BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 53, Number 4, October 2016, Pages 617–642
http://dx.doi.org/10.1090/bull/1539
Article electronically published on June 24, 2016

AN INTRODUCTION TO LARGE DEVIATIONS FOR RANDOM GRAPHS

SOURAV CHATTERJEE

Abstract. This article gives an overview of the emerging literature on large deviations for random graphs. Written for the general mathematical audience, the article begins with a short introduction to the theory of large deviations. This is followed by a description of some large deviation questions about random graphs and an outline of the recent progress on this topic. A more elaborate discussion follows, with a brief account of graph limit theory and its application in constructing a large deviation theory for dense random graphs. The role of Szemerédi's regularity lemma is explained, together with a sketch of the proof of the main large deviation result and some examples. Applications to exponential random graph models are briefly touched upon. The remainder of the paper is devoted to large deviations for sparse graphs. Since the regularity lemma is not applicable in the sparse regime, new tools are needed. Fortunately, there have been several new breakthroughs that managed to achieve the goal by an indirect method. These are discussed, together with an exposition of the underlying theory. The last section contains a list of open problems.

Contents

Received by the editors April 22, 2016.
2010 *Mathematics Subject Classification.* Primary 60F10, 05C80; Secondary 60C05, 05A20.
The author's research was partially supported by NSF grant DMS-1441513.

©2016 American Mathematical Society

1. Large deviations

The theory of large deviations aims to study two things: (a) the probabilities of rare events, and (b) the conditional probabilities of various events given that some rare event has occurred (that is, what would the world look like if some rare event happens?). Often, the second question is more interesting than the first, but it is usually essential to answer the first question to be able to understand how to approach the second.

By way of illustration, consider the following simple example. Toss a fair coin n times, where n is a large number. Under normal circumstances, you expect to get approximately $n/2$ heads. Also, you expect to get roughly $n/4$ pairs of consecutive heads. However, suppose that the following rare event occurs: the tosses yield $\geq 2n/3$ heads. General-purpose tools from the theory of large deviations allows us to compute that the probability of this rare event is

$$e^{-n\log(2^{5/3}/3)(1+o(1))} \tag{1.1}$$

as $n \to \infty$. Moreover, it can be shown that if this rare event has occurred, then it is highly likely that there are approximately $4n/9$ pairs of consecutive heads instead of the usual $n/4$.

So how is the estimate (1.1) obtained? The argument goes roughly as follows. Let $X_1, \ldots, X_n$ be independent random variables, such that $\mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1) = 1/2$ for each i. Then the number of heads in n tosses of a fair coin can be modeled by the sum $S_n := X_1 + \cdots + X_n$. For any $\theta \geq 0$,

$$\begin{aligned}
\mathbb{P}(S_n \geq 2n/3) &= \mathbb{P}(e^{\theta S_n} \geq e^{2\theta n/3}) \\
&\leq \frac{\mathbb{E}(e^{\theta S_n})}{e^{2\theta n/3}} \quad \text{(Markov's inequality)} \\
&= \frac{\mathbb{E}(\prod_{i=1}^n e^{\theta X_i})}{e^{2\theta n/3}} = \frac{\prod_{i=1}^n \mathbb{E}(e^{\theta X_i})}{e^{2\theta n/3}} \quad \text{(independence)} \\
&= e^{-2\theta n/3}\left(\frac{1+e^\theta}{2}\right)^n.
\end{aligned}$$

Optimizing over θ gives the desired upper bound. To prove the lower bound, take some $\epsilon > 0$ and define a random variable Z as

$$Z := \begin{cases} 1 & \text{if } 2n/3 \leq S_n \leq (1+\epsilon)2n/3, \\ 0 & \text{otherwise.} \end{cases}$$

Then for any $\theta \geq 0$,

$$\begin{aligned}
\mathbb{P}(S_n \geq 2n/3) &\geq \mathbb{E}(Z) \\
&\geq e^{-\theta(1+\epsilon)2n/3}\mathbb{E}(e^{\theta S_n} Z) \quad \text{(since } S_n \leq (1+\epsilon)2n/3 \text{ when } Z \neq 0) \\
&= e^{-\theta(1+\epsilon)2n/3}\left(\frac{1+e^\theta}{2}\right)^n \frac{\mathbb{E}(e^{\theta S_n} Z)}{\mathbb{E}(e^{\theta S_n})}.
\end{aligned}$$

The proof is completed by showing that if θ is chosen to be the same number that optimized the upper bound and ϵ is sent to zero sufficiently slowly as $n \to \infty$, then

$$\frac{\mathbb{E}(e^{\theta S_n} Z)}{\mathbb{E}(e^{\theta S_n})} = e^{o(n)}.$$

Establishing the above claim is the most nontrivial part of the whole argument, but is by now standard. This is sometimes called the "change of measure trick".

The above example has a built-in linearity, which allowed us to explicitly compute $\mathbb{E}(e^{\theta S_n})$. Generalizing this idea, classical large deviation theory possesses a collection of powerful tools to deal with linear functionals of independent random variables, random vectors, random functions, random probability measures, and other abstract random objects. The classic text of Dembo and Zeitouni [23] contains an in-depth introduction to this broad area.

2. The problem with nonlinearity

Despite the remarkable progress with linear functionals, there are no general tools for large deviations of nonlinear functionals. Nonlinearity arises naturally in many contexts. For instance, the analysis of real-world networks has been one of the most popular scientific endeavors in the last two decades, and rare events on networks are often nonlinear in nature. This is demonstrated by the following simple example.

Construct a random graph on n vertices by putting an undirected edge between any two with probability p, independently of each other. This is known as the Erdős–Rényi $G(n,p)$ model, originally defined in [25]. The model is too simplistic to be a model for any real-world network, but it has many nice mathematical properties and has led to the developments of many new techniques in combinatorics and probability theory over the years. One can ask the following large deviation questions about this model:

(a) What is the probability that the number of triangles in a $G(n,p)$ random graph is at least $1+\delta$ times the expected value of the number of triangles, where δ is some given number?
(b) What is the most likely structure of the graph if we know that the above rare event has occurred?

This is an example of a nonlinear problem because the number of triangles in $G(n,p)$ is a degree three polynomial of independent random variables. To see this, let $\{1,\dots,n\}$ be the set of vertices, and let X_{ij} be the random variable that is 1 if the edge $\{i,j\}$ is present in the graph and 0 if not. Then $(X_{ij})_{1\le i<j\le n}$ are independent random variables, and the number of triangles is nothing but

$$\frac{1}{6}\sum_{i,j,k=1}^{n} X_{ij}X_{jk}X_{ki}\,,$$

which is a polynomial of degree three. Until even a few years ago, large deviations theory did not have the tools to answer such basic questions about nonlinear functions of independent random variables, although a number of powerful concentration inequalities were available for computing upper and lower bounds on tail probabilities [35, 37, 48, 50].

3. Recent developments

The large deviation theory for the Erdős–Rényi random graph was developed fairly recently in [20], taking to completion a program initiated in the unpublished manuscript [7]. The theory brought together ideas from classical large deviations

theory and tools from combinatorics and graph theory, such as Szemerédi's regularity lemma and the theory of graph limits. The calculations dictated by the theory led to surprising conclusions, even in the simplest of applications such as the following. Let $T_{n,p}$ be the number of triangles in the Erdős–Rényi graph $G(n,p)$. What is the most likely structure of the graph if the rare event

$$E := \{T_{n,p} \geq (1+\delta)\mathbb{E}(T_{n,p})\}$$

happens, where δ is a given positive constant? For instance, are all the extra triangles likely to be arising from a small subset of vertices with high connectivity amongst themselves? Or do they occur because the graph has an excess number of edges spread uniformly?

Surprisingly, the large deviation theory of [20] implies that both scenarios can happen. If p is smaller than a threshold, then there exist $0 < \delta_1 < \delta_2$ such that if $0 < \delta \leq \delta_1$ or $\delta \geq \delta_2$, then conditional on the event E, the graph behaves like $G(n,r)$ for some $r > p$; and if $\delta_1 < \delta < \delta_2$, then the conditional structure is *not* like an Erdős–Rényi graph.

In other words, if the number of triangles exceeds the expected value by a little bit or by a lot, then the most likely scenario is that there is an excess number of edges spread uniformly; and if the surplus amount belongs to a middle range, then the structure of the graph is likely to be inhomogeneous. There is probably no way that the above result could have been guessed from intuition; it was derived purely from a set of mathematical formulas.

The general theory of [20] and its main results are described in Section 5, after a brief introduction to graph limit theory in Section 4.

The large deviation theory for the Erdős–Rényi model has been extended to other, more realistic models of random graphs. For example, it was applied to exponential random graph models in [19] and a number of subsequent papers. These models are widely used in the analysis of real social networks. A brief discussion of applications to exponential random graphs is given in Section 6.

Despite its successes, the theory developed in [20] has one serious limitation: it applies only to dense graphs. A graph is called dense if the average vertex degree is comparable to the total number of vertices (recall that the number of neighbors of a vertex is called its degree). For example, in the Erdős–Rényi model with $n = 10000$ and $p = .3$, the average degree is roughly 3000. This is not true for real networks, which are usually sparse. Unfortunately, the graph theoretic tools used for the analysis of large deviations for random graphs are useful only in the dense setting. Despite considerable progress in developing a theory of sparse graph limits [6, 11, 12], there is still no result that fully captures the power of Szemerédi's lemma in the sparse setting. In the absence of such tools, a nascent theory of "nonlinear large deviations", developed in [17], has been helpful in solving some questions about large deviations for sparse random graphs [5, 40]. This theory is discussed in Sections 7 through 10.

4. Graph limit theory

A beautiful unifying theory of graph limits has been developed by Laszlo Lovász and coauthors in recent years [8–10, 38, 39]. For connections with the theory of exchangeable arrays in probability theory, see [1, 24, 27, 29]. This section contains a brief review of some of the basic definitions and results from this theory.

Let $\{G_n\}_{n\geq 1}$ 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 (that is, edge-preserving maps from $V(H)$ into $V(G)$, where $V(H)$ and $V(G)$ are the vertex sets). As an example, note that if H is a triangle, then $\hom(H,G)$ is the number of triangles in G multiplied by six.

The number of homomorphisms 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.

Suppose that $t(H,G_n)$ tends to a limit $t(H)$ for every H. Then Lovász and Szegedy [39] 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, as follows. If H is a simple graph on $\{1,2,\ldots,k\}$ with edge set $E(H)$ and $f \in \mathcal{W}$, let

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

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 4.1 (Limit of Erdős–Rényi graphs)**.** Fix $p \in (0,1)$ and let $G_{n,p}$ be a random graph from the Erdős-Rényi $G(n,p)$ model. For any fixed graph H, it is not difficult to show that with probability 1,

$$t(H,G_{n,p}) \to p^{|E(H)|} \quad \text{as } n\to\infty.$$

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 $\{G_{n,p}\}_{n\geq 1}$ converges with probability 1 to the nonrandom limit function $f(x,y) \equiv p$ as $n\to\infty$.

The elements of $\mathcal{W}$ are sometimes called *graphons*. A finite simple graph G on $\{1,\ldots,n\}$ can be represented as a graphon f^G in a natural way:

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

Note that this allows *all* simple graphs, irrespective of the number of vertices, to be represented as elements of the single abstract space $\mathcal{W}$. The starting point of graph limit theory is to define a suitable topology on this space. The first step in defining this topology is to recall the cut distance of Frieze and Kannan [26]:

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

where the supremum is taken over all measurable subsets S and T of $[0,1]$. The next step is to introduce an equivalence relation on $\mathcal{W}$, by declaring 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 $\{g_\sigma\}$. Let $\widetilde{\mathcal{W}}$ be the quotient space,

and let τ denote the quotient map $g \mapsto \widetilde{g}$. Since $d_\square$ is invariant under σ, one can define on $\widetilde{\mathcal{W}}$ an induced metric $\delta_\square$ 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 $(\widetilde{\mathcal{W}}, \delta_\square)$ a metric space. This is the abstract space of graph limits. To any finite graph G, 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 4.2 ([9]). *A sequence of graphs $\{G_n\}_{n\ge1}$ 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 4.3 ([39]). *The space $\widetilde{\mathcal{W}}$ is compact under the metric $\delta_\square$.*

The main ingredient in the proofs of the above results is the famous regularity lemma of Szemerédi [47]. This will be discussed in the next section.

5. Large deviations for dense random graphs

We will continue to use the notations and terminologies introduced in the previous section. Fix $p \in (0,1)$. For $u \in [0,1]$, let

$$I_p(u) := u \log \frac{u}{p} + (1-u) \log \frac{1-u}{1-p}, \tag{5.1}$$

with the convention that $0 \log 0 = 0$. For $h \in \mathcal{W}$, let

$$I_p(h) := \int_{[0,1]^2} I_p(h(x,y))\, dx\, dy. \tag{5.2}$$

Finally, for $\widetilde{h} \in \widetilde{\mathcal{W}}$, let $I_p(\widetilde{h}) := I_p(h)$ where h is any element of $\widetilde{h}$. A lemma in [20] shows that the right-hand side does not depend on the choice of h in $\widetilde{h}$, which ensures that the above definition makes sense.

The Erdős–Rényi $G(n,p)$ model induces a probability measure $\widetilde{\mathbb{P}}_{n,p}$ on the space $\widetilde{\mathcal{W}}$ through the map $G \to \widetilde{G}$. The main result of [20] is the following large deviation principle for $\widetilde{\mathbb{P}}_{n,p}$.

Theorem 5.1 ([20]). *For any closed set $\widetilde{F} \subseteq \widetilde{\mathcal{W}}$,*

$$\limsup_{n\to\infty} \frac{2}{n^2} \log \widetilde{\mathbb{P}}_{n,p}(\widetilde{F}) \le - \inf_{\widetilde{h}\in\widetilde{F}} I_p(\widetilde{h}),$$

and for any open set $\widetilde{U} \subseteq \widetilde{\mathcal{W}}$,

$$\liminf_{n\to\infty} \frac{2}{n^2} \log \widetilde{\mathbb{P}}_{n,p}(\widetilde{U}) \ge - \inf_{\widetilde{h}\in\widetilde{U}} I_p(\widetilde{h}).$$

Although the above theorem looks like a result specifically about the $G(n,p)$ model, this is somewhat misleading. Theorem 5.1 actually allows us to approximately count the number of simple graphs on n vertices that have any given property, as long as the property is nicely behaved with respect to the cut metric. This can be made precise as follows. 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

$$I(u) := u \log u + (1-u)\log(1-u)\,.$$

For any $\widetilde{h} \in \widetilde{\mathcal{W}}$, let

$$I(\widetilde{h}) := \int_{[0,1]^2} I(h(x,y))\,dx\,dy\,, \tag{5.3}$$

where h is any element of $\widetilde{h}$. The following corollary can be easily derived from Theorem 5.1, by taking $p = 1/2$.

Corollary 5.2. *For any measurable $\widetilde{A} \subseteq \widetilde{\mathcal{W}}$,*

$$\begin{aligned} -\inf_{\widetilde{h}\in \mathrm{cl}(\widetilde{A})} I(\widetilde{h}) &\geq \limsup_{n\to\infty} \frac{2\log|\widetilde{A}_n|}{n^2} \\ &\geq \liminf_{n\to\infty} \frac{2\log|\widetilde{A}_n|}{n^2} \geq -\inf_{\widetilde{h}\in \mathrm{int}(\widetilde{A})} I(\widetilde{h})\,, \end{aligned}$$

where $\mathrm{cl}(\widetilde{A})$ *is the closure of* $\widetilde{A}$ *and* $\mathrm{int}(\widetilde{A})$ *is the interior of* $\widetilde{A}$.

Under very special circumstances, the variational problems of Theorem 5.1 and Corollary 5.2 are known to have explicit solutions. For example, it follows from Corollary 5.2 and some results in [20] that the number of graphs on n vertices with at least tn^3 triangles is

$$e^{\frac{1}{2}n^2 f(t)(1+o(1))}$$

as $n \to \infty$, where

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

On the other hand, for the number of graphs with *at most* tn^3 triangles, such an explicit formula can be obtained if t is sufficiently away from zero, and it can also be shown that this formula *does not* hold if t is sufficiently close to zero. As of now, there is no explicit formula for small t. See Zhao [52] for the most advanced results about the lower tail problem.

Theorem 5.1 gives estimates of the probabilities of rare events related to an Erdős–Rényi graph. However, it does not answer the second type of question mentioned in the introduction, that is, given that some particular rare event has occurred, what does the graph look like? The answer may be roughly described as follows. If $G_{n,p}$ is a random graph from the $G(n,p)$ model, $\widetilde{F}$ is a closed subset of $\widetilde{\mathcal{W}}$, and we condition on the event that $\widetilde{G}_{n,p} \in \widetilde{F}$, then it is very likely that $\widetilde{G}_{n,p}$ is close to one of the elements of $\widetilde{F}$ that minimize I_p in $\widetilde{F}$. The precise result goes as follows. Let $\widetilde{F}$ and $G_{n,p}$ be as above. Let $\widetilde{F}^*$ be the set of minimizers of I_p in $\widetilde{F}$. Suppose that

$$\inf_{\widetilde{h}\in \mathrm{int}(\widetilde{F})} I_p(\widetilde{h}) = \inf_{\widetilde{h}\in \widetilde{F}} I_p(\widetilde{h}) > 0\,. \tag{5.4}$$

It was proved in [20] that I_p is lower semicontinuous on $\widetilde{\mathcal{W}}$, and we know that $\widetilde{\mathcal{W}}$ is compact by Theorem 4.3. Therefore, $\widetilde{F}^*$ is nonempty. For any $\widetilde{h} \in \widetilde{\mathcal{W}}$, let

$$\delta_\square(\widetilde{h}, \widetilde{F}^*) := \inf_{\widetilde{f}\in\widetilde{F}^*} \delta_\square(\widetilde{h}, \widetilde{f})\,. \tag{5.5}$$

Theorem 5.3 ([20]). *In the above setting, the following inequality holds for each $n \ge 1$ and $\epsilon > 0$:*

$$\mathbb{P}(\delta_\square(\widetilde{G}_{n,p}, \widetilde{F}^*) \ge \epsilon \mid \widetilde{G}_{n,p} \in \widetilde{F}) \le e^{-C(\epsilon,\widetilde{F})n^2},$$

where $C(\epsilon, \widetilde{F})$ is a positive constant depending only on ϵ and $\widetilde{F}$.

In particular, it follows that if $\widetilde{F}^*$ contains only one element, $\widetilde{h}^*$, then the conditional distribution of $\widetilde{G}_{n,p}$ given $\widetilde{G}_{n,p} \in \widetilde{F}$ converges to the point mass at $\widetilde{h}^*$ as $n \to \infty$, giving a conditional law of large numbers.

Let us now see a sketch of the proof of Theorem 5.1. The result can be proved by standard techniques for the weak topology on $\widetilde{\mathcal{W}}$. However, the weak topology is not very interesting. For example, subgraph counts are not continuous with respect to the weak topology. The large deviation principle under the topology of the cut metric (Theorem 5.1) does not follow via standard methods.

The main tool for proving Theorem 5.1 is Szemerédi's regularity lemma. One version of Szemerédi's lemma goes as follows. Let $G = (V, E)$ be a simple graph of order n (recall that the number of vertices of a graph is called its order). For any $X, Y \subseteq V$, let $e_G(X, Y)$ be the number of X-Y edges of G and let

$$\rho_G(X, Y) := \frac{e_G(X, Y)}{|X||Y|},$$

where $|X|$ and $|Y|$ are the sizes of X and 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 $|X| \ge \epsilon|A|$ and $|Y| \ge \epsilon|B|$ satisfy

$$|\rho_G(X, Y) - \rho_G(A, B)| \le \epsilon.$$

The concept of ϵ-regularity tries to capture the notion that the edges going between A and B behave like randomly distributed edges with density $\rho_G(A, B)$.

A partition $\{V_0, \ldots, V_K\}$ of V is called an ϵ-regular partition of G if it satisfies the following conditions:

(i) $|V_0| \le \epsilon n$.
(ii) $|V_1| = |V_2| = \cdots = |V_K|$.
(iii) All but at most ϵK^2 of the pairs (V_i, V_j), $1 \le i < j \le K$, are ϵ-regular.

Theorem 5.4 (Szemerédi's regularity lemma [47]). *Given any $\epsilon > 0$ and $m \ge 1$, there exists $M = M(\epsilon, m)$ such that every graph of order $\ge M$ admits an ϵ-regular partition $\{V_0, \ldots, V_K\}$ for some $K \in [m, M]$.*

Roughly speaking, Szemerédi's lemma says that any large graph G may be partitioned into blocks of equal size (plus one exceptional block of small size) so that the edges going between the blocks behave like randomly distributed edges, and the number of blocks depends only on the desired degree of randomness and not on the size of the graph. The key to proving Theorem 5.1 is to formulate a precise version of this statement. This goes as follows.

Choose a small number $\epsilon > 0$. Suppose that $G = (V, E)$ is a simple graph of order n with ϵ-regular partition $\{V_0, \ldots, V_K\}$, as in Theorem 5.4. Let $G' = (V, E')$ be a random graph with independent edges where a vertex $u \in V_i$ is connected to a vertex $v \in V_j$ with probability $\rho_G(V_i, V_j)$. Let $\widetilde{\mu}$ be the probability measure on $\widetilde{\mathcal{W}}$ induced by G'. The main step in the proof of Theorem 5.1, proved using Szemerédi's lemma, is that $\widetilde{\mu}(\widetilde{B}) \approx 1$, where $\widetilde{B}$ is a small ball around $\widetilde{G}$ in the $\delta_\square$ metric and the radius of $\widetilde{B}$ depends only on ϵ and not on the size of G. In

other words, the random graph G' is close to the given graph G in the $\delta_\square$ metric. This gives a precise meaning to the sentence from the previous paragraph that a large graph can be partitioned into blocks with approximately randomly distributed edges between the blocks.

To complete the sketch of the proof of Theorem 5.1, let f be the probability density of $\widetilde{\mathbb{P}}_{n,p}$ with respect to $\widetilde{\mu}$. Since the edges in both $G(n,p)$ and G' are independent, this probability density is easy to compute. Since $\widetilde{B}$ is a ball of small radius, we get

$$\widetilde{\mathbb{P}}_{n,p}(\widetilde{B}) \approx f(\widetilde{G})\widetilde{\mu}(\widetilde{B}) \approx f(\widetilde{G})\,.$$

Since the space $\widetilde{\mathcal{W}}$ is compact (Theorem 4.3), this allows us to estimate $\widetilde{\mathbb{P}}(\widetilde{A})$ for any nice set $\widetilde{A}$ by approximating $\widetilde{A}$ as a finite union of small balls.

Example 5.5 (Triangles in dense Erdős–Rényi graphs)**.** For an application of Theorem 5.1 to a concrete problem, consider the number of triangles $T_{n,p}$ in a $G(n,p)$ random graph. Recall that $\mathcal{W}$ is the space of symmetric measurable functions from $[0,1]^2$ into $[0,1]$. For each $f \in \mathcal{W}$, let

$$T(f) := \frac{1}{6}\int_{[0,1]^3} f(x,y)f(y,z)f(z,x)\,dx\,dy\,dz\,,$$

and let $I_p(f)$ be as in (5.2). For each $p \in (0,1)$ and $t \geq 0$, let

$$\phi_p(t) := \inf\{I_p(f) : f \in \mathcal{W},\, T(f) \geq t\}. \tag{5.6}$$

In [20] the following result was proved using Theorem 5.1.

Theorem 5.6 ([20])**.** *For each $p \in (0,1)$ and each $t \geq 0$,*

$$\lim_{n\to\infty} \frac{2}{n^2}\log \mathbb{P}(T_{n,p} \geq tn^3) = -\phi_p(t).$$

Moreover, the infimum is attained in the variational problem (5.6)*.*

The way to deduce Theorem 5.6 from Theorem 5.1 is to transform the problem into a question about the probability measure $\widetilde{\mathbb{P}}_{n,p}$. This is quite straightforward. The next step is to apply Theorem 5.1 to an appropriate pair of sets $(\widetilde{F}, \widetilde{U})$ and show that the upper and lower bounds match.

As a consequence of Theorem 4.3, one can show that the infimum in (5.6) is attained. Moreover, the minimizing functions determine the behavior of the graph conditional on the event $\{T_{n,p} \geq tn^3\}$. Theorem 4.2 is instrumental in proving such claims. For example, if the minimization problem is solved by constant functions, then the conditional behavior continues to be like that of an Erdős–Rényi graph but with a different edge probability. On the other hand, if the minimizers are all nonconstant functions, then the conditional behavior is not like that of an Erdős–Rényi graph. The precise statements and proofs of these claims can be formalized using Theorem 5.3.

The variational problem (5.6) was analyzed in [20] and [41]. One of the most surprising findings, already mentioned briefly in Section 3, is that if p is smaller than a threshold, then there exists a nonempty interval (t_1, t_2) such that the minimization problem is solved by constant functions when $t \notin (t_1, t_2)$, and it is solved by nonconstant functions when $t \in (t_1, t_2)$. The existence of this interval was proved in [20] and the values of t_1 and t_2 were computed in [41]. The theorem proved in [41] applies to the homomorphism density of any regular graph; specialized to triangles, it says the following.

Theorem 5.7 ([41]). *Take any $p \in (0,1)$ and $t \in (p^3/6, 1)$. Let $r := (6t)^{1/3}$. If the point $(r^2, I_p(r))$ lies on the convex minorant of the function $J_p(x) := I_p(x^{1/2})$, then there is a unique solution of the variational problem* (5.6)*, and it is the constant function $f(x,y) \equiv r$. On the other hand, if $(r^2, I_p(r))$ does not lie on the convex minorant of J_p, then any solution of* (5.6) *is nonconstant.*

Applications of Theorem 5.1 to exponential random graphs are briefly discussed in the next section before moving on to large deviations for sparse random graphs.

6. Exponential random graphs

Let $\mathcal{G}_n$ be the set of all simple graphs on n labeled vertices. A variety of probability models on this set can be presented in exponential form

$$p_\beta(G) = \exp\biggl(\sum_{i=1}^k \beta_i T_i(G) - \psi(\beta)\biggr),$$

where $\beta = (\beta_1, \ldots, \beta_k)$ is a vector of real parameters, $T_1, T_2, \ldots, T_k$ are real-valued functions on $\mathcal{G}_n$, and $\psi(\beta)$ is the normalizing constant. Usually, T_i are taken to be counts of various subgraphs, for example $T_1(G) =$ number of edges in G, $T_2(G) =$ number of triangles in G, etc. These are known as exponential random graph models (ERGM). These models are widely used in the study of social networks, but were generally mathematically intractable until recently. A general technique for proving theorems about dense exponential random graphs, based on the large deviation theory for dense Erdős–Rényi random graphs, was proposed in [19]. The mathematical literature on exponential random graphs has grown considerably since the publication of [19]. Since this is somewhat disjoint from the literature of large deviations, we will not attempt to survey these developments here. For a comprehensive survey of the ERGM literature up to the publication of [19], see [19]. For a few mathematical results preceding [19], see [4, 18]. For a nonexhaustive list of subsequent developments, see [2, 3, 30, 31, 41–46, 51]. The discussion in this section will be limited to a basic result from [19] and one easy example. We will continue to use the notation introduced in Section 4.

Let T be a real-valued continuous function on the space $\widetilde{\mathcal{W}}$. Define a probability mass function p_n on $\mathcal{G}_n$ induced by T as

$$p_n(G) := e^{n^2(T(\widetilde{G}) - \psi_n)},$$

where ψ_n is a constant such that the total mass of p_n is 1. Explicitly,

$$\psi_n = \frac{1}{n^2} \log \sum_{G \in \mathcal{G}_n} e^{n^2 T(\widetilde{G})}.$$

The coefficient n^2 is meant to ensure that ψ_n tends to a nontrivial limit as n tends to infinity. This setup gives an abstract formulation of exponential random graphs in the language of graph limits, but with the usual limitation that it makes sense only for dense graphs. The following theorem was proved in [19].

Theorem 6.1 ([19]). *Let T and ψ_n be as above. Let I be the function defined in equation* (5.3)*. Then*

$$\lim_{n\to\infty} \psi_n = \sup_{\widetilde{h} \in \widetilde{\mathcal{W}}} \Bigl(T(\widetilde{h}) - \frac{1}{2} I(\widetilde{h})\Bigr). \tag{6.1}$$

The evaluation of the normalizing constant is an important problem in statistical applications of exponential random graphs, because it is required for computing maximum likelihood estimates. Incidentally, even the existence of the limit in Theorem 6.1 has an important consequence. Suppose that a computer program can evaluate the exact value of the normalizing constant for moderate sized n. Then if n is large, one can choose a scaled-down model with a smaller number of nodes, and use the exact value of the normalizing constant in the scaled-down model as an approximation to the normalizing constant in the larger model.

Theorem 6.1 gives an asymptotic formula for ψ_n; however, it says nothing about the behavior of a random graph drawn from the exponential random graph model. Some aspects of this behavior can be described as follows. Let $\widetilde{F}^*$ be the subset of $\widetilde{\mathcal{W}}$ where $T(\widetilde{h}) - \frac{1}{2}I(\widetilde{h})$ is maximized. By the compactness of $\widetilde{\mathcal{W}}$, the continuity of T and the lower semicontinuity of I (proved in [20]), $\widetilde{F}^*$ is a nonempty compact set. Let G_n be a random graph on n vertices drawn from the exponential random graph model defined by T. The following theorem shows that for n large, $\widetilde{G}_n$ must lie close to $\widetilde{F}^*$ with high probability. In particular, if $\widetilde{F}^*$ is a singleton set, then the theorem gives a weak law of large numbers for $\widetilde{G}_n$.

Theorem 6.2 ([19]). *Let $\widetilde{F}^*$ and G_n be defined as in the above paragraph. Then for any $\eta > 0$, there exist $C, \gamma > 0$ such that for any n,*

$$\mathbb{P}(\delta_\square(\widetilde{G}_n, \widetilde{F}^*) > \eta) \le Ce^{-n^2\gamma}.$$

There is no general method for efficiently solving the variational problem (6.1), either analytically or computationally. In some special cases, however, the problem yields an explicit solution. One such example, worked out in [19], is the following.

Example 6.3. Let $H_1, \dots, H_k$ be finite simple graphs, where H_1 is the complete graph on two vertices (that is, just a single edge) and each H_i contains at least one edge. Let $\beta_1, \dots, \beta_k$ be k real numbers. For any $h \in \mathcal{W}$, let

$$T(h) := \sum_{i=1}^{k} \beta_i t(H_i, h), \tag{6.2}$$

where $t(H_i, h)$ is the homomorphism density of H_i in h. The functional T extends naturally to $\widetilde{\mathcal{W}}$, and is continuous on $\widetilde{\mathcal{W}}$ by Theorem 4.2. For any finite simple graph G that has at least as many nodes as the largest of the H_i's,

$$T(\widetilde{G}) = \sum_{i=1}^{k} \beta_i t(H_i, G),$$

where $t(H_i, G)$ is the homomorphism density of H_i in G. For example, if $k = 2$, H_2 is a triangle, and G has at least three nodes, then

$$T(\widetilde{G}) = 2\beta_1 \frac{\text{number of edges in } G}{n^2} + 6\beta_2 \frac{\text{number of triangles in } G}{n^3}.$$

The following theorem says that when T is of the form (6.2) and $\beta_2, \dots, \beta_k$ are nonnegative, the variational problem of Theorem 6.1 can be reduced to a simple maximization problem in one real variable. The theorem moreover says that each solution of the variational problem is a constant function, and there are only a finite number of solutions. By Theorem 6.2, this implies that when $\beta_2, \dots, \beta_k$

are nonnegative, exponential random graphs from this class of models behave like random graphs drawn from a finite mixture of Erdős–Rényi models.

Theorem 6.4 ([19]). *Let $H_1, \ldots, H_k$ and T be as above. Suppose that the parameters $\beta_2, \ldots, \beta_k$ are nonnegative. Let ψ_n be the normalizing constant of the exponential random graph model induced by T on the set of simple graphs on n vertices. Then*

$$\lim_{n\to\infty} \psi_n = \sup_{0\le u\le 1}\biggl(\sum_{i=1}^k \beta_i u^{e(H_i)} - \frac{1}{2}I(u)\biggr), \tag{6.3}$$

where I is the function defined in (5.3) *and $e(H_i)$ is the number of edges in H_i. Moreover, there are only a finite number of solutions of the variational problem of Theorem* 6.1 *for this T, and each solution is a constant function, where the constant solves the scalar maximization problem* (6.3).

7. Large deviations for sparse random graphs

Recall the Erdős–Rényi model $G(n,p)$. Let $T_{n,p}$ be the number of triangles in a $G(n,p)$ random graph, as before. The behavior of the upper tail probabilities of $T_{n,p}$, especially when $p \to 0$ as $n \to \infty$, has been the subject of intense investigation for many years. After a series of successively improving suboptimal results by various authors, a big advance was made by Kim and Vu [36] and simultaneously by Janson et al. [28] who showed that if $p \ge n^{-1}\log n$, then

$$e^{-c_1(\delta)n^2p^2\log(1/p)} \le \mathbb{P}(T_{n,p} \ge (1+\delta)\mathbb{E}(T_{n,p})) \le e^{-c_2(\delta)n^2p^2},$$

where $c_1(\delta)$ and $c_2(\delta)$ are constants that depend only on δ. It took several more years to remove the logarithmic discrepancy between the exponents on the two sides. It was finally established in [16] and independently in [21, 22] that when $p \ge n^{-1}\log n$,

$$e^{-c_1(\delta)n^2p^2\log(1/p)} \le \mathbb{P}(T_{n,p} \ge (1+\delta)\mathbb{E}(T_{n,p})) \le e^{-c_2(\delta)n^2p^2\log(1/p)}.$$

This still left open the question of determining the dependence of the exponent on δ. We have seen the solution of this question in the previous section when p is fixed. But this theory does not carry over to the case where $p \to 0$ as $n \to \infty$, especially if p decays like a negative power of n. Partly motivated by this question, a general theory of nonlinear large deviations was proposed in [17]. Using this theory, Lubetzky and Zhao [40] proved the following theorem, which fully solved the upper tail large deviation problem for $T_{n,p}$ when p goes to zero slower than $n^{-1/42}$.

Theorem 7.1 ([40]). *If $T_{n,p}$ is the number of triangles in $G(n,p)$, then as $n \to \infty$ and $p \to 0$ slower than $n^{-1/42}$,*

$$\mathbb{P}(T_{n,p} \ge (1+\delta)\mathbb{E}(T_{n,p})) = \exp\biggl(-(1+o(1))\min\biggl\{\frac{\delta^{2/3}}{2}, \frac{\delta}{3}\biggr\}n^2p^2\log\frac{1}{p}\biggr).$$

It is conjectured that this result holds when $p \to 0$ slower than $n^{-1/2}$ (see [40]). The above theorem has been generalized by Bhattacharya et al. [5], who got the following beautiful result. Take any finite simple graph H with maximum degree Δ. Let H^* be the induced subgraph of H on all vertices whose degree in H is Δ. Recall that an independent set in a graph is a set of vertices such that no two are connected

by an edge. Also, recall that a graph is called regular if all its vertices have the same degree, and it is irregular otherwise. Define a polynomial

$$P_{H^*}(x) := \sum_k i_{H^*}(k) x^k \,,$$

where $i_{H^*}(k)$ is the number of k-element independent sets in H^*. The main result of [5] is the following.

Theorem 7.2 ([5]). *Let H be a connected finite simple graph on k vertices with maximum degree $\Delta \geq 2$. Then for any $\delta > 0$, there is a unique positive number $\theta = \theta(H, \delta)$ that solves $P_{H^*}(\theta) = 1 + \delta$, where P_{H^*} is the polynomial defined above. Let $H_{n,p}$ be the number of homomorphisms (defined in Section 4) of H into an Erdős–Rényi $G(n,p)$ random graph. Then there is a constant $\alpha_H > 0$ depending only on H, such that if $n \to \infty$ and $p \to 0$ slower than $n^{-\alpha_H}$, then for any $\delta > 0$,*

$$\mathbb{P}(H_{n,p} \geq (1+\delta)\mathbb{E}(H_{n,p})) = \exp\left(-(1+o(1))c(\delta)n^2 p^\Delta \log\frac{1}{p}\right),$$

where

$$c(\delta) = \begin{cases} \min\{\theta, \frac{1}{2}\delta^{2/k}\} & \text{if } H \text{ is regular,} \\ \theta & \text{if } H \text{ is irregular.} \end{cases}$$

The formula given in Theorem 7.2 is more than just a formula. It gives a hint at the conditional structure of the graph and at the nature of phase transitions as δ varies. Unlike the dense case, it is hard to give a precise meaning to claims about the conditional structure in the sparse setting due to the lack of an adequate sparse graph limit theory. For a detailed discussion, see [5].

The paper [5] also gives a number of examples where the coefficient $c(\delta)$ in Theorem 7.2 can be explicitly computed. For instance, if $H = C_4$, the cycle of length four, then

$$c(\delta) = \begin{cases} \frac{1}{2}\sqrt{\delta} & \text{if } \delta < 16, \\ -1 + \sqrt{1 + \frac{1}{2}\delta} & \text{if } \delta \geq 16. \end{cases}$$

Theorem 7.1 is also a special case of Theorem 7.2.

8. The low complexity gradient condition

The purpose of this section is to begin to describe the theory developed in [17] that leads to the results of the previous section. The initial part of the discussion will be kept intentionally imprecise so as to convey the ideas smoothly without getting bogged down in technical details and notational complexities. The precise statement of the main result is given in in the next section.

Take a smooth $f : [0,1]^n \to \mathbb{R}$. Let $Y = (Y_1, \ldots, Y_n)$ be a vector of independent 0-1 random variables with $\mathbb{P}(Y_i = 1) = p$. The goal is to find an approximation for the upper tail probability $\mathbb{P}(f(Y) \geq tn)$ when t is bigger than $n^{-1}\mathbb{E}(f(Y))$. Recall the function $I_p : [0,1] \to \mathbb{R}$ defined in (5.1). For $x = (x_1, \ldots, x_n)$, define

$$I_p(x) := \sum_{i=1}^n I_p(x_i) \,.$$

For each $t \in \mathbb{R}$, let

$$\phi_p(t) := \frac{1}{n} \inf\{I_p(x) \,:\, x \in [0,1]^n,\, f(x) \geq tn\} \,. \tag{8.1}$$

In many problems, it turns out that

$$\mathbb{P}(f(Y) \geq tn) \approx e^{-\phi_p(t)n}. \tag{8.2}$$

The approximation (8.2) holds in great generality for linear functions. Theorem 5.6 gives a nonlinear example. Theorems 7.1 and 7.2 are consequences of this approximation, where $\phi_p(t)$ is explicitly computable in a limit. The main result of [17], building on ideas developed in [14, 15, 18], gives a sufficient condition under which (8.2) is valid for a general nonlinear f. The condition may be roughly stated as follows. Let ∇f be the gradient of f, so that ∇f is a map from $[0,1]^n$ into $\mathbb{R}^n$. Then we need the image of $[0,1]^n$ under the map ∇f to have *low complexity*, in the sense that it can be covered by a relatively small number of balls of an appropriate radius. Here *small* means $e^{o(n)}$. This is called the *low complexity gradient condition* in [17]. A different way of putting this is to say that the value of $\nabla f(x)$ may be approximately encoded by $o(n)$ bits of information.

To understand this, let $f(x) = \frac{1}{2}\|x\|^2$, where $\|x\|$ is the Euclidean norm of x. Then $\nabla f(x) = x$, and so $\nabla f([0,1]^n) = [0,1]^n$. Since the typical distance between points in $[0,1]^n$ is of order $\sqrt{n}$, it is appropriate to consider balls whose radii are of order $\sqrt{n}$ when measuring the complexity on $[0,1]^n$. An $\epsilon\sqrt{n}$-net of $[0,1]^n$ is a collection of points in $[0,1]^n$ such that any point of $[0,1]^n$ is within distance $\epsilon\sqrt{n}$ from one of these points. It is not difficult to prove that for any fixed ϵ, an $\epsilon\sqrt{n}$-net of $[0,1]^n$ must have size at least $e^{C(\epsilon)n}$, where $C(\epsilon)$ is a positive constant that depends on ϵ. Therefore the gradient of f does not have low complexity.

On the other hand, suppose that $f : [0,1]^n \to \mathbb{R}$ is a linear map. Then $\nabla f(x)$ is the same for every x, and hence the image of $[0,1]^n$ under ∇f consists of a single point. This is a set of the lowest possible complexity, and therefore the low complexity gradient condition is satisfied by linear functions. We will see less trivial examples of functions satisfying the low complexity gradient condition below, after formally defining a measure of the complexity of ∇f. This measure of complexity is chiefly for expositional purposes; the main theorem of [17], presented later in this section, does not make any direct use of this complexity measure.

Implicitly, we have a sequence of functions rather than a single function; that is, the function f in (8.2) depends on n and varies as n varies. We will assume that f is scaled with n in such a way that sizes of the partial derivatives of f do not go to zero or infinity as $n \to \infty$. Consequently, the typical diameter of $\nabla f([0,1]^n)$ is of order $\sqrt{n}$. Keeping this in mind, the complexity of $\nabla f([0,1]^n)$ may be defined as follows. For each ϵ, let $\mathcal{D}(\epsilon)$ be an $\epsilon\sqrt{n}$-net for the image of $[0,1]^n$ under ∇f. As explained before, this is a finite set of points such that for any $x \in [0,1]^n$, there exists a point $z \in \mathcal{D}(\epsilon)$ such that the Euclidean distance between $\nabla f(x)$ and z is less than $\epsilon\sqrt{n}$. Let $|\mathcal{D}(\epsilon)|$ denote the size of $\mathcal{D}(\epsilon)$. Suppose that $\mathcal{D}(\epsilon)$ is optimized to have the minimum possible size. Define

$$C(f) := \inf_{\epsilon > 0}\left(\epsilon + \frac{\log|\mathcal{D}(\epsilon)|}{n}\right).$$

Note that $C(f)$ is small if and only if there exists a small ϵ such that $\log|\mathcal{D}(\epsilon)| \ll n$. The number $C(f)$ can therefore be used as a measure of the complexity of the gradient of f when f scales with n is such a way that the sizes of its partial derivatives do not tend to zero or infinity as $n \to \infty$. We have already observed that $C(f) = 0$ when f is linear. Let us now see some more examples.

Example 8.1. For $x = (x_1, \ldots, x_n) \in [0,1]^n$, let

$$f(x) := \frac{1}{n} \sum_{1 \le i < j \le n} x_i x_j \,.$$

Then for each i,

$$\frac{\partial f}{\partial x_i} = \frac{1}{n} \sum_{j \ne i} x_j = -\frac{x_i}{n} + \frac{1}{n} \sum_{j=1}^{n} x_j \,.$$

Note that the sizes of the partial derivatives of f are not blowing up or tending to zero as $n \to \infty$. Let D be the set of all vectors in $\mathbb{R}^n$ that are of the form $(k/n, k/n, \ldots, k/n)$ for some $0 \le k \le n-1$. The above formula shows that for any $x \in [0,1]^n$, there exists $z \in D$ such that the Euclidean distance between $\nabla f(x)$ and z is bounded by $2n^{-1/2}$. To see this, just take $z = (k/n, \ldots, k/n)$ where k is the integer part of $\sum_j x_j$. Thus, the set D can serve as $\mathcal{D}(\epsilon)$ for this f, for $\epsilon = 2n^{-1}$. Since $|\mathcal{D}(\epsilon)| = n$, we get

$$C(f) \le \frac{2 + \log n}{n} \,,$$

which shows that f satisfies the low complexity gradient condition. This is easier to understand via the language of encoding $\nabla f(x)$: the value of $\nabla f(x)$ may be approximately encoded by the single quantity $n^{-1} \sum_j x_j$ and therefore needs $O(\log n)$ bits.

Example 8.2. Take any $n \ge 2$, and for $x = (x_1, \ldots, x_n) \in [0,1]^n$, let

$$f(x) := \sum_{i=1}^{n-1} x_i x_{i+1} \,.$$

Then

$$\frac{\partial f}{\partial x_i} = \begin{cases} x_2 & \text{if } i = 1, \\ x_{i-1} + x_{i+1} & \text{if } 2 \le i \le n-1, \\ x_{n-1} & \text{if } i = n. \end{cases}$$

This f, again, is scaled in such a way that the sizes of its partial derivatives remain stable as $n \to \infty$. Now, given any $z \in \{0,1,2\}^n$, it is easy to prove using the above formula that there exists $x \in \{0,1\}^n$ such that $\nabla f(x)$ and z agree on at least $n/3$ coordinates. This, in turn, can be used to show that $C(f)$ does not tend to zero as $n \to \infty$. Therefore, this f does not have a gradient of low complexity. Again, this may be easier to understand by observing that to encode $\nabla f(x)$, we need to essentially have information about all the coordinates of x, which requires n bits.

Example 8.3 (Triangles in $G(n,p)$)**.** Take any n such that $n = m(m-1)/2$ for some positive integer $m \ge 2$. Denote elements of $\mathbb{R}^n$ as $x = (x_{ij})_{1 \le i < j \le m}$. Define $f : \mathbb{R}^n \to \mathbb{R}$ as

$$f(x) = \frac{1}{m} \sum_{i,j,k=1}^{m} x_{ij} x_{jk} x_{ki} \,,$$

using the conventions that $x_{ii} = 0$ for each i and $x_{ij} = x_{ji}$ when $i > j$. Note that when x_{ij} is the (i,j)-th entry of the adjacency matrix of a graph on m vertices, $f(x)$ is equal to the number of triangles in the graph multiplied by $6/m$. Let us now see why $C(f)$ is small. The complete proof is somewhat lengthy, so we will only see a sketch of the proof. For details, see [17].

First, a simple computation gives

$$\frac{\partial f}{\partial x_{ij}} = \frac{3}{m}\sum_{k=1}^{m} x_{ik}x_{jk}\,. \tag{8.3}$$

The above formula shows that the sizes of the partial derivatives of f remain stable as $n \to \infty$. To establish the low complexity gradient condition, we need to show that there is a set of size $e^{o(n)}$ such that for any $x \in [0,1]^n$, $\nabla f(x)$ is within Euclidean distance $o(\sqrt{n})$ from one of the elements of this set. This is proved as a consequence of two key observations:

1. By equation (8.3),

$$\|\nabla f(x) - \nabla f(y)\|^2 = \frac{9}{m^2}\sum_{i,j,k,l}(x_{ik}x_{jk} - y_{ik}y_{jk})(x_{il}x_{jl} - y_{il}y_{jl})\,.$$

Expand the brackets on the right-hand and consider one pair of terms in the expansion, say,

$$\frac{1}{m^2}\sum_{i,j,k,l}(x_{ik}x_{jk}x_{il}x_{jl} - x_{ik}x_{jk}y_{il}y_{jl})\,.$$

This term may be written in a telescoping manner as

$$\frac{1}{m^2}\sum_{i,j,k,l} x_{ik}x_{jk}x_{il}(x_{jl} - y_{jl}) + \frac{1}{m^2}\sum_{i,j,k,l} x_{ik}x_{jk}(x_{il} - y_{il})y_{jl}\,.$$

Let $M(x)$ be the matrix whose (i,j)-th entry is x_{ij}. Consider the first sum. If i and k are fixed, then the sum in j and l is a quadratic form of the matrix $M(x) - M(y)$. This shows that the first sum is bounded by $m\|M(x) - M(y)\|_{\mathrm{op}}$, where $\|A\|_{\mathrm{op}}$ denotes the L^2 operator norm of a matrix A. Similarly bounding other terms, we get

$$\|\nabla f(x) - \nabla f(y)\|^2 \le Cm\|M(x) - M(y)\|_{\mathrm{op}}\,, \tag{8.4}$$

where C is a universal constant.

2. Let $\mathcal{M}$ denote the set of $m \times m$ symmetric matrices with entries in $[0,1]$. This set has low complexity in operator norm, in the sense that it has a subset of size $e^{o(m^2)}$ such that for any $M \in \mathcal{M}$, there is some N in this subset for which $\|M - N\|_{\mathrm{op}} = o(m)$. To see this, observe that if $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of M, then $\sum_i \lambda_i^2 = \mathrm{Tr}(M^2) \le m^2$. This implies that the kth largest eigenvalue is bounded by $mk^{-1/2}$. Therefore there is a rank k matrix N such that $\|M - N\|_{\mathrm{op}} \le mk^{-1/2}$. Since a matrix of rank k is determined by k eigenvectors and eigenvalues, it is easy to see that the set of rank k matrices has low complexity. Letting k grow slowly with m, this allows us to establish the low complexity of $\mathcal{M}$. The proof of the low complexity of ∇f is completed by combining this information with the inequality (8.4).

Example 8.4 (Three-term arithmetic progressions)**.** Index the elements of $[0,1]^n$ as $x = (x_i)_{i \in \mathbb{Z}/n\mathbb{Z}}$, and define $f : [0,1]^n \to \mathbb{R}$ as

$$f(x) := \frac{1}{n}\sum_{i,j \in \mathbb{Z}/n\mathbb{Z}} x_i x_{i+j} x_{i+2j}\,,$$

where the sums $i+j$ and $i+2j$ are carried out modulo n. Note that if A is a subset of $\mathbb{Z}/n\mathbb{Z}$ and $x_i = 1$ if $i \in A$ and 0 otherwise, then $f(x)$ is equal to the number of three-term arithmetic progressions in A divided by n. Now,

$$\frac{\partial f}{\partial x_i} = \frac{1}{n} \sum_{j \in \mathbb{Z}/n\mathbb{Z}} (x_{i+j}x_{i+2j} + x_{i-j}x_{i+j} + x_{i-2j}x_{i-j}) .$$

This shows that the sizes of the partial derivatives of f remain stable as $n \to \infty$. It turns out that this f, too, satisfies the low complexity gradient condition. The proof is not short enough to be presented here in full detail (see [17] for that), but the main idea may be easily explained as follows. The discrete Fourier transform $\hat{x}$ of a point $x \in \mathbb{R}^{\mathbb{Z}/n\mathbb{Z}}$ is defined as

$$\hat{x}_j := \frac{1}{\sqrt{n}} \sum_k x_k e^{2\pi \mathrm{i} jk/n} ,$$

where $\mathrm{i} = \sqrt{-1}$. The following inequality was proved in [17]:

$$\|\nabla f(x) - \nabla f(y)\|^2 \le C n^{1/2} \max_i |\hat{x}_i - \hat{y}_i| .$$

Using this inequality, it is possible to show that if we know a few of the large Fourier coefficients of x, then the vector $\nabla f(x) = (f_1(x), \dots, f_n(x))$ can be approximately recovered. This allows us to establish the low complexity gradient condition for f.

9. Nonlinear large deviations

Having discussed several examples of functions satisfying the low complexity gradient condition, let us now review the main result of [17], which says that the approximation (8.2) holds for such functions. Let us begin with a limiting statement that applies when p is fixed and n tends to infinity. This is not pertinent to large deviations for sparse random graphs, but gives a nice, clean result that encapsulates the essence of the low complexity gradient condition. This result is powerful enough to imply the large deviation results for dense random graphs.

For each n, let $f_n : [0,1]^n \to \mathbb{R}$ be a twice differentiable function. For each n, i, and j, let $c_{n,i,j}$ be a uniform upper bound on the mixed partial derivative

$$\left| \frac{\partial^2 f_n}{\partial x_i \partial x_j} \right| .$$

Additionally, assume that a and b are constants such that for each n, $|f_n|$ is bounded by an and the absolute values of all the first-order partial derivatives of f_n are bounded by b. Let $\phi_{n,p}$ be defined according to the formula (8.1) applied to f_n, and assume that $\phi_p(t) := \lim_{n\to\infty} \phi_{p,n}(t)$ exists for each t.

Theorem 9.1. *Consider the setting introduced above. Let $C(f_n)$ be the complexity of the gradient of f_n, as defined in Section 8. Suppose that as $n \to \infty$, the following conditions hold:*

$$(9.1) \quad C(f_n) = o(1), \quad \sum_{i,j} c_{n,i,j} = o(n^2), \quad \sum_{i,j} c_{n,i,j}^2 = o(n), \quad \sum_i c_{n,i,i} = o(n) .$$

Take any $p \in (0,1)$ and let $Y = (Y_1, \dots, Y_n)$ be a vector of i.i.d. random variables with $\mathbb{P}(Y_i) = p = 1 - \mathbb{P}(Y_i = 0)$. Let ϕ_p be defined as above. Then for any t where

ϕ_p is continuous and finite,

$$\lim_{n\to\infty} \frac{\log \mathbb{P}(f_n(Y) \ge tn)}{n} = -\phi_p(t)\,.$$

Revisiting the examples of Section 8, recall that in three out of the four cases we verified or at least sketched that $C(f_n) \to 0$ as $n \to \infty$. It is easy to check in each of these cases that the remaining three conditions in (9.1) are valid. For instance, if

$$f_n(x) = \frac{1}{n} \sum_{1\le i<j\le n} x_i x_j\,,$$

then $c_{n,i,i} = 0$ and $c_{n,i,j} = 1/n$ for $i \ne j$. Therefore, (9.1) holds and so we can apply Theorem 9.1 for this function.

Theorem 9.1 is a straightforward corollary of the main theorem of [17], which is stated below. This theorem gives a quantitative error bound instead of a limiting result, which can be used to analyze situations where $p \to 0$ as $n \to \infty$.

Fix some $n \ge 1$. Let $\|f\|$ denote the supremum norm of $f : [0,1]^n \to \mathbb{R}$. Suppose that $f : [0,1]^n \to \mathbb{R}$ is twice continuously differentiable in $(0,1)^n$, such that f and all its first- and second-order derivatives extend continuously to the boundary. For each i and j, let

$$f_i := \frac{\partial f}{\partial x_i} \quad \text{and} \quad f_{ij} := \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

Define

$$a := \|f\|, \quad b_i := \|f_i\|, \quad \text{and} \quad c_{ij} := \|f_{ij}\|\,.$$

Given $\epsilon > 0$, let $\mathcal{D}(\epsilon)$ be a finite subset of $\mathbb{R}^n$ such that for all $x \in \{0,1\}^n$, there exists $d = (d_1, \ldots, d_n) \in \mathcal{D}(\epsilon)$ such that

$$\sum_{i=1}^n (f_i(x) - d_i)^2 \le n\epsilon^2.$$

The following result gives an error bound for the approximation (8.2) in terms of the quantities a, b_i, c_{ij} and the sizes of the sets $\mathcal{D}(\epsilon)$.

Theorem 9.2 ([17])**.** *Take f as above, $p \in (0,1)$, Y as in Theorem* 9.1*, and ϕ_p as in* (5.6)*. Then, for any $\delta > 0$, $\epsilon > 0$, and $t \in \mathbb{R}$,*

$$\log \mathbb{P}(f(Y) \ge tn) \le -\phi_p(t-\delta)n + \text{complexity term} + \text{smoothness term},$$

where with a, b, c_{ij}, $\mathcal{D}(\epsilon)$ defined above,

$$\text{complexity term} := \frac{1}{4}\Big(n\sum_{i=1}^n \beta_i^2\Big)^{1/2}\epsilon + 3n\epsilon + \log\left(\frac{4\phi_p(t)(\frac{1}{n}\sum_{i=1}^n b_i^2)^{1/2}}{\delta\epsilon}\right) + \log\left|\mathcal{D}\left(\frac{\delta\epsilon}{4\phi_p(t)}\right)\right|, \quad \textit{and}$$

$$\text{smoothness term} := 4\Bigg(\sum_{i=1}^n(\alpha\gamma_{ii} + \beta_i^2) + \frac{1}{4}\sum_{i,j=1}^n\big(\alpha\gamma_{ij}^2 + \beta_i\beta_j\gamma_{ij} + 4\beta_i\gamma_{ij}\big)\Bigg)^{1/2} + \frac{1}{4}\Big(\sum_{i=1}^n\beta_i^2\Big)^{1/2}\Big(\sum_{i=1}^n\gamma_{ii}^2\Big)^{1/2} + 3\sum_{i=1}^n\gamma_{ii} + \log 2\,,$$

where

$$\alpha := n\phi_p(t) + n|\log p| + n|\log(1-p)|\,,$$
$$\beta_i := \frac{2\phi_p(t)b_i}{\delta} + |\log p| + |\log(1-p)|\,, \text{ and}$$
$$\gamma_{ij} := \frac{2\phi_p(t)c_{ij}}{\delta} + \frac{6\phi_p(t)b_ib_j}{n\delta^2}\,.$$

Moreover,

$$\log \mathbb{P}(f(Y) \geq tn) \geq -\phi_p(t+\delta_0)n - \epsilon_0 n - \log 2\,,$$

where

$$\epsilon_0 := \frac{1}{\sqrt{n}}\left(4 + \left|\log\frac{p}{1-p}\right|\right)$$

and

$$\delta_0 := \frac{2}{n}\left(\sum_{i=1}^{n}(ac_{ii} + b_i^2)\right)^{1/2}.$$

When applying Theorem 9.2 to a given problem, one needs to first compute an error bound depending on some specific choice of ϵ and δ, and then optimize the resulting bound over all possible values of ϵ and δ.

Example 9.3 (Triangles in $G(n,p)$)**.** To get a flavor of the consequences of Theorem 9.2, let us look at what it says for triangles in $G(n,p)$. Recall the function $I_p : [0,1] \to \mathbb{R}$ defined in (5.1). For $x = (x_{ij})_{1\leq i<j\leq n} \in [0,1]^{n(n-1)/2}$, define

$$I_p(x) := \sum_{1\leq i<j\leq n} I_p(x_{ij})$$

and

$$T(x) := \frac{1}{6}\sum_{i,j,k=1}^{n} x_{ij}x_{jk}x_{ki}\,,$$

where $x_{ji} = x_{ij}$ and $x_{ii} = 0$. For $u > 1$ define

$$\psi_p(u) := \inf\{I_p(x) : T(x) \geq u\,\mathbb{E}(T_{n,p})\}\,,$$

where $T_{n,p}$ is the number of triangles in $G(n,p)$. The following result was proved in [17]. The complexity calculations of Example 8.3 are crucial for this result.

Theorem 9.4 ([17])**.** *For $u > 1$, n sufficiently large (depending only on u), and $n^{-1/6} \leq p \leq 1 - n^{-1}$,*

$$1 - \frac{c\log n}{n^{1/6}p^2} \leq \frac{\psi_p(u)}{-\log \mathbb{P}(T_{n,p} \geq u\,\mathbb{E}(T_{n,p}))} \leq 1 + \frac{C(\log n)^{33/29}}{n^{1/29}p^{42/29}}\,,$$

where c and C are constants that depend only on u.

Theorem 9.4 is a nonasymptotic result. To get an asymptotic statement, note that

$$\frac{\psi_p(u)}{-\log \mathbb{P}(T_{n,p} \geq u\,\mathbb{E}(T_{n,p}))} \to 1$$

if $n \to \infty$ and $p \to 0$ slower than $n^{-1/42}(\log n)^{11/14}$. Theorem 7.1 was proved in [41] by analyzing the asymptotic behavior of $\psi_p(u)$.

Example 9.5 (Three-term arithmetic progressions)**.** Fix $n \in \mathbb{N}$ and $p \in (0,1)$. Let A be a random subset of $\mathbb{Z}/n\mathbb{Z}$, constructed by keeping each element with probability p, and dropping with probability $1-p$, independently of each other. Recall the function $I_p : [0,1] \to \mathbb{R}$ defined in equation (5.1). For $x = (x_i)_{i \in \mathbb{Z}/n\mathbb{Z}}$, let

$$I_p(x) := \sum_{i \in \mathbb{Z}/n\mathbb{Z}} I_p(x_i)\,.$$

The following large deviation result about the number of three-term arithmetic progressions in A was proved in [17]. Again, the complexity calculations of Example 8.4 are used for proving this theorem.

Theorem 9.6 ([17])**.** *Let A be a random subset of $\mathbb{Z}/n\mathbb{Z}$, constructed as above. Let X be the number of pairs $(i,j) \in (\mathbb{Z}/n\mathbb{Z})^2$ such that $\{i, i+j, i+2j\} \subseteq A$. Let I_p be defined as above. Let*

$$\theta_p(u) := \inf\biggl\{I_p(x) : x \in [0,1]^{\mathbb{Z}/n\mathbb{Z}} \text{ such that } \sum_{i,j \in \mathbb{Z}/n\mathbb{Z}} x_i x_{i+j} x_{i+2j} \ge u\,\mathbb{E}(X)\biggr\}\,.$$

Suppose that $n^{-1/162} \le p \le 1 - n^{-1}$. Then for any $u > 1$,

$$1 - c\,n^{-1/6} p^{-6} \log n \le \frac{\theta_p(u)}{-\log \mathbb{P}(X \ge u\,\mathbb{E}(X))} \le 1 + C n^{-1/29} p^{-162/29} (\log n)^{33/29}\,,$$

where C and c are constants that depend only on u.

This theorem gives an approximation for the upper tail probabilities of the number of three-term arithmetic progressions in random subsets of $\mathbb{Z}/n\mathbb{Z}$ when p is either fixed or decays slower than $n^{-1/162}(\log n)^{33/162}$ as $n \to \infty$. Note that when $p = 1/2$, calculating these upper tail probabilities is the same problem as counting the number of subsets of $\mathbb{Z}/n\mathbb{Z}$ that contain more than a given number of three-term progressions.

The study of arithmetic progressions in subsets of integers has a long and storied history, most of which is concerned with questions of existence. An excellent survey of old and new results is available in [49]. Counting the number of sets with a given number of arithmetic progressions, or understanding the typical structure of sets that contain lots of progressions, are challenges of a different type, falling within the purview of large deviations theory. Recently a certain amount of interest has begun to grow around the resolution of such questions, quickly leading to the realization that the conventional theory of large deviations will not provide the answers. The most pertinent papers are the recent articles on probabilistic properties of the so-called *nonconventional averages* [13, 32–34]. For example, [13] gives a large deviation principle for sums of the type $\sum x_i x_{2i}$.

10. The naive mean field approximation

Theorem 9.2 is a consequence of a more general theorem about normalizing constants. In a nutshell, the theorem says that the so-called *naive mean field approximation* of statistical physics is valid when the low complexity gradient condition holds. The purpose of this section is to give the precise statement of this theorem and to present a sketch of its proof.

Take any twice differentiable $f : [0,1]^n$ and let a, b_i, c_{ij}, and $\mathcal{D}(\epsilon)$ be as in Theorem 9.2. For $x = (x_1, \dots, x_n) \in [0,1]^n$, let

$$I(x) := \sum_{i=1}^{n} (x_i \log x_i + (1 - x_i) \log(1 - x_i)) .$$

Let

$$F := \log \sum_{x \in [0,1]^n} e^{f(x)} .$$

One version of the naive mean field approximation is that

$$F \approx \sup_{x \in [0,1]^n} (f(x) - I(x)) .$$

The following theorem, proved in [17], gives a sufficient condition for the validity of this approximation. As far as we know, there is no other general sufficient condition for the validity of the naive mean field approximation.

Theorem 10.1 ([17])**.** *Let all notation be as above. Then for any $\epsilon > 0$,*

$$\begin{aligned} F \le \sup_{x \in [0,1]^n} (f(x) - I(x)) + \frac{1}{4}\Big(n \sum_{i=1}^{n} b_i^2\Big)^{1/2} \epsilon + 3n\epsilon + \log |\mathcal{D}(\epsilon)| \\ + 4\Bigg(\sum_{i=1}^{n} (a c_{ii} + b_i^2) + \frac{1}{4} \sum_{i,j=1}^{n} \big(a c_{ij}^2 + b_i b_j c_{ij} + 4 b_i c_{ij}\big)\Bigg)^{1/2} \\ + \frac{1}{4}\Big(\sum_{i=1}^{n} b_i^2\Big)^{1/2} \Big(\sum_{i=1}^{n} c_{ii}^2\Big)^{1/2} + 3 \sum_{i=1}^{n} c_{ii} + \log 2 , \end{aligned}$$

and

$$F \ge \sup_{x \in [0,1]^n} (f(x) - I(x)) - \frac{1}{2} \sum_{i=1}^{n} c_{ii} .$$

As in Theorem 9.2, the error bound needs to optimized over ϵ when applying Theorem 10.1. Theorem 9.2 is deduced from Theorem 10.1 by replacing f in Theorem 9.2 with a function g which is a smooth approximation of the function that equals 1 where $f(x) \ge tn$ and 0 where $f(x) < tn$.

To finish the discussion, let us now see a sketch of the proof of Theorem 10.1. This sketch is given with the intention of being helpful to someone who may try to solve the open problems listed in the next section. Let $X = (X_1, \dots, X_n)$ be a random vector that has probability density proportional to $e^{f(x)}$ on $\{0,1\}^n$ with respect to the counting measure. For each i, define a function $\hat{x}_i : [0,1]^n \to [0,1]$ as

$$\hat{x}_i(x) = \mathbb{E}(X_i \mid X_j = x_j,\ 1 \le j \le n,\ j \ne i) .$$

Let $\hat{x} : [0,1]^n \to [0,1]^n$ be the vector-valued function whose ith coordinate function is $\hat{x}_i$. Let $\hat{X} = \hat{x}(X)$. The first step in the proof is to show that if the smoothness term is small, then $f(X) \approx f(\hat{X})$ with high probability. To show this, define

$$h(x) := f(x) - f(\hat{x}(x)) .$$

Let $u_i(t,x) := f_i(tx + (1-t)\hat{x}(x))$, so that

$$h(x) = \int_0^1 \sum_{i=1}^{n} (x_i - \hat{x}_i(x)) u_i(t,x)\, dt .$$

Thus, if $D := f(X) - f(\hat{X})$, then

$$\mathbb{E}(D^2) = \int_0^1 \sum_{i=1}^n \mathbb{E}((X_i - \hat{X}_i)u_i(t, X)D)\, dt\,. \tag{10.1}$$

Let $X^{(i)}$ denote the random vector $(X_1, \ldots, X_{i-1}, 0, X_{i+1}, \ldots, X_n)$ and let $D_i := h(X^{(i)})$. Then note that $u_i(t, X^{(i)})D_i$ is a function of the random variables $(X_j)_{j\neq i}$ only. Therefore since $\hat{X}_i = \mathbb{E}(X_i \mid (X_j)_{j\neq i})$,

$$\mathbb{E}((X_i - \hat{X}_i)u_i(t, X^{(i)})D_i) = 0\,.$$

Thus,

$$\begin{aligned}&\mathbb{E}((X_i - \hat{X}_i)u_i(t, X)D)\\ &= \mathbb{E}((X_i - \hat{X}_i)u_i(t, X)D) - \mathbb{E}((X_i - \hat{X}_i)u_i(t, X^{(i)})D_i)\,.\end{aligned}$$

If the smoothness term is small, then $u_i(t, X) \approx u_i(t, X^{(i)})$ and $D \approx D_i$. Together with the identity (10.1), this shows that $f(X) \approx f(\hat{X})$ with high probability.

Next, define a function $g : [0,1]^n \times [0,1]^n \to \mathbb{R}$ as

$$g(x, y) := \sum_{i=1}^n (x_i \log y_i + (1 - x_i)\log(1 - y_i))\,.$$

By a similar argument as above, it is possible to show that if the smoothness term is small, then with high probability,

$$g(X, \hat{X}) \approx g(\hat{X}, \hat{X}) = I(\hat{X})\,.$$

Let A be the set of all x where $f(x) \approx f(\hat{x}(x))$ and $g(x, \hat{x}(x)) \approx I(\hat{x}(x))$. Since $X \in A$ with high probability,

$$\frac{\sum_{x\in A} e^{f(x)}}{\sum_{x\in\{0,1\}^n} e^{f(x)}} \approx 1\,.$$

Therefore

$$\begin{aligned}F &= \log \sum_{x\in\{0,1\}^n} e^{f(x)} \approx \log \sum_{x\in A} e^{f(x)}\\ &\approx \log \sum_{x\in A} e^{f(\hat{x}(x)) - I(\hat{x}(x)) + g(x,\hat{x}(x))}\,.\end{aligned}$$

Now let ϵ be a small positive number. Using the set $\mathcal{D}(\epsilon)$, it is easy to produce a set $\mathcal{D}'(\epsilon) \subseteq [0,1]^n$ such that $|\mathcal{D}(\epsilon)| = |\mathcal{D}'(\epsilon)|$, and for each x there exists $p \in \mathcal{D}'(\epsilon)$ such that $\hat{x}(x) \approx p$. For each $p \in \mathcal{D}'(\epsilon)$, let $\mathcal{P}(p)$ be the set of all $x \in \{0,1\}^n$ such that $\hat{x}(x) \approx p$. The crucial fact is that for any $p \in [0,1]^n$,

$$\sum_{x\in\{0,1\}^n} e^{g(x,p)} = 1\,.$$

Therefore,

$$\begin{aligned}
&\log \sum_{x\in A} e^{f(\hat{x}(x))-I(\hat{x}(x))+g(x,\hat{x}(x))} \\
&\quad \le \log \sum_{p\in\mathcal{D}'(\epsilon)} \sum_{x\in\mathcal{P}(p)} e^{f(\hat{x}(x))-I(\hat{x}(x))+g(x,\hat{x}(x))} \\
&\quad \approx \log \sum_{p\in\mathcal{D}'(\epsilon)} \sum_{x\in\mathcal{P}(p)} e^{f(p)-I(p)+g(x,p)} \\
&\quad \le \log \sum_{p\in\mathcal{D}'(\epsilon)} e^{f(p)-I(p)} \le \log|\mathcal{D}'(\epsilon)| + \sup_{p\in[0,1]^n} (f(p)-I(p)) .
\end{aligned}$$

This completes the sketch of the proof for the upper bound. The lower bound is much more straightforward. Take any $y \in [0,1]^n$. Let $Y = (Y_1, \dots, Y_n)$ be independent 0-1 random variables, with $\mathbb{P}(Y_i = 1) = y_i$. Then by Jensen's inequality,

$$\begin{aligned}
\sum_{x\in\{0,1\}^n} e^{f(x)} &= \sum_{x\in\{0,1\}^n} e^{f(x)-g(x,y)+g(x,y)} \\
&= \mathbb{E}(e^{f(Y)-g(Y,y)}) \\
&\ge \exp(\mathbb{E}(f(Y)-g(Y,y))) \\
&= \exp(\mathbb{E}(f(Y))-I(y)) .
\end{aligned}$$

Then, by using arguments similar to those employed for the upper bound, one can prove that if the error term in the lower bound is small, then $\mathbb{E}(f(Y)) \approx f(y)$. Since this is true for any y, this completes the sketch of the proof of the lower bound.

11. Open problems

There are many open problems about large deviations for random graphs since the subject is still in a stage of development. The following is a partial list of the most important questions.

1. Produce explicit nonconstant solutions of the variational problems arising from applications of Theorem 5.1. Currently, only the existence of nonconstant solutions is known in certain regimes, but explicit nonconstant solutions, or mathematically provable qualitative properties of nonconstant solutions, are unknown. This is important for understanding the conditional behavior of dense random graphs if some rare event takes place.
2. As a concrete instance of the above general question, analyze the behavior of the nonconstant solutions of (5.6).
3. Improve the main nonlinear large deviation result (Theorem 9.2), so that results like Theorem 7.1 can be proved when p tends to zero at an optimal rate.
4. As an example of the above, show that Theorem 7.1 holds when $p \to 0$ slower than $n^{-1/2}$.
5. Develop a sparse regularity lemma and a sparse graph limit theory that is powerful enough to prove results like Theorem 7.1 and Theorem 7.2. In fact, a reasonable test for the completeness of a sparse graph limit theory is whether it can lead to a solution of the large deviation question for sparse Erdős–Rényi random graphs. This is because analyzing the large deviation behavior of $G(n,p)$ for small p requires a full understanding of *all* possible sparse graph structures rather than focusing a small subset of graphs with nice properties.

6. Extend the large deviation results for three-term arithmetic progressions (Example 9.5) to longer progressions. The reader may recall that discrete Fourier analysis was used in the analysis of three-term progressions. The method does not seem to extend easily to longer progressions. It is possible that higher order Fourier analysis (Gowers norms) or the hypergraph regularity lemma may be needed for longer progressions.
7. Find explicit solutions to the variational problems coming from arithmetic progressions in the spirit of Theorems 5.6, 7.1, and 7.2 of this survey.
8. In addition to the above problems, there are many related open problems of similar flavor about exponential random graphs. In particular, nonconstant solutions of the variational problem (6.1) are generally not known, except for the solution of a related problem in [30]. Sparse exponential random graphs are still out of the reach of mathematical results, except for some progress in [17].

Acknowledgments

I thank Amir Dembo, Jafar Jafarov, Jun Yan, and Yufei Zhao for carefully reading the first draft and giving useful comments, and Rick Durrett for suggesting that I write this survey.

About the author

Sourav Chatterjee is professor of mathematics and statistics at Stanford University. He is interested in probability theory and its applications in mathematics, statistics, and physics.

References

[1] David J. Aldous, *Representations for partially exchangeable arrays of random variables*, J. Multivariate Anal. **11** (1981), no. 4, 581–598, DOI 10.1016/0047-259X(81)90099-3. MR637937

[2] David Aristoff and Charles Radin, *Emergent structures in large networks*, J. Appl. Probab. **50** (2013), no. 3, 883–888, DOI 10.1239/jap/1378401243. MR3102521

[3] David Aristoff and Lingjiong Zhu, *Asymptotic structure and singularities in constrained directed graphs*, Stochastic Process. Appl. **125** (2015), no. 11, 4154–4177, DOI 10.1016/j.spa.2015.06.004. MR3385599

[4] S. Bhamidi, G. Bresler and A. Sly, *Mixing time of exponential random graphs*, In 2008 IEEE 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS), 2008, 803–812.

[5] B. Bhattacharya, S. Ganguly, E. Lubetzky and Y. Zhao, *Upper tails and independence polynomials in random graphs*. arXiv preprint arXiv:1507.04074, 2015.

[6] Béla Bollobás and Oliver Riordan, *Metrics for sparse graphs*, Surveys in combinatorics 2009, London Math. Soc. Lecture Note Ser., vol. 365, Cambridge Univ. Press, Cambridge, 2009, pp. 211–287. MR2588543

[7] E. Bolthausen, F. Comets and A. Dembo, *Large deviations for random matrices and random graphs*, Private communication, 2003.

[8] Christian Borgs, Jennifer Chayes, László Lovász, Vera T. Sós, and Katalin Vesztergombi, *Counting graph homomorphisms*, Topics in discrete mathematics, Algorithms Combin., vol. 26, Springer, Berlin, 2006, pp. 315–371, DOI 10.1007/3-540-33700-8_18. MR2249277

[9] C. Borgs, J. T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi, *Convergent sequences of dense graphs. I. Subgraph frequencies, metric properties and testing*, Adv. Math. **219** (2008), no. 6, 1801–1851, DOI 10.1016/j.aim.2008.07.008. MR2455626

[10] C. Borgs, J. T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi, *Convergent sequences of dense graphs II. Multiway cuts and statistical physics*, Ann. of Math. (2) **176** (2012), no. 1, 151–219, DOI 10.4007/annals.2012.176.1.2. MR2925382

[11] C. Borgs, J. T. Chayes, H. Cohn and Y. Zhao, *An L^p theory of sparse graph convergence I: limits, sparse random graph models, and power law distributions*. arXiv preprint arXiv:1401.2906, 2014.
[12] C. Borgs, J. T. Chayes, H. Cohn and Y. Zhao, *An L^p theory of sparse graph convergence II: LD convergence, quotients, and right convergence*. arXiv preprint arXiv:1408.0744, 2014.
[13] Gioia Carinci, Jean-René Chazottes, Cristian Giardinà, and Frank Redig, *Nonconventional averages along arithmetic progressions and lattice spin systems*, Indag. Math. (N.S.) **23** (2012), no. 3, 589–602, DOI 10.1016/j.indag.2012.05.010. MR2948646
[14] Sourav Chatterjee, *Concentration inequalities with exchangeable pairs*, ProQuest LLC, Ann Arbor, MI, 2005. Thesis (Ph.D.)–Stanford University. MR2707160
[15] Sourav Chatterjee, *Stein's method for concentration inequalities*, Probab. Theory Related Fields **138** (2007), no. 1-2, 305–321, DOI 10.1007/s00440-006-0029-y. MR2288072
[16] Sourav Chatterjee, *The missing log in large deviations for triangle counts*, Random Structures Algorithms **40** (2012), no. 4, 437–451, DOI 10.1002/rsa.20381. MR2925306
[17] Chatterjee, S. and Dembo, A. (2014), Nonlinear large deviations. To appear in *Adv. Math.*
[18] Sourav Chatterjee and Partha S. Dey, *Applications of Stein's method for concentration inequalities*, Ann. Probab. **38** (2010), no. 6, 2443–2485, DOI 10.1214/10-AOP542. MR2683635
[19] Sourav Chatterjee and Persi Diaconis, *Estimating and understanding exponential random graph models*, Ann. Statist. **41** (2013), no. 5, 2428–2461, DOI 10.1214/13-AOS1155. MR3127871
[20] Sourav Chatterjee and S. R. S. Varadhan, *The large deviation principle for the Erdős–Rényi random graph*, European J. Combin. **32** (2011), no. 7, 1000–1017, DOI 10.1016/j.ejc.2011.03.014. MR2825532
[21] B. DeMarco and J. Kahn, *Upper tails for triangles*, Random Structures Algorithms **40** (2012), no. 4, 452–459, DOI 10.1002/rsa.20382. MR2925307
[22] B. Demarco and J. Kahn, *Tight upper tail bounds for cliques*, Random Structures Algorithms **41** (2012), no. 4, 469–487, DOI 10.1002/rsa.20440. MR2993131
[23] Amir Dembo and Ofer Zeitouni, *Large deviations techniques and applications*. (Corrected reprint of the second (1998) edition), Stochastic Modelling and Applied Probability, vol. 38, Springer-Verlag, Berlin, 2010. MR2571413
[24] Persi Diaconis and Svante Janson, *Graph limits and exchangeable random graphs*, Rend. Mat. Appl. (7) **28** (2008), no. 1, 33–61. MR2463439
[25] P. Erdős and A. Rényi, A., *On the evolution of random graphs*, Publ. Math. Inst. Hungar. Acad. Sci., **5** (1960), 17–61. MR0125031
[26] Alan Frieze and Ravi Kannan, *Quick approximation to matrices and applications*, Combinatorica **19** (1999), no. 2, 175–220, DOI 10.1007/s004930050052. MR1723039
[27] D. N. Hoover, *Row-column exchangeability and a generalized model for probability*, Exchangeability in probability and statistics (Rome, 1981), North-Holland, Amsterdam–New York, 1982, pp. 281–291. MR675982
[28] Svante Janson, Krzysztof Oleszkiewicz, and Andrzej Ruciński, *Upper tails for subgraph counts in random graphs*, Israel J. Math. **142** (2004), 61–92, DOI 10.1007/BF02771528. MR2085711
[29] Olav Kallenberg, *Probabilistic symmetries and invariance principles*, Probability and its Applications (New York), Springer, New York, 2005. MR2161313
[30] R. Kenyon, C. Radin, K. Ren and L. Sadun, *Multipodal Structure and Phase Transitions in Large Constrained Graphs*. arXiv preprint arXiv:1405.0599, (2014).
[31] R. Kenyon M. Yin, *On the asymptotics of constrained exponential random graphs*. arXiv preprint arXiv:1406.3662, 2014.
[32] Yuri Kifer, *Nonconventional limit theorems*, Probab. Theory Related Fields, **148** (2010), no. 1-2, 71–106, DOI 10.1007/s00440-009-0223-9. MR2653222
[33] Yuri Kifer and S. R. S. Varadhan, *Nonconventional limit theorems in discrete and continuous time via martingales*, Ann. Probab. **42** (2014), no. 2, 649–688, DOI 10.1214/12-AOP796. MR3178470
[34] Yuri Kifer and S. R. S. Varadhan, *Nonconventional large deviations theorems*, Probab. Theory Related Fields **158** (2014), no. 1-2, 197–224, DOI 10.1007/s00440-013-0481-4. MR3152784
[35] Jeong Han Kim and Van H. Vu, *Concentration of multivariate polynomials and its applications*, Combinatorica **20** (2000), no. 3, 417–434, DOI 10.1007/s004930070014. MR1774845

[36] J. H. Kim and V. H. Vu, *Divide and conquer martingales and the number of triangles in a random graph*, Random Structures Algorithms **24** (2004), no. 2, 166–174, DOI 10.1002/rsa.10113. MR2035874

[37] Rafał Latała, *Estimation of moments of sums of independent real random variables*, Ann. Probab. **25** (1997), no. 3, 1502–1513, DOI 10.1214/aop/1024404522. MR1457628

[38] László Lovász, *Large networks and graph limits*, American Mathematical Society Colloquium Publications, vol. 60, American Mathematical Society, Providence, RI, 2012. MR3012035

[39] László Lovász and Balázs Szegedy, *Limits of dense graph sequences*, J. Combin. Theory Ser. B **96** (2006), no. 6, 933–957, DOI 10.1016/j.jctb.2006.05.002. MR2274085

[40] E. Lubetzky and Y. Zhao, *On the variational problem for upper tails of triangle counts in sparse random graphs*. arXiv preprint arXiv:1402.6011, 2014.

[41] Eyal Lubetzky and Yufei Zhao, *On replica symmetry of large deviations in random graphs*, Random Structures Algorithms **47** (2015), no. 1, 109–146, DOI 10.1002/rsa.20536. MR3366814

[42] Charles Radin, Kui Ren, and Lorenzo Sadun, *The asymptotics of large constrained graphs*, J. Phys. A **47** (2014), no. 17, 175001, 20, DOI 10.1088/1751-8113/47/17/175001. MR3197554

[43] Charles Radin and Lorenzo Sadun, *Phase transitions in a complex network*, J. Phys. A **46** (2013), no. 30, 305002, 12, DOI 10.1088/1751-8113/46/30/305002. MR3083277

[44] Charles Radin and Lorenzo Sadun, *Singularities in the entropy of asymptotically large simple graphs*, J. Stat. Phys. **158** (2015), no. 4, 853–865, DOI 10.1007/s10955-014-1151-3. MR3311483

[45] Charles Radin and Mei Yin, *Phase transitions in exponential random graphs*, Ann. Appl. Probab. **23** (2013), no. 6, 2458–2471, DOI 10.1214/12-AAP907. MR3127941

[46] Cosma Rohilla Shalizi and Alessandro Rinaldo, *Consistency under sampling of exponential random graph models*, Ann. Statist. **41** (2013), no. 2, 508–535, DOI 10.1214/12-AOS1044. MR3099112

[47] Endre Szemerédi, *Regular partitions of graphs* (English, with French summary), Problèmes combinatoires et théorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976), Colloq. Internat. CNRS, vol. 260, CNRS, Paris, 1978, pp. 399–401. MR540024

[48] Michel Talagrand, *Concentration of measure and isoperimetric inequalities in product spaces*, Inst. Hautes Études Sci. Publ. Math. **81** (1995), 73–205. MR1361756

[49] Terence Tao and Van Vu, *Additive combinatorics*, Cambridge Studies in Advanced Mathematics, vol. 105, Cambridge University Press, Cambridge, 2006. MR2289012

[50] V. H. Vu, *Concentration of non-Lipschitz functions and applications*, Random Structures Algorithms **20** (2002), no. 3, 262–316, DOI 10.1002/rsa.10032. Probabilistic methods in combinatorial optimization. MR1900610

[51] Mei Yin, *Critical phenomena in exponential random graphs*, J. Stat. Phys. **153** (2013), no. 6, 1008–1021, DOI 10.1007/s10955-013-0874-x. MR3131507

[52] Y. Zhao, *On the lower tail variational problem for random graphs*. arXiv preprint arXiv:1502.00867, 2015.

Department of Statistics, Stanford University
E-mail address: souravc@stanford.edu