

Rigorous result about degeneracy

Theorem (Chatterjee & Diaconis, 2011)

Let G_n be a random graph from the edge-triangle model. Fix $\beta_1 < 0$. Let

$$c_1 := \frac{e^{\beta_1/2}}{1 + e^{\beta_1/2}}, \quad c_2 := 1 + \frac{1}{\beta_1}.$$

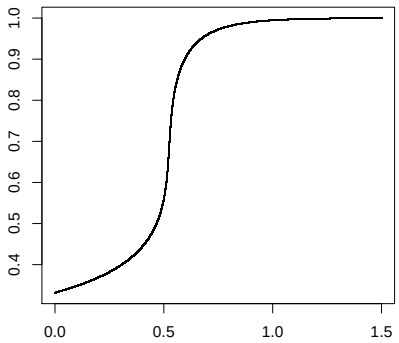
Suppose $|\beta_1|$ is so large that $c_1 < c_2$. Let $e(G_n)$ be the number of edges in G_n and let $f(G_n) := e(G_n)/\binom{n}{2}$ be the edge density. Then there exists $q(\beta_1)$ such that if $\beta_2 < q(\beta_1)$, then as $n \to \infty$,

$$\mathbb{P}(f(G_n) > c_1) \to 0,$$

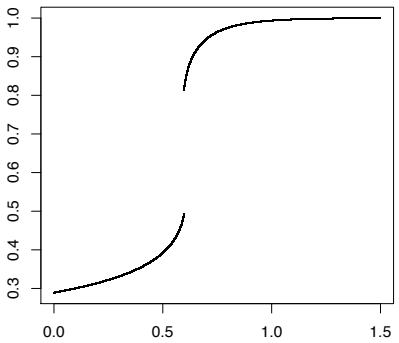
and if $\beta_2 > q(\beta_1)$, then $\mathbb{P}(f(G_n) < c_2) \to 0$.

Remark. The difference in the values of c_1 and c_2 can be quite striking even for relatively small values of β_1. For example, $\beta_1 = -10$ gives $c_1 \simeq 0.007$ and $c_2 = 0.9$.

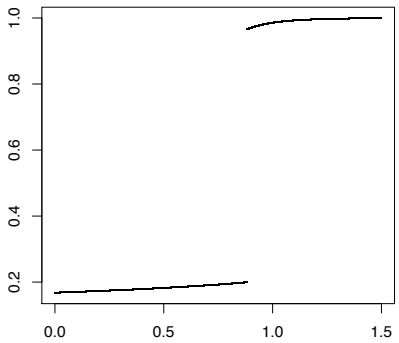
Phase transitions and degeneracy



(b) $\beta_1 = -0.35$



(a) $\beta_1 = -0.45$



(b) $\beta_1 = -0.8$

Figure: Plot of asymptotic edge density (on y-axis) vs. β_2 (on x-axis) for three different values of β_1.

More progress on this recently by Charles Radin and Mei Yin.

General formula for ψ_n

Theorem (Chatterjee & Diaconis, 2011)

For any β_1 and β_2,

$$\lim_{n\to\infty} \psi_n(\beta_1,\beta_2)$$
$$= \sup_f \Biggl(\frac{\beta_1}{2} \iint f(x,y)\,dx\,dy + \frac{\beta_2}{6} \iiint f(x,y) f(y,z) f(z,x)\,dx\,dy\,dz$$
$$- \frac{1}{2} \iint \Bigl\{ f(x,y)\log f(x,y) + (1-f(x,y))\log(1-f(x,y)) \Bigr\} dx\,dy \Biggr),$$

where the supremum is over all measurable $f : [0,1]^2 \to [0,1]$ satisfying $f(x,y) = f(y,x)$ and the integrals are from 0 to 1.

Remarks. (a) The symmetric phase is where the maximizer is a constant function. (b) There is a general version of this theorem in our paper which applies to essentially all exponential random graph models.

First step: counting graphs with a given property

- $2^{n(n-1)/2}$ simple graphs on n vertices.
- Question: Given a property P and an integer n, roughly how many of these graphs have property P?
- For example, P may be: #triangles $\geq tn^3$, where t is a given constant.
- How this helps: The ability to count the number of graphs with a given number of triangles and a given number of edges will lead to the evaluation of the normalizing constant in the edge-triangle model.
- To make any progress, need to assume some regularity on P. For example, we may demand that P be continuous or at least measurable with respect to some metric.
- What metric? What space?

An abstract topological space of graphs

- Beautiful unifying theory developed by Laszlo Lovász and coauthors V. T. Sós, B. Szegedy, C. Borgs, J. Chayes, K. Vesztergombi, A. Schrijver and M. Freedman in the last six years. Related to earlier works of Aldous, Hoover, Kallenberg.
- Let G_n be a sequence of simple graphs whose number of nodes tends to infinity.
- For every fixed simple graph H, let $\hom(H, G)$ denote the number of homomorphisms of H into G (i.e. edge-preserving maps $V(H) \to V(G)$, where $V(H)$ and $V(G)$ are the vertex sets).
- This number is normalized to get the homomorphism density

$$t(H, G) := \frac{\hom(H, G)}{|V(G)|^{|V(H)|}}.$$

This gives the probability that a random mapping $V(H) \to V(G)$ is a homomorphism.

Abstract space of graphs contd.

- Suppose that $t(H, G_n)$ tends to a limit $t(H)$ for every H.
- Then Lovász & Szegedy proved that there is a natural "limit object" in the form of a function $f \in \mathcal{W}$, where $\mathcal{W}$ is the space of all measurable functions from $[0,1]^2$ into $[0,1]$ that satisfy $f(x,y) = f(y,x)$ for all x, y.
- Conversely, every such function arises as the limit of an appropriate graph sequence.
- This limit object determines all the limits of subgraph densities: if H is a simple graph with k vertices, then

$$t(H, f) = \int_{[0,1]^k} \prod_{(i,j) \in E(H)} f(x_i, x_j)\, dx_1 \cdots dx_k.$$

- A sequence of graphs $\{G_n\}_{n \geq 1}$ is said to converge to f if for every finite simple graph H,

$$\lim_{n \to \infty} t(H, G_n) = t(H, f).$$

Example

- Consider the Erdős-Rényi random graph $G(n,p)$. Each edge is present with probability p, independent of other edges.
- For any fixed graph H,

$$t(H, G(n,p)) \longrightarrow p^{|E(H)|} \quad \text{almost surely as } n \longrightarrow \infty.$$

- On the other hand, if f is the function that is identically equal to p, then $t(H,f) = p^{|E(H)|}$.
- Thus, the sequence of random graphs $G(n,p)$ converges almost surely to the non-random limit function $f(x,y) \equiv p$ as $n \longrightarrow \infty$.

Abstract space of graphs contd.

- The elements of $\mathcal{W}$ are sometimes called 'graphons'.
- A finite simple graph G on n vertices can also be represented as a graphon f^G is a natural way:

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

- Note that this allows *all* simple graphs, irrespective of the number of vertices, to be represented as elements of the single abstract space $\mathcal{W}$.
- So, what is the topology on this space?

The cut metric

- For any $f, g \in \mathcal{W}$, Frieze and Kannan defined the cut distance:

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

- Introduce an equivalence relation on $\mathcal{W}$: say that $f \sim g$ if $f(x,y) = g_\sigma(x,y) := g(\sigma x, \sigma y)$ for some measure preserving bijection σ of $[0,1]$.
- Denote by $\widetilde{g}$ the closure in $(\mathcal{W}, d_{\square})$ of the orbit $\{g_\sigma\}$.
- The quotient space is denoted by $\widetilde{\mathcal{W}}$ and τ denotes the natural map $g \to \widetilde{g}$.
- Since $d_{\square}$ is invariant under σ one can define on $\widetilde{\mathcal{W}}$ the natural distance $\delta_{\square}$ by

$$\delta_{\square}(\widetilde{f}, \widetilde{g}) := \inf_{\sigma} d_{\square}(f, g_\sigma) = \inf_{\sigma} d_{\square}(f_\sigma, g) = \inf_{\sigma_1, \sigma_2} d_{\square}(f_{\sigma_1}, g_{\sigma_2})$$

making $(\widetilde{\mathcal{W}}, \delta_{\square})$ into a metric space.

Cut metric and graph limits

To any finite graph G, we associate the natural graphon f^G and its orbit $\widetilde{G} = \tau f^G = \widetilde{f}^G \in \widetilde{\mathcal{W}}$. One of the key results of the theory is the following:

Theorem (Borgs, Chayes, Lovász, Sós & Vesztergombi)

A sequence of graphs $\{G_n\}_{n\geq 1}$ converges to a limit $f \in \mathcal{W}$ if and only if $\delta_\square(\widetilde{G}_n, \widetilde{f}) \to 0$ as $n \to \infty$.

Another important result is:

Theorem (Lovász & Szegedy)

The space $\widetilde{\mathcal{W}}$ is compact under the metric $\delta_\square$.

Counting graphs with a given property

- For any Borel set $\widetilde{A} \subseteq \widetilde{\mathcal{W}}$, let
$$\widetilde{A}_n := \{\widetilde{h} \in \widetilde{A} : \widetilde{h} = \widetilde{G} \text{ for some } G \text{ on } n \text{ vertices}\}.$$
- Let $I(u) := \frac{1}{2} u \log u + \frac{1}{2}(1-u)\log(1-u)$.
- For any $\widetilde{h} \in \widetilde{\mathcal{W}}$, let $I(\widetilde{h}) := \int\int I(h(x,y))dxdy$, where h is any element of $\widetilde{h}$.

Theorem (Chatterjee & Varadhan, 2010)

The function I is well-defined and lower-semicontinuous on $\widetilde{\mathcal{W}}$. For any measurable $\widetilde{A} \subseteq \widetilde{\mathcal{W}}$,

$$-\inf_{\widetilde{h} \in \mathit{closure}(\widetilde{A})} I(\widetilde{h}) \geq \limsup_{n\to\infty} \frac{\log |\widetilde{A}_n|}{n^2}$$
$$\geq \liminf_{n\to\infty} \frac{\log |\widetilde{A}_n|}{n^2} \geq -\inf_{\widetilde{h} \in \mathit{interior}(\widetilde{A})} I(\widetilde{h})$$

A simple application

- Under very special circumstances, the variational problem is known to have an explicit solution.
- For example, we can prove that the number of graphs on n vertices with at least tn^3 triangles is $e^{n^2 f(t)(1+o(1))}$, where

$$f(t) = \begin{cases} \frac{1}{2}\log 2 & \text{if } 0 \le t < \frac{1}{48} \\ -I((6t)^{1/3}) & \text{if } \frac{1}{48} \le t < \frac{1}{6} \\ -\infty & \text{if } t \ge \frac{1}{6}. \end{cases}$$

- On the other hand, for the number of graphs with *at most* tn^3 triangles, we can prove such an explicit formula for t sufficiently away from zero, and can show that the formula *does not hold* sufficiently close to zero. But could not derive an explicit formula for small t.

Method of proof

- Counting graphs with a given property is essentially the same as proving a Large Deviation Principle (LDP) for the Erdős-Rényi random graph $G(n,p)$. For example,

$$\begin{aligned}&\#\text{graphs on } n \text{ vertices satisfying } P\\ &= 2^{n(n-1)/2}\mathbb{P}(G(n,1/2) \text{ satisfies } P).\end{aligned}$$

- The LDP can be proved by standard techniques for the weak topology on $\widetilde{\mathcal{W}}$. (Fenchel-Legendre transforms, Gärtner-Ellis theorem, etc.)
- However, the weak topology is not very interesting. For example, subgraph counts are not continuous with respect to the weak topology.
- The LDP for the topology of the cut metric does not follow via standard methods.

Szemerédi's lemma

- Let $G = (V, E)$ be a simple graph of order n.
- For any $X, Y \subseteq V$, let $e_G(X, Y)$ be the number of X-Y edges of G and let
$$\rho_G(X, Y) := \frac{e_G(X, Y)}{|X||Y|}$$
- Call a pair (A, B) of disjoint sets $A, B \subseteq V$ ϵ-regular if all $X \subseteq A$ and $Y \subseteq B$ with $|X| \geq \epsilon|A|$ and $|Y| \geq \epsilon|B|$ satisfy $|\rho_G(X, Y) - \rho_G(A, B)| \leq \epsilon$.
- A partition $\{V_0, \ldots, V_K\}$ of V is called an ϵ-regular partition of G if it satisfies the following conditions: (i) $|V_0| \leq \epsilon n$; (ii) $|V_1| = |V_2| = \cdots = |V_K|$; (iii) all but at most ϵK^2 of the pairs (V_i, V_j) with $1 \leq i < j \leq K$ are ϵ-regular.

Theorem (Szemerédi's lemma)

Given $\epsilon > 0$, $m \geq 1$ there exists $M = M(\epsilon, m)$ such that every graph of order $\geq M$ admits an ϵ-regular partition $\{V_0, \ldots, V_K\}$ for some $K \in [m, M]$.

Finishing the proof using Szemerédi's lemma

- Suppose G is a graph of order n with ϵ-regular partition $\{V_0, \ldots, V_K\}$.
- Let G' be the random graph with independent edges where a vertex $u \in V_i$ is connected to a vertex $v \in V_j$ with probability $\rho_G(V_i, V_j)$.
- Using Szemerédi's regularity lemma, one can prove that $\delta_\square(G, G') \simeq 0$ with high probability if K and n are appropriately large and ϵ is small.
- Let f be the probability density of the law of $G(n,p)$ with respect to the law of G'. (This is easily computed; gives rise to the entropy function.) Then

$$\mathbb{P}(G(n,p) \approx G) \approx f(G)\mathbb{P}(G' \approx G) \approx f(G).$$

- Since the space $\widetilde{\mathcal{W}}$ is compact, this allows us to approximate $\mathbb{P}(G(n,p) \in A)$ for any nice set A by approximating A as a finite union of small balls.

Solution of general ERGMs

- Let $T : \mathcal{W} \to \mathbb{R}$ be a bounded continuous function on the pseudometric space $(\mathcal{W}, \delta_\square)$.
- Let $\mathcal{G}_n$ denote the set of simple graphs on n vertices.
- Then T induces a probability mass function p_n on $\mathcal{G}_n$:

$$p_n(G) := \exp(n^2 T(G) - n^2 \psi_n),$$

where ψ_n is the normalizing constant.

Theorem (Chatterjee & Diaconis, 2011)

For $h \in \mathcal{W}$, define

$$I(h) := \frac{1}{2} \iint [h(x,y) \log h(x,y) + (1 - h(x,y)) \log(1 - h(x,y))] dx dy.$$

Then $\lim_{n\to\infty} \psi_n = \sup_{h \in \mathcal{W}} (T(h) - I(h))$.

Summary

- Theory of graph limits gives a framework for proving the Large Deviation Principle for the Erdős-Rényi random graph $G(n,p)$.
- The LDP for $G(n,p)$ can be used to count the number of graphs with a given property, such as a prescribed number of edges and triangles.
- The graph counting theorem allows us to evaluate normalizing constants and maximum likelihood estimates in exponential random graph models and understand their qualitative behavior. The solutions involve variational problems.
- The general theorems can be specialized in simple situations to give useful byproducts such as proofs of degeneracy and other qualitative phenomena observed by practitioners.
- Another application: limits of graphs with a given degree sequence (joint work with Persi Diaconis and Allan Sly).
- Main open question: How to solve the variational problems in complicated models?

A future direction: alternating sign ERGMs

- As we saw, the edge-triangle model does not exhibit transitivity.
- Alternating sign ERGMs, introduced by Snijders et. al. (2006), attempt to do this.
- A one-parameter example:

$$p_\beta(G) \propto \exp\left(\beta E - \frac{\beta L}{n} + \frac{\beta \Delta}{n}\right),$$

  where E, L and Δ are the number of edges, 2-stars and triangles in G.
- In this model, we can prove that if β is large, then the vertices automatically divide into two groups of roughly equal size, so that two vertices in the same group are connected by an edge with probability $\simeq 1$, while two vertices in different groups are connected with probability $\simeq 1/2$.