Large deviation principle of subgraph statistics of complex networks

Zezheng Song

August 21, 2018

Abstract

According to Central Limit Theorem(CLT), the average of the sum of a large amount of independent random variables, after an appropriate normalization, converges to a Gaussian distribution. However, in this report, we summarize some main results from recent papers on the large deviation problem, in which case the random variables are extremely far from the mean, and the probability of observing a large deviation is usually exponentially small. We also apply the large deviation principle to the case of complex networks to learn about the structure of the graphs.

1 Introduction

Nowadays, there are more and more studies focusing on complex networks. Complex networks arise in many situations in our daily life, such as in the setting of social networks, biological networks, etc. When studying networks, we use mathematical models from random graphs to describe them. Beginning from the pioneering Erdős–Rényi random graphs, people have developed much more models, such as exponential random graphs and stochastic block models. In order to better understand networks, it is extremely important to find the structure underlying them. This brings up the idea of graph partitioning, where we partition the nodes into several different groups, so nodes within the same group have a closer relationship with each other than with nodes outside the group. Under different models, people have developed many community-detection algorithms.

For practical reasons, it is sometimes impossible to have all the information of the whole networks, but we can use some subgraph statistics to infer the big picture. In this work, we focus on the large deviation theory for subgraph statistics, which could provide crucial information for the networks when a specific statistics lies at the upper tail of the distribution. In other words, we are interested in the structure of the graph conditioned on a large deviation.

This work is organized as following: In section 2, we discuss large deviation principle in different graph models and settings. In section 3, we will summarize some results from certain applications. In section 4, we conclude the main idea behind these works in the literature.

2 Large deviation under certain settings

2.1 Erdős–Rényi Uniform random graph

In uniform model, all graphs with n vertices and m edges are distributed uniformly. The result and work in this setting are mainly contributed by Dembo and Lubetzky[2].

Let $\mathcal{W}$ denote the space of all symmetric measurable functions, and let $\mathcal{W}_0$ be the space of graphons. Define a cut-norm on the graphon space $\mathcal{W}$,.

$$\delta_{\square}(W_1, W_2) := \inf_{\sigma} \|W_1 - W_2^{\sigma}\|_{\square}. \tag{1}$$

where σ denotes a measure-preserving map from $[0,1] \to [0,1]$, and W^{σ} denotes the graphon $W^{\sigma}(x,y) = W(\sigma(x), \sigma(y))$.

Upon taking the quotient *w.r.t.* the equivalence relation $W_1 \sim W_2$ iff $\delta_{\square}(W_1, W_2) = 0$, we get a compact metric space $(\tilde{W}_0, \delta_{\square})$. For any finite simple graph $H = (V(H), E(H))$ with $V(H) = \{1, ..., k\}$, its subgraph density in $W \in \mathcal{W}$, is

$$t_H(W) := \int_{[0,1]^k} \prod_{(i,j)\in E(H)} W(x_i, x_j) \mathrm{d}x_1 \cdots \mathrm{d}x_k. \tag{2}$$

Define

$$I_p(x) := \frac{x}{2} \log \frac{x}{p} + \frac{1-x}{2} \log \frac{1-x}{1-p} \quad \text{for } p \in (0,1) \text{ and } x \in [0,1]. \tag{3}$$

which could be extended to $\mathcal{W}_\prime$ by $I_p(W) := \int_{[0,1]^2} I_p(W(x,y))dxdy$ for $W \in \mathcal{W}_\prime$.

Define

$$W_0{}^{(p)} := \{W \in W_0 : \|W\|_1 = p\} \qquad \text{and} \qquad \tilde{W_0}^{(p)} = \{W \in \tilde{W}_0 : \|W\|_1 = p\} \tag{4}$$

where $\|W\| = \int |W(x,y)| dxdy$

Theorem 1 *Fix $0 < p < 1$ and let $m_n \in \mathbb{N}$ be such that $\frac{m_n}{\binom{n}{2}} \to p$ as $n \to \infty$. Let $G_n \sim \mathcal{G}(n, m_n)$. Then the sequence $\mathbb{P}(G_n \in \cdot)$ obeys the LDP in the space $(\tilde{W}_0, \delta_\square)$ with the good rate function J_p, where $J_p(W) = I_p(W)$ if $W \in \tilde{W}_0{}^{(p)}$ and ∞ otherwise. That is, for any closed set $F \in \tilde{W}_0$,*

$$\limsup_{n\to\infty} n^{-2} \log \mathbb{P}(G_n \in F) \le - \inf_{W\in F} J_p(W), \tag{5}$$

$$\liminf_{n\to\infty} n^{-2} \log \mathbb{P}(G_n \in U) \ge - \inf_{W\in U} J_p(W). \tag{6}$$

Define

$$\psi_H(p,r) := \inf \left\{ I_p(W) : W \in \tilde{\mathcal{W}}_0^{(p)}, t_H(W) \ge r \right\}. \tag{7}$$

Theorem 2 *Fixing a subgraph H and $0 < p < 1$, let $r_H \in (t_H(p), 1]$ denote the largest r for which the collection of graphons in (7) are nonempty.*

(a) For any $m_n \in \mathbb{N}$ such at $\frac{m_n}{\binom{n}{2}} \to p$ as $n \to \infty$ and right-continuity point $r \in [0, r_H]$ of $t \mapsto \psi_H(p,t)$, the random graph $G_n \sim \mathcal{G}(n, m_n)$ satisfies

$$\lim_{n\to\infty} n^{-2} \log \mathbb{P}\left(t_H(G_n) \ge r\right) = -\psi_H\left(p, r\right) \tag{8}$$

(b) For any (p,r) in part (a), and every $\epsilon > 0$ there is $C = C\left(H, \epsilon, p, r\right) > 0$ so that for all n large enough

$$\mathbb{P}\left(\delta_\square(G_n, F_*) \ge \epsilon | t_H\left(G_n\right) \ge r\right) \le e^{-Cn^2} \tag{9}$$

2.2 Erdős–Rényi graph

In this case,both the setting and the result are similar to the previous one. Therefore, we will only mention the main conclusion here. The result and work here are contributed by Chatterjee and Varadhan[1].

This section is dedicated to proving a large deviation principle for $G(n,p)$ random graph, where n denotes the number of nodes of the graph and p denotes the probability of any two nodes are connected. As in the previous subsection, we will use the same topological space$\widetilde{\mathcal{W}}$ and the cut distance $(\tilde{W}, \delta_\square)$. Finally, we define a rate function $I_p : [0,1] \to \mathbb{R}$ to be

$$I_p(x) := \frac{1}{2} u \log \frac{u}{p} + \frac{1}{2}(1-u) \log \frac{1-u}{1-p} \tag{10}$$

Similar to the previous subsection, we will extend the domain of I_p to W as

$$I_p(h) := \int_0^1 \int_0^1 I_p\left(h\left(x,y\right)\right) dxdy \tag{11}$$

Theorem 3 *For each fixed $p \in (0,1)$, the sequence $\widetilde{\mathbb{P}}_{n,p}$ obeys a large deviation principle in the space $\widetilde{\mathcal{W}}$ (equipped with the cut metric) with rate function I_p defined by (10). Explicitly, this means that for any closed set $\widetilde{F} \subset \widetilde{\mathcal{W}}$,*

$$\limsup_{n\to\infty} \frac{1}{n^2} \log \widetilde{\mathbb{P}}_{n,p}(\tilde{F}) \leq -\inf_{\tilde{h}\in\tilde{F}} I_p(\tilde{h}), \tag{12}$$

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

$$\liminf_{n\to\infty} \frac{1}{n^2} \log \widetilde{\mathbb{P}}_{n,p}(\tilde{U}) \geq -\inf_{\tilde{h}\in\tilde{U}} I_p(\tilde{h}). \tag{13}$$

This theorem estimates the probability of rare events for $G(n,p)$. However, the following theorem gives an answer to the question that what does the graph look like when some particular rare event has occurred.

Theorem 4 *Take any $p \in (0,1)$. Let $\widetilde{F}$ be a closed subset of $\widetilde{\mathcal{W}}$. Let $\widetilde{F}^*$ be the subset of $\widetilde{F}$ where I_p is minimized. Then $\widetilde{F}^*$ is non-empty and compact, and for each n, and each $\epsilon > 0$,*

$$\mathbb{P}\left(\delta_\square(G(n,p), \tilde{F}^*) \geq \epsilon | G(n,p) \in \tilde{F}\right) \leq e^{-C(\epsilon,\tilde{F})n^2} \tag{14}$$

where $C(\epsilon, \widetilde{F})$ is a positive constant depending only on ϵ and $\widetilde{F}$. In particular, if $\widetilde{F}^*$ contains only one element $\tilde{h}^*$, then the conditional distribution of $G(n,p)$ given $G(n,p) \in \widetilde{F}$ converges to the point mass at $\tilde{h}^*$ as $n \to \infty$.

3 Applications

3.1 Application to triangle counts

In this section, we provide a specific example from the work of Chatterjee and Varadhan[1] to see the power of the theory. Suppose $T_{n,p}$ be the number of triangles in $G(n,p)$. We want to compute the large deviation rate function for the upper tail of $T_{n,p}$ when p remains fixed and $n \to \infty$. In other words, given $p \in [0,1]$ and $\epsilon > 0$, we wish to evaluate the limit

$$\lim_{n\to\infty} \frac{1}{n^2} \log \mathbb{P}\left(T_{n,p} \geq (1+\epsilon)\mathbb{E}(T_{n,p})\right). \tag{15}$$

This problem has been open for a long time and was recently solved. The first result was made in and we now know that given $p \in (0,1)$, there exists $p^3 < t^{'} \leq t^{''} < \frac{1}{6}$ such that for all $t \in (\frac{p^3}{6}, t^{'}) \cup (t^{"}, \frac{1}{6})$,

$$\lim_{n\to\infty} \frac{1}{n^2} \log \mathbb{P}\left(T_{n,p} \geq tn^3\right) = -I_p((6t)^{1/3}). \tag{16}$$

However, this formula does not cover all values of (p,t), so in this section we want to improve it.

Surprisingly, the solution to the following variational problem could improve the above answer. Define

$$T(f) := \frac{1}{6}\int_0^1 \int_0^1 \int_0^1 f(x,y)f(y,z)f(z,x)\mathrm{d}x\mathrm{d}y\mathrm{d}z. \tag{17}$$

and for each $p \in [0,1]$ and $t \in [0, \frac{1}{6})$, let

$$\phi(p,t) := \inf\left\{I_p(f) : f \in \mathcal{W}, T(f) \geq t\right\}. \tag{18}$$

and for $t \geq \frac{1}{6}$, let $\phi(p,t) = \infty$. The following theorem gives a particular illustration for this specific example.

Theorem 5 *Let $G(n,p)$ be the Erdős–Rényi random graph on n vertices with edge probability p. Let $T_{n,p}$ denote the number of triangles in $G(n,p)$. Let ϕ be defined as above. Then for each $p \in (0,1)$ and each $t \geq 0$,*

$$\lim_{n\to\infty} \frac{1}{n^2} \log \mathbb{P}\left(T_{n,p} \geq tn^3\right) = -\phi(p,t) \tag{19}$$

4 Conclusion and Remark

From the previous sections, we can reach a conclusion to solve general large deviation problems. In general, we begin by defining a rate function on a certain topological space, and then we formulate a variational problem. The second step is to solve this variational problem, which is usually difficult and problem-dependent.

During the Moncrief Undergraduate Summer Internship of 2018, I mainly read some papers on the topic of complex networks and random graphs. In this summer report, I summarized some main results from these works. After this summer internship program, I will continue to study and research on this subject.

5 Acknowledgements

I am deeply thankful to my advisor, Kui Ren, for supervising my reading and research over the summer, discussing interesting problems with me and constantly providing me with tremendous help and support. His passion, guidance and encouragement has been invaluable to me. I am grateful to have him as both a mentor and a friend. This work would be impossible without his support for me throughout my undergraduate career.

References

[1] S. Chatterjee and S. R. S. Varadhan. *The large deviation principle for the Erdos-Renyi random graph*[J]. European J. Combin., 32(7):1000–1017, 2011.

[2] A. Dembo and E. Lubetzky.*A large deviation principle for the Erdos-Renyi uniform random graph*[J].