S&DS 684: Statistical Inference on Graphs Fall 2018

Lecture 1: Introduction & Max Clique in Erdős-Rényi graphs

Lecturer: Yihong Wu *Scribe: Brandon Chow, August 29, 2018 [Ed. Sep 8]*

1.1 Introduction

1.1.1 Basic Definitions

A **graph** $G = (V, E)$ consists of

- A **vertex set** V. With loss of generality (WLOG), we shall assume $V = [n] \equiv \{0, 1, \ldots, n\}$ for some positive integer n.
- An **edge set** $E \subset \binom{V}{2}$. Each element of E is an edge $e = (i, j)$ (unordered pair). We say i and j are connected and write $i \sim j$ if $(i, j) \in E$.

For the most part, we will be focusing on graphs that are *undirected* (i.e., edges do not have orientation) and *simple* (i.e., no multi-edges or self-loops).

Alternatively, one can also represent a graph as an **adjacency matrix** $A = (A_{ij})_{i,j\in[n]}$, which is an $n \times n$ symmetric binary matrix with zero diagonal. In particular, for a simple and undirected graph $G = (V, E)$, the entries A_{ij} are defined as:

$$A_{ij} = \mathbb{1}\{i \sim j\} = \begin{cases} 1 & (i, j) \in E \\ 0 & \text{o.w.} \end{cases}.$$

Some basic concepts of graphs are defined as follows:

- The **neighborhood** of a given vertex $v \in V$ is defined as $N(v) = \{u \in V : u \sim v\}$, i.e., it is the set of vertices (neighbors) that are connected with v.
- The **degree** of v is defined as $d_v = |N(v)|$, i.e., the number of neighbors of v.
- **Induced subgraph**: For any $S \subset V$, the subgraph induced by S is defined as the graph $G[S] = (S, E_S)$, where $E_S \triangleq \{(u, v) \in E : u, v \in S\}$.
- A **clique** is a **complete** subgraph. A graph is complete iff every pair of vertices in the graph are connected.

1.1.2 Sample topics

The goal of statistical inference is to using data to make informed decisions (hypotheses testing, estimation, etc). The usual framework of statistical inference is the following:

$$\underbrace{\theta \in \Theta}_{\text{parameter}} \mapsto \underbrace{X}_{\text{data}} \mapsto \underbrace{\hat{\theta}}_{\text{estimate}}.$$

The theoretical objectives of this class are two-fold:

1. Understand and characterize the fundamental (statistical) limits: What is possible/impossible information-theoretically?
2. Can statistical limits be attained computationally efficiently, e.g., in polynomial time? If yes, how? If not, why?

In this course,

- Data = graphs;
- Parameter = hidden (latent, or planted) structure;
- We will focus on large-graph limit (number of vertices $\to \infty$).

As a preview, we briefly describe two models that we will study below: the **Planted Clique Model** and the **Stochastic Block Model**.

The Planted Clique Model Let V be a vertex set and $n = |V|$, and let $k \leq n$ be a given positive integer. The edge set E in a graph $G = (V, E)$ is generated in the following manner:

1. A set S of k vertices is selected out of n vertices to form a clique (all possible edges between them are added to E).
2. Remaining edges are added independently with probability $\frac{1}{2}$.

Given the resulting graph $G = (V, E)$, the goal is to find the planted (hidden) clique S.

To start, notice that this set up follows a classical statistical framework: a sample (here, the graph G) is generated from a distribution (i.e., the random process described above), and we want to estimate a parameter of that distribution (here, the set S) via the sample (here, G).

A decision-theoretic setting is to consider the minimax framework for the worst-case analysis, in which the goal is to find an estimator $\widehat{S} = \widehat{S}(G)$ that correctly recover S with probability close to 1, regardless of the true set S used to generate the graph G. In other words,

$$\min_{S \in \binom{[n]}{k}} \P_S\Big[\widehat{S}(G) = S\Big] \approx 1,$$

where $\P_S$ denotes the law of G conditioned on the location of the planted clique S. Alternatively, one can consider the more relaxed Bayesian setting, assuming S is drawn uniformly at random. Equivalently, this amounts to finding an $\widehat{S}$ that preforms well on average:

$$\mathbb{E}_{S \sim \mathrm{Unif}\left(\binom{[n]}{k}\right)} \P_S\Big[\widehat{S}(G) = S\Big] \approx 1.$$

Remark 1.1. For problems with symmetry, these two formulations are often equivalent, in the sense that

$$\sup_{\widehat{S}} \min_{S \in \binom{[n]}{k}} \P_S\Big[\widehat{S}(G) = S\Big] = \sup_{\widehat{S}} \mathbb{E}_{S \sim \mathrm{Unif}\left(\binom{[n]}{k}\right)} \P_S\Big[\widehat{S}(G) = S\Big].$$

This follows from the permutation invariance of the model, which implies the least favorable prior is uniform.

The Stochastic Block Model (SBM) Given a vertex set V, suppose V can be partitioned into two "communities" of equal size. Community membership is represented by a vector

$$\sigma = (\sigma_1, \ldots, \sigma_n) \in \{\pm 1\}^n,$$

where $\sigma_i = \sigma_j$ means that i and j belong to the same community, and $\sum_{i=1}^n \sigma_i = 0$ because the size of the two communities are equal. An edge between two vertices $i, j \in V$ is added to E according to the following probabilities:

$$\P\big[(i,j) \in E\big] = \begin{cases} p & \sigma_i = \sigma_j \\ q & \sigma_i \neq \sigma_j \end{cases},$$

where $0 \leq p, q \leq 1$ (note that p, q need not sum to 1). Thus, in this model, in-group ties and out-group ties have a different probability of forming. There are also several different statistical inference tasks associated with this problem that SBMs address. For example, if p and q are known, then our goal could be to estimate the parameter σ. Or, if p and q are unknown, then we may be interested in jointly estimating $p, q,$ *and* σ.

1.2 Asymptotic Behavior of Max Clique in $G(n, \frac{1}{2})$

We start with the ensemble of the Erdős-Rényi graph: $G \sim G(n,p)$ is a graph on n vertices where each pair of vertices is connected independently with probability p. Next, as a warmup, we will focus on the behavior of the maximum size of a clique in $G(n, \frac{1}{2})$.

In particular, let $G_n \sim G(n, \frac{1}{2})$. Define its clique number $\omega(G_n) \triangleq$ size of the max clique in G_n. We will show that $\omega(G_n) \approx 2\log_2 n$ for large n: for any $\epsilon > 0$, with high probability (whp),

$$\omega(G_n) \leq (2+\epsilon)\log_2 n, \tag{1.1}$$

$$\omega(G_n) \geq (2-\epsilon)\log_2 n. \tag{1.2}$$

In other words, $\frac{\omega(G_n)}{\log_2 n} \to 2$ in probability.

1.2.1 Proof of (1.1): First moment method

Let any $\epsilon > 0$ be given. We will show that $\P\big[\omega(G_n) \geq (2+\epsilon)\log_2 n\big] \to 0$.

To start, consider any positive integer k, as well as any $S \subset [n]$ where $|S| = k$. Notice that there are $\binom{k}{2}$ possible edges that can form between the k vertices in S, meaning that:

$$P(G_n[S] \text{ is a } k\text{-clique}) = 2^{-\binom{k}{2}}.$$

And, there are $\binom{n}{k}$ different sets of k vertices in a graph with n vertices. So, by the union bound,

$$\P(\exists S \subset [n] : G_n[S] \text{ is a } k\text{-clique}) \leq \binom{n}{k} 2^{-\binom{k}{2}}.$$

Now, let $k_0 = (2+\epsilon)\log_2 n$. Again by the union bound, we have that:

$$\begin{aligned}
\P(\omega(G) \geq k_0) &\leq \sum_{k=k_0}^{n} \binom{n}{k} 2^{-\binom{k}{2}} \\
&\overset{(a)}{\leq} \sum_{k=k_0}^{n} \left(n 2^{-\frac{(k_0-1)}{2}}\right)^k \\
&\leq \sum_{k=k_0}^{\infty} \left(n 2^{-\frac{(k_0-1)}{2}}\right)^k \overset{(b)}{\leq} 2(n 2^{-\frac{(k_0-1)}{2}})^{k_0},
\end{aligned}$$

where (a) follows from $\binom{n}{k} \leq n^k$ and $k_0 \leq k$, (b) follows from $n 2^{-\frac{k_0-1}{2}} = \sqrt{2} n^{-\epsilon/2} < 1/2$ for sufficiently large n.

1.2.2 Proof of (1.2): Second moment method

We will now show that $\lim_{n\to\infty} \P[\omega(G_n) \geq k] \to 1$, where $k \triangleq (2-\epsilon)\log_2 n$. Define:

$$T_n \triangleq \#\text{ of cliques of size } k \text{ in } G_n = \sum_{|S|=k} \mathbb{1}\{G_n[S] \text{ is a } k \text{ clique}\}. \tag{1.3}$$

Note that if a graph contains at least one clique of size k, then the max clique must be of size $\geq k$, implying that $\P[\omega(G_n) \geq k] \geq \P[T_n > 0]$. So, to show that $\P[\omega(G_n) \geq k] \to 1$ as $n \to \infty$, it suffices to show instead that $\P[T_n > 0] \to 1$.

Intuition

But, before trying to prove that $\P[T_n > 0] \to 1$, let's first build some intuition. What we computed in the union bound is in fact computing the *first moment* of $\mathbb{E}[T_n]$. By linearity of expectation, we have

$$\mathbb{E}[T_n] = \binom{n}{k} 2^{-\binom{k}{2}}. \tag{1.4}$$

Clearly, when $k = (2+\epsilon)\log_2 n$, $\mathbb{E}[T_n] \ll 0$, which implies that $\mathbb{P}[T_n > 0] \ll 0$ since T_n is integer-valued. As T_n is a positive random variable, it tempting to think that a sufficient condition for $\P[T_n > 0] \gg 0$ is $\mathbb{E}[T_n] \gg 0$. However, this direction is generally false: a counterexample would be a distribution that places almost *all* of its probability mass at zero, and the remaining very *small* amount of probability mass at, say, 10^{100}. Indeed, while the expected value of a random variable with this distribution would be very large, the probability that this random variable is non-zero would still be very small.

So, to show that $\P[T_n > 0]$ is large, it won't be enough to show that $\mathbb{E}[T_n]$ is large. What to do? Well, one way to characterize the distribution in the counterexample above is that it has very *high variance*. If we can show that the variance of T_n is not so large, then that would essentially show that T_n's distribution does not assign low probability to extremely high valued integers, essentially ruling out counterexamples like the one previously entertained. Will this be enough?

Second Moment Method

As it turns out, this approach works and is called the ***Second Moment Method***. Briefly, suppose X_n is a non-negative, integer-valued random variable. In this approach, one shows that $\P[X_n > 0] \to 1$ by showing that:

$$\mathrm{Var}[X_n] = o(\mathbb{E}^2[X_n]),$$

where Var stands for variance. Since we are going to apply the Second Moment Method to show that $\P[T_n > 0] \to 1$, let's first take a small detour to prove it works for the general random variable X_n describe above. And, the first step in doing so will be to prove the Paley-Zygmund inequality.

Lemma 1.1 (Paley-Zygmund Inequality)**.** *Let $X \geq 0$ be a random variable with $0 < \mathbb{E}[X^2] < \infty$. Then for any $0 \leq c \leq 1$,*

$$\P(X > c\mathbb{E}[X]) \geq (1-c)^2 \frac{\mathbb{E}^2[X]}{\mathbb{E}[X^2]} = (1-c)^2 \frac{\mathbb{E}^2[X]}{\mathbb{E}^2[X] + \mathrm{Var}[X]}. \tag{1.5}$$

Proof. First, note that:

$$\mathbb{E}[X] = \mathbb{E}[X\mathbb{1}\{X \leq c\mathbb{E}[X]\}] + \mathbb{E}[X\mathbb{1}\{X > c\mathbb{E}[X]\}] \leq c\mathbb{E}[X] + \mathbb{E}[X\mathbb{1}\{X > c\mathbb{E}[X]\}],$$

meaning that $(1-c)\mathbb{E}[X] \leq \mathbb{E}[X\mathbb{1}\{X > c\mathbb{E}[X]\}]$. Next, note that by Cauchy Swartz:

$$\mathbb{E}[X\mathbb{1}\{X > c\mathbb{E}[X]\}] \leq \sqrt{\mathbb{E}[X^2]}\sqrt{\P(X > c\mathbb{E}[X])}.$$

Thus:

$$(1-c)^2\mathbb{E}^2[X] \leq \mathbb{E}[X^2]\P(X > c\mathbb{E}[X]),$$

which implies the desired inequality. □

To show that the Second Moment Method works, notice that choosing $c = 0$ in the Paley Zygmund inequality gives us

$$\P(X_n > 0) \geq \frac{\mathbb{E}^2[X_n]}{\mathbb{E}^2[X_n] + \mathrm{Var}[X_n]} = \frac{1}{1 + \frac{\mathrm{Var}[X_n]}{\mathbb{E}^2[X_n]}},$$

so if $\mathrm{Var}[X_n] = o(\mathbb{E}^2[X_n])$, then $\P(X_n > 0) \to 1$, as desired.

Applying the Second Moment Method

We now return to our original goal of showing that $\P[T_n > 0] \to 1$, which we shall prove via the Second Moment Method. In particular, we need to show that $\mathrm{Var}[T_n] = o(\mathbb{E}^2[T_n])$. To start, notice

that:

$$\begin{aligned}
\mathrm{Var}[T_n] &= \mathrm{Var}\left[\sum_{|S|=k} \mathbb{1}\left\{G_n[S] \text{ is a } k \text{ clique}\right\}\right] \\
&= \sum_{\substack{S,S' \\ |S|=|S'|=k}} \mathrm{Cov}\Big[\mathbb{1}\left\{G_n[S] \text{ is a } k \text{ clique}\right\}, \mathbb{1}\left\{G_n[S'] \text{ is a } k \text{ clique}\right\}\Big] \\
&\overset{(a)}{=} \sum_{\substack{|S\cap S'|\geq 2 \\ |S|=|S'|=k}} \mathrm{Cov}\Big[\mathbb{1}\left\{G_n[S] \text{ is a } k \text{ clique}\right\}, \mathbb{1}\left\{G_n[S'] \text{ is a } k \text{ clique}\right\}\Big] \\
&\leq \sum_{\substack{|S\cap S'|\geq 2 \\ |S|=|S'|=k}} \P\Big[\text{ both } G_n[S] \text{ and } G_n[S'] \text{ are } k \text{ cliques}\Big],
\end{aligned}$$

where (a) follows from the fact that, for any two vertex sets S and S'. If $|S \cap S'| \leq 1$ (at most one node shared between S and S'), then the set of edges formed among nodes in S are *disjoint* from the set of edges formed among nodes in S'. Thus, by independence, the covariance is zero.

Now, for any given pair of sets S,S', let $\ell = |S \cap S'|$. In order for S and S' to *both* be k-cliques, there are a total of $2\binom{k}{2} - \binom{l}{2}$ possible edges that must be formed (think: inclusion-exclusion principle), so we have

$$\mathrm{Var}[T_n] = \sum_{\ell=2}^{k} \left|\left\{(S,S') : |S| = |S'| = k, |S \cap S'| = \ell\right\}\right| \cdot 2^{-2\binom{k}{2}+\binom{\ell}{2}} \tag{1.6}$$

$$= \sum_{\ell=2}^{k} \binom{n}{k}\binom{k}{\ell}\binom{n-k}{k-\ell} \cdot 2^{-2\binom{k}{2}+\binom{\ell}{2}}, \tag{1.7}$$

where the last step follows from the following reasoning: there are $\binom{n}{k}$ ways of picking a set S of k vertices from a graph on n vertices. And, for each such set S, there are exactly $\binom{k}{\ell}$ ways to pick ℓ nodes from S that will also be part of another set S'. Once S and the nodes of S that will be shared with S' have been determined, it remains to pick from S^c the remaining $k - l$ nodes of S', and there are exactly $\binom{n-k}{k-\ell}$ ways of doing that.

At this point, one can analyze the above sum by brute force, focusing on the exponent of each term. Next we present a more "statistician's approach". Note that the counting step in (1.6) is precisely how hypergeometric distribution (sampling without replacement) arises. Indeed, if we have an urn of n balls among which k balls are red, let H denote the number of red balls if we draw k balls from the urn uniformly at random without replacements. Then $H \sim \text{Hypergeometric}(n,k,k)$. Thus, we can express the same quantity in terms of H as follows:

$$\begin{aligned}
\frac{\mathrm{Var}[T_n]}{\mathbb{E}^2[T_n]} &= \sum_{\ell=2}^{k} \frac{\binom{k}{\ell}\binom{n-k}{k-\ell}}{\binom{n}{k}} \cdot 2^{\binom{\ell}{2}} \leq \sum_{\ell=2}^{k} \frac{\binom{k}{\ell}\binom{n-k}{k-\ell}}{\binom{n}{k}} \cdot 2^{\ell k/2} \\
&= \mathbb{E}[2^{kH/2}\mathbb{1}\{H \geq 2\}] \leq \mathbb{E}[2^{kH/2}] - \P[H=0].
\end{aligned} \tag{1.8}$$

Next we will show that both $\P[H=0] \to 1$ and $\mathbb{E}[2^{kH/2}] \to 1$. Indeed,

$$\P[H=0] = \frac{\binom{n-k}{k}}{\binom{n}{k}} = \left(1-\frac{k}{n}\right)\left(1-\frac{k}{n-1}\right)\cdots\left(1-\frac{k}{n-k+1}\right) \to 1,$$

since $k = (2-\epsilon)\log_2 n = o(\sqrt{n})$.

To bound the generating function, we use the comparison between sampling with replacements (binomial) and sampling without replacements (hypergeometric). The following result of Hoeffding (proved in the homework) will be useful in several places in this course:

Lemma 1.2 (Hoeffding's lemma)**.** Binom$(k, \frac{k}{n})$ *dominates* Hypergeometric(n, k, k) *in the order of convex functions. In other words, if* $B \sim$ *Binomial*$(k, \frac{k}{n})$*, then* $\mathbb{E}[f(H)] \le \mathbb{E}[f(B)]$ *for all convex functions* f*.*

Using this lemma, we have

$$\mathbb{E}[2^{kH/2}] \le \mathbb{E}[2^{kB/2}] = \left(1 + \frac{k}{n}\left(2^{\frac{k}{2}} - 1\right)\right)^k \le \exp\left(\frac{k^2}{n}\left(2^{\frac{k}{2}} - 1\right)\right) \to 1.$$

since $k^2 2^{\frac{k}{2}} \ll n$ by the assumption that $k = (2-\epsilon)\log_2 n$.

To summarize, we have shown that $\dfrac{\mathrm{Var}[T_n]}{\mathbb{E}^2[T_n]} \to 0$. By Paley-Zygmund (Lemma 1.1), it follows that $\P[T_n > 0] \to 1$, i.e., $\P[\omega(G_n) \ge (2-\epsilon)\log_2 n] \to 1$, so we've proven the desiderata.

Remark 1.2. Note that in computing the second moment, (1.8) can be equivalently written as

$$\frac{\mathrm{Var}[T_n]}{\mathbb{E}^2[T_n]} = \mathbb{E}[2^{k|S\cap S'|/2}\mathbb{1}\left\{|S \cap S'| \ge 2\right\}],$$

where S and S' are independent random k-sets drawn uniformly. This is something we will frequently encounter in computing the second moment, which typically involves *two independent copies* of the same randomness and their *overlap* $|S \cap S'|$.

Remark 1.3. As a small aside, we can further show that not only there exists a clique of size $k = (2-\epsilon)\log_2 n$, there are an *abundance* of them. Indeed, by (1.4) and using $\binom{n}{k} \ge (\frac{n}{k})^k$, we have

$$\mathbb{E}[T_n] = \binom{n}{k} 2^{-\binom{k}{2}} \ge \left(\frac{n}{k}\right)^k 2^{-\binom{k}{2}} = n^{\Omega(\log n)} \to \infty.$$

By Lemma 1.1, we have $T_n > o(\mathbb{E}[T_n])$ with probability $1 - o(1)$. This shows that there exists superpolynomially many cliques of size $(2-\epsilon)\log_2 n$. Unfortunately, the best polynomial-time algorithm can only guarantee to find a clique of size $(1-\epsilon)\log_2 n$ with high probability. We will discuss this next time.

References