随机图: 概率与图论的邂逅

胡洁洋

1 问题的引入: 从社交网络到随机图

在信息化高度发达的今天, 每个人都生活在一个巨大的社交网络中. 我们的每一次点赞、分享, 甚至一次简单的聊天, 都在无形中连接着这个网络的节点. 试想在一个大型社交平台上, 随机选择两个人, 他们之间是否是好友? 这样的关系网络又是否表现出整体的规律性? 我们又该如何量化这个复杂网络的行为?

社交网络的规模通常非常庞大——拥有数亿甚至数十亿个用户. 每个用户可以被看作是一个网络中的"节点", 每一对好友关系则构成了一条"边". 然而, 用户之间是否相互连接往往具有随机性: 有人拥有数千好友, 有人只有寥寥数人. 如此复杂而动态的系统背后隐藏了什么样的结构特性? 如果样本足够多, 我们确实可以将一对用户是否为好友视为随机事件, 运用我们学过的概率论知识构建模型.

为了简单起见, 我们可以假设社交网络的每一对用户有概率 p 成为好友, 且这个概率对于每一对用户都是相同、独立的, 这样我们就抽象出了 **Erdős–Rényi 随机图模型** *(Erdős–Rényi random graph model)* .

定义 1.1 (Erdős–Rényi 随机图模型)**.** *给定 $n \in \mathbb{N}^*$ 和 $p \in [0,1]$, 设 $V := [n]$, 随机图 $G = (V, E)$ 由以下方式生成: 对 G 的任意两个不同顶点 x, y, 边 $\{x, y\}$ 有 p 的概率在 E 中, 且与其他边的生成独立. 称此图模型为 n 个顶点, 密度为 p 的 Erdős–Rényi 随机图模型. 我们记 $G \sim \mathbb{G}_{n,p}$.*

在这个模型的假设下, 我们就可以开始着手研究我们关心的话题, 比如: 网络中是否可能存在"孤岛", 即完全没有连接的用户群体; 网络是否完全连通, 即所有用户都能通过某些好友路径间接联系; 若 p 足够小, 网络可能会破碎成多个孤立的子图, 甚至有孤立点出现; 若 p 足够大, 则可能形成一个"巨型连通分支". 这些现象之间有怎样的临界点? 比如, 当 p 较小的时候, 图中可能几乎没有边; 当 p 接近于 1 时, 图中大多数的节点会被连接, 很大可能所有人都处在同一个网络之中, 而没有"信息孤岛".

本文仅涉及最为浅显的数学分析和概率论知识, 尝试运用朴素的概率工具研究 ER 模型最为基础的话题, 揭示随机图最为基本的数学规律. 在开始之前, 让我们先回顾一下我们拥有的概率论知识, 发展一套趁手的工具, 供我们随时取用.

2 随机图问题的工具箱

随机变量的矩能够提供信息, 比如我们可以利用矩得到尾部概率的估计. 所谓尾部概率, 就是 $\mathbb{P}(X \geq x)$(通常 x 远大于 X 的期望) 和 $\mathbb{P}(X \leq x)$(通常 x 远小于 X 的期望), 我们将前者称为**上尾 (或右尾) 概率** *(upper/right tail probability)* , 将后者称为**下尾 (或左尾) 概率** *(lower/left tail probability)* .

一阶矩和二阶矩能对尾概率提供上下界估计, 由此衍生出一阶矩方法和二阶矩方法, 本节尝试介绍之.

2.1 一阶矩方法

所谓**一阶矩方法** *(first moment method)* 看起来非常平凡, 但我们会发现其在随机图的估计中非常有效. 一阶矩方法的强大之处源于数学期望非平凡的性质: 定义于联合概率空间上的随机变量 $X_1, X_2, \cdots, X_n$ 的一阶矩存在, 则

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i],$$

此式成立不需要其他任何条件, 特别地, 就算它们之间两两不独立, 此式仍然成立, 所以我们可以通过将随机变量 X 分解为若干小的 (示性) 随机变量, 从而方便地求出 $\mathbb{E}[X]$.

定理 2.1 (一阶矩方法)**.** *对于非负整值随机变量 X, 有*

$$\mathbb{P}(X > 0) \leq \mathbb{E}[X].$$

证明. 由 Markov 不等式,

$$\mathbb{P}(X > 0) = \mathbb{P}(X \geq 1) \leq \frac{\mathbb{E}[X]}{1} = \mathbb{E}[X].$$

□

此定理的使用范式为: 如果我们想证明随着 n 的增大, 某个“坏”事件以接近 1 的概率不出现, 我们就可以令随机变量 X_n 为其出现的次数 (可以显式地写为一些示性随机变量之和), 如果我们证明了 $\mathbb{E}[X_n] \to 0$, 当 $n \to \infty$, 那么 $\mathbb{P}(X_n > 0) \to 0$.

将定理 2.1写为事件的形式, 我们有

推论 2.2. *令 $B_n = A_{n,1} \cup A_{n,2} \cup \cdots \cup A_{n,m_n}$, 其中 $A_{n,1}, A_{n,2}, \cdots, A_{n,m_n}$ 是事件. 若记*

$$\mu_n \coloneqq \sum_{i=1}^{m_n} \mathbb{P}(A_{n,i}),$$

则 $\mathbb{P}(B_n) \leq \mu_n$. 特别地, 若 $\mu_n \to 0$ 当 $n \to \infty$, 则 $\mathbb{P}(B_n) \to 0$ 当 $n \to \infty$.

证明. 令 $X = \sum_{i=1}^{m_n} \mathbf{1}_{A_{n,i}}$, 由定理 2.1即得证. □

2.2 二阶矩方法

一阶矩方法给出了非负整值随机变量为正概率的上界估计, 那么**二阶矩方法** *(second moment method)* 就可以给出此概率的下界估计.

一个非常朴素的方法是通过 Chebyshev 不等式直接得到下界:

$$\mathbb{P}(X>0)=1-\mathbb{P}(X=0)\geq 1-\mathbb{P}(|X-\mathbb{E}[X]|\geq \mathbb{E}[X])\geq 1-\frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2}. \tag{1}$$

这个不等式实际上在处理大部分情况已经完全足够用, 但通过下面的 **Paley-Zygmund 不等式**, 我们还能给出更紧的版本.

定理 2.3 (Paley-Zygmund)**.** *对于非负随机变量 X 和实数 $\theta\in(0,1)$, 有*

$$\mathbb{P}(X\geq \theta\mathbb{E}[X])\geq (1-\theta)^2\frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}.$$

证明. 由 Cauchy-Schwarz 不等式,

$$\begin{aligned}\mathbb{E}[X]&=\mathbb{E}[X;X<\theta\mathbb{E}[X]]+\mathbb{E}[X;X\geq\theta\mathbb{E}[X]]\\&\leq \theta\mathbb{E}[X]+\mathbb{E}[X\mathbf{1}_{X\geq\theta\mathbb{E}[X]}]\\&\leq \theta\mathbb{E}[X]+\sqrt{\mathbb{E}[X^2]\mathbb{E}[\mathbf{1}_{X\geq\theta\mathbb{E}[X]}]}\\&=\theta\mathbb{E}[X]+\sqrt{\mathbb{E}[X^2]\mathbb{P}(X\geq\theta\mathbb{E}[X])},\end{aligned}$$

整理即得. □

由此我们马上有:

定理 2.4 (二阶矩方法)**.** *对非负随机变量 X, 有*

$$\mathbb{P}(X>0)\geq\frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}. \tag{2}$$

证明. 在定理 2.3中, 令 $\theta\downarrow 0$ 即得证. □

简单的计算表明, (2) 比 (1) 紧一些, 我们将定理 2.4称为二阶矩方法, 由于定理 2.4是定理 2.3的简单推论, 有时我们会将定理 2.3称为一般的二阶矩方法.

有了二阶矩方法, 我们只需计算 $\mathbb{E}[X]^2$ 和 $\mathbb{E}[X^2]$ 的比值即可给出 $\mathbb{P}(X>0)$ 的下界估计, 特别地, 在随机图方面, 我们可以更加"工具化"一下这个不等式, 也就是下面事件版本的二阶矩方法.

推论 2.5. *令 $B_n=A_{n,1}\cup A_{n,2}\cup\cdots\cup A_{n,m_n}$, 其中 $A_{n,1},A_{n,2},\cdots,A_{n,m_n}$ 是事件. 记 $i\overset{n}{\sim}j$, 若 $A_{n,i}$ 和 $A_{n,j}$ 不独立. 记*

$$\mu_n := \sum_{i=1}^{m_n} \mathbb{P}(A_{n,i}), \qquad \gamma_n := \sum_{i \overset{n}{\sim} j} \mathbb{P}(A_{n,i} \cap A_{n,j}),$$

则若 $\mu_n \to \infty$ 且 $\gamma_n = O(\mu_n^2)$ 当 $n \to \infty$, 则

$$\liminf_{n\to\infty} \mathbb{P}(B_n) > 0,$$

进一步, 若 $\gamma_n = o(\mu_n^2)$, 则

$$\lim_{n\to\infty} \mathbb{P}(B_n) = 1.$$

证明. 令 $X_n = \sum_{i=1}^{m_n} \mathbf{1}_{A_{n,i}}$, 分别计算 $\mathbb{E}[X_n]$ 和 $\mathbb{E}[X_n^2]$. 一方面,

$$\mathbb{E}[X_n] = \sum_{i=1}^{m_n} \mathbb{E}[\mathbf{1}_{n,i}] = \sum_{i=1}^{m_n} \mathbb{P}(A_{n,i}) = \mu_n,$$

另一方面,

$$\begin{aligned}
\mathbb{E}[X_n^2] &= \sum_{i=1}^{m_n} \mathbb{E}[\mathbf{1}_{n,i}] + \sum_{i,j} \mathbb{E}[\mathbf{1}_{n,i}\mathbf{1}_{n,j}] \\
&= \mu_n + \sum_{i,j} \mathbb{P}(A_{n,i} \cap A_{n,j}) \\
&= \mu_n + \sum_{i,j} \mathbb{P}(A_{n,i} \cap A_{n,j} - \mathbb{P}(A_{n,i})\mathbb{P}(A_{n,j})) + \sum_{i,j} \mathbb{P}(A_{n,i})\mathbb{P}(A_{n,j}) \\
&\le \mu_n + \sum_{i \overset{n}{\sim} j} \mathbb{P}(A_{n,i} \cap A_{n,j}) + \mu_n^2 \\
&= \mu_n + \mu_n^2 + \gamma_n.
\end{aligned}$$

于是,

$$\mathbb{P}(B_n) = \mathbb{P}(X_n > 0) \ge \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]} \ge \frac{\mu_n^2}{\mu_n + \mu_n^2 + \gamma_n} = \left(1 + \frac{1}{\mu_n} + \frac{\gamma_n}{\mu_n^2}\right)^{-1},$$

代入定理假设即证. □

注意此推论是要比定理 2.4弱的, 因为在计算 $\mathbb{E}[X_n^2]$ 时我们进行了放缩. 所以如果用推论 2.5得不到结果的时候, 可以再试试直接验证定理 2.4, 在后面的例子中也有所体现.

2.3 Chernoff-Cramér 方法

Chebyshev 不等式给出了随机变量尾概率平方反比的估计, 但对于很多情况而言, 显然不是最好的. 以标准正态分布为例, 即若 $X \sim N(0,1)$, 简单的计算表明

$$\mathbb{P}(|X| \ge \beta) \sim \sqrt{\frac{2}{\pi}} \beta^{-1} \exp\left(-\frac{\beta^2}{2}\right),$$

而右边要远好于 β^{-2} 的估计.

为了得到指数级别的估计, 我们需要引入**矩母函数** *(moment generating function)*.

定义 2.6. 对于随机变量 X, 定义其矩母函数

$$M_X(s) \coloneqq \mathbb{E}[e^{sX}] = \sum_{k\geq 0} \frac{s^k}{k!}\mathbb{E}[X^k]$$

若 $\mathbb{E}[e^{sX}]$ 有限.

由此, 我们有 Chernoff-Cramér 界.

引理 2.7 (Chernoff-Cramér 界)**.** 设随机变量 X, 且对存在 s_0, 使得对于任意 $s \in (-s_0, s_0)$, $M_X(s) < \infty$. 那么对任意 $\beta > 0$ 和 $s \in (0, s_0)$,

$$\mathbb{P}(X \geq \beta) \leq \exp(-s\beta + \Psi_X(s)),$$

这里 $\Psi_X(s) \coloneqq \log M_X(s)$ 为累积生成函数.

证明. 由 Markov 不等式,

$$\mathbb{P}(X \geq \beta) = \mathbb{P}(e^{sX} \geq e^{s\beta}) \leq \frac{\mathbb{E}[e^{sX}]}{e^{s\beta}} = \exp(-s\beta + \Psi_X(s)).$$

□

回到上面的标准正态分布, 我们熟知标准正态分布的矩母函数为

$$M_X(s) = \exp\left(\frac{s^2}{2}\right),$$

于是

$$\mathbb{P}(X \geq \beta) \leq \inf_{s>0} \exp\left(-s\beta + \frac{s^2}{2}\right) = \exp\left(-\frac{\beta^2}{2}\right),$$

给出的结果就比之前好得多.

2.3.1 次高斯随机变量

除了正态分布, 还有很大一类随机变量的尾概率也会有如此渐近性态, 如果我们归纳一下共性, 可以得到以下定义.

定义 2.8. 对随机变量 X, 设其期望为 μ, 若对某个 $\nu > 0$, 满足对任意 $s > 0$,

$$\Psi_{X-\mu}(s) \leq \frac{s^2\nu}{2},$$

则称 X 是参数为 ν 的次高斯随机变量 *(sub-Gaussian random variable)*, 记为 $X \in \mathrm{s}\mathcal{G}(\nu)$.

由定义, 我们立即得到

$$\mathbb{P}(X - \mu \geq \beta) \vee \mathbb{P}(X - \mu \leq -\beta) \leq \exp\left(-\frac{\beta^2}{2\nu}\right).$$

这里我们用到了 $-X$ 也是次高斯的.

定理 2.9 (Hoeffding 不等式)**.** *设相互独立的随机变量 $X_1, X_2, \cdots, X_n$ 满足 $X_i \in \mathrm{s}\mathcal{G}(\nu_i)$, $1 \le i \le n$, 这里 $0 < \nu_i < \infty$, 并给定 $w_1, w_2, \cdots, w_n \in \mathbb{R}$, $S_n = \sum_{i \le n} w_i X_i$, 则*

$$S_n \in \mathrm{s}\mathcal{G}\left(\sum_{i=1}^{n} w_i^2 \nu_i\right).$$

特别地,

$$\mathbb{P}(S_n - \mathbb{E}[S_n] > \beta) \le \exp\left(-\frac{\beta^2}{2\sum_{i=1}^{n} w_i^2 \nu_i^2}\right).$$

定理 2.9的证明是平凡的, 因为可以直接写出 S_n 的矩母函数.

2.3.2 次指数随机变量

可惜事无完美, 并非我们关心的所有随机变量都是次高斯随机变量. 比如对于 $X \sim N(0,1)$, 我们有

$$M_{X^2-1}(s) = \frac{1}{e^s(1-2s)^{1/2}}, \quad 0 < s < \frac{1}{2},$$

但当 $s \ge \frac{1}{2}$ 时, $M_{X^2-1}(s) = \infty$. 然而 Taylor 展开告诉我们在 $|s| < \frac{1}{4}$ 时,

$$\Psi_{X^2-1}(s) \le 2s^2,$$

所以是可以看作是"部分"次高斯的. 我们称这样的随机变量为**次指数随机变量** *(sub-exponential random variable)*.

定义 2.10. *对随机变量 X, 其期望为 μ, 称之为参数 (ν, α) 的次指数随机变量, 若对任意 $|s| < \frac{1}{\alpha}$,*

$$\Psi_{X-\mu}(s) \le \frac{s^2\nu}{2},$$

这里 $\nu, \alpha > 0$. 记为 $X \in \mathrm{s}\mathcal{E}(\nu, \alpha)$.

比如, 在上面的例子中, $X^2 \in \mathrm{s}\mathcal{E}(4,4)$.

再次使用引理 2.7, 有对任意 $\beta > 0$,

$$\mathbb{P}(X - \mu \ge \beta) \le \begin{cases} \exp\left(-\dfrac{\beta^2}{2\nu}\right), & 0 < \beta \le \dfrac{\nu}{\alpha}; \\ \exp\left(-\dfrac{\beta}{2\alpha}\right), & \beta > \dfrac{\nu}{\alpha}. \end{cases}$$

类似 Hoeffding 不等式, 次指数随机变量也有如下结果.

定理 2.11 (Bernstein 不等式)**.** *设相互独立的随机变量 $X_1, X_2, \cdots, X_n$ 满足 $X_i \in \mathrm{s}\mathcal{E}(\nu_i, \alpha_i)$, $1 \le i \le n$, 这里 $0 < \nu_i, \alpha_i < \infty$, 并给定 $w_1, w_2, \cdots, w_n \in \mathbb{R}$, $S_n = \sum_{i \le n} w_i X_i$, 则*

$$S_n \in \mathrm{s}\mathcal{E}\left(\sum_{i=1}^{n} w_i^2 \nu_i, \max_i |w_i| \alpha_i\right),$$

特别地,

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq \beta) \leq \begin{cases} \exp\left(-\dfrac{\beta^2}{2\sum_{i=1}^n w_i^2 \nu_i}\right), & 0 < \beta \leq \dfrac{\sum_{i=1}^n w_i^2 \nu_i}{\max_i |w_i|\alpha_i}; \\ \exp\left(-\dfrac{\beta}{2\max_i |w_i|\alpha_i}\right), & \beta > \dfrac{\sum_{i=1}^n w_i^2 \nu_i}{\max_i |w_i|\alpha_i}. \end{cases}$$

定理 2.11的证明同样是平凡的, 不再指出.

对于有界随机变量, 它是次高斯的, 同时也是次指数的. 应用定理 2.11, 有

定理 2.12 (有界随机变量的 Bernstein 不等式). *设相互独立的随机变量 $X_1, X_2, \cdots, X_n$ 满足 X_i 的期望为 μ_i, 方差为 σ_i^2, 且对某个 $0 < c < \infty$, $|X_i - \mu_i| \leq c$, 令 $S_n = \sum_{i=1}^n X_i$, 则*

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq \beta) \leq \begin{cases} \exp\left(-\dfrac{\beta^2}{4\sum_{i=1}^n \sigma_i^2}\right), & 0 < \beta \leq \dfrac{\sum_{i=1}^n \sigma_i^2}{c}; \\ \exp\left(-\dfrac{\beta}{4c}\right), & \beta > \dfrac{\sum_{i=1}^n \sigma_i^2}{c}. \end{cases}$$

证明. 只需证明 $X_i \in \mathrm{s}\mathcal{E}(2\sigma_i^2, 2c)$. 对 $k \geq 2$,

$$\mathbb{E}[|X_i - \mu_i|^k] \leq c^{k-2}\mathbb{E}[(X_i - \mu_i)^2] = c^{k-2}\sigma_i^2,$$

于是

$$\begin{aligned} \mathbb{E}(e^{s(X_i - \mu_i)}) &= 1 + \sum_{k=2}^{\infty} \frac{s^k}{k!}\mathbb{E}[(X_i - \mu_i)^k] \\ &\leq 1 + \sum_{k=2}^{\infty} \frac{s^k}{k!} c^{k-2}\sigma_i^2 \\ &\leq 1 + \frac{s^2\sigma_i^2}{2} + \frac{s^2\sigma_i^2}{6}\sum_{k=1}^{\infty}(cs)^k \\ &= 1 + \frac{s^2\sigma_i^2}{2} + \frac{s^2\sigma_i^2}{6}\frac{cs}{1-cs} \\ &\leq 1 + s^2\sigma_i^2 \\ &\leq \exp\left(s^2\sigma_i^2\right), \end{aligned}$$

当 $s \leq \frac{1}{2c}$, 进而结论得证. □

实际上, 也有有界版本的 Hoeffding 不等式, 但是在 β 比较小时会弱于定理 2.12, 这与我们的直觉不同: 好像次高斯要比次指数强, 但得到的不等式却弱于后者.

二阶矩方法和 Chernoff-Cramér 方法的比较

从上面可以看出, Chernoff-Cramér 方法对于离散和连续的随机变量均适用, 而且得到的尾概率量级是优于二阶矩方法的, 但其限制条件比较多, 可能需要计算出矩母函数才能给出估计, 然而对于一些互相依赖、不独立的随机变量之和, 计算矩母函数可能就没那么现实了. 而二阶矩方法对于处理离散概率的情况就会比较方便.

3 探索随机图: 阈值和尾概率

拥有了这些工具, 探索随机图的道路便会平坦许多.

我们首先来关注所谓"阈值"现象. 前面提到, 随机图的一些现象会随着概率的增大而发生变化, 我们希望找到一个变化的临界点. 将其严格化, 我们引入**阈值函数** *(threshold function)*. 设 $G_n \sim \mathbb{G}_{n,p_n}$, $n \in \mathbb{N}^*$, 对于图的某种性质 P, 若函数 $r(n)$ 满足

$$\lim_{n\to\infty} \mathbb{P}_{n,p_n}(G\text{有性质}P) = \begin{cases} 0, & p_n \ll r(n), \\ 1, & p_n \gg r(n), \end{cases}$$

则称 r 是图性质 P 的阈值函数.

3.1 最大团数阈值

定理 3.1. 性质 G_n包含完全子图K_4 的一个阈值函数为 $n^{-2/3}$.

证明. 记 X_n 为 $G_n \sim \mathbb{G}_{n,p_n}$ 中的 4− 团的个数, 则

$$\mathbb{E}[X_n] = \binom{n}{4} p_n^6 = \Theta(n^4 p_n^6).$$

先考虑 p_n 足够小的一侧. 由一阶矩方法, 若 $p_n \ll n^{-2/3}$, 则 $\mathbb{P}(G_n\text{包含完全子图}K_4) \to 0$.

再考虑另一侧. 记 $m_n = \binom{n}{4}$, $A_{n,1}, A_{n,2}, \cdots, A_{n,m_n}$ 为一列事件, 将这 n 个顶点组成的所有四元组任意排序, 其中 $A_{n,i}$ 表示第 i 组顶点的子图构成完全图, 若用推论 2.5的记号, 我们来计算 μ_n, γ_n.

首先

$$\mu_n = \mathbb{E}[X_n] = \binom{n}{4} p_n^6 = \Theta(n^4 p_n^6).$$

另一方面, 当 $p_n \gg n^{-2/3}$ 时, 先考虑何时 $i \overset{n}{\sim} j$. 当且仅当 i 组和 j 组的完全子图有公共边, 也就是它们之间有至少两个公共点的时候 $A_{n,i}$ 与 $A_{n,j}$ 不独立, 因此

$$\begin{aligned}
\gamma_n &= \sum_{i \overset{n}{\sim} j} \mathbb{P}(A_{n,i} \cap A_{n,j}) \\
&= \sum_i \left(\sum_{j\text{组},i\text{组有三个公共点}} \mathbb{P}(A_{n,i} \cap A_{n,j}) + \sum_{j\text{组},i\text{组有两个公共点}} \mathbb{P}(A_{n,i} \cap A_{n,j}) \right) \\
&= \sum_i \left(\sum_{j\text{组},i\text{组有三个公共点}} p_n^9 + \sum_{j\text{组},i\text{组有两个公共点}} p_n^{11} \right) \\
&= \sum_i \left(\binom{4}{3}\binom{n-4}{1} p_n^9 + \binom{4}{2}\binom{n-4}{2} p_n^{11} \right) \\
&= \Theta(n^5 p_n^9) + \Theta(n^6 p_n^{11}) \\
&= o(\mu_n^2),
\end{aligned}$$

而 $\mu_n \to \infty$ 当 $n \to \infty$, 于是由推论 2.5, $\mathbb{P}(G_n\text{包含完全子图}K_4) \to 1$. □

3.2 连通性阈值

我们从开始孤立点的存在性开始.

定理 3.2. 性质"没有孤立点"的一个阈值函数为 $\frac{\log n}{n}$.

证明. 设 X_n 为随机图 $G_n \sim \mathbb{G}_{n,p_n}$ 的孤立点数, 则当 $p_n \gg \frac{\log n}{n}$ 时,

$$\mathbb{E}[X_n] = n(1-p_n)^{n-1} \leq \exp(\log n - (n-1)p_n) \to 0.$$

由一阶矩方法,

$$\mathbb{P}(X_n > 0) \to 0, \qquad n \to \infty.$$

另一方面, 设 $B_n = A_{n,1} \cup A_{n,2} \cup \cdots \cup A_{n,n}$, 其中 $A_{n,i}$ 为事件"顶点 i 是孤立点". 若用推论 2.5的记号, 我们来计算 μ_n, γ_n.

类似上一例,

$$\mu_n = \mathbb{E}[X_n] = n(1-p_n)^{n-1} \sim \exp(\log n - np_n) \to \infty, \qquad n \to \infty.$$

再计算 γ_n. 对于任意的 $i \neq j$, $A_{n,i}, A_{n,j}$ 不独立, 于是

$$\begin{aligned}
\gamma_n &= \sum_{i\neq j} \mathbb{P}(A_{n,i} \cap A_{n,j}) \\
&= \sum_{i\neq j} (1-p_n)^{2n-3} \\
&= n(n-1)(1-p_n)^{2n-3} \\
&\neq o(\mu_n^2),
\end{aligned}$$

因此我们不能用推论 2.5的方法, 于是直接求 $\mathbb{E}[X_n^2]$. 注意 $\mathbb{E}[X_n^2] = \mu_n + \gamma_n$, 于是

$$\frac{\mathbb{E}[X_n]^2}{\mathbb{E}[X_n^2]} = \frac{\mu_n^2}{\mu_n + \gamma_n} = \frac{n^2(1-p_n)^{2n-2}}{n(1-p_n)^{n-1} + n(n-1)(1-p_n)^{2n-3}} \to 1, \qquad n \to \infty,$$

于是

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

由夹逼定理,

$$\mathbb{P}(B_n) = \mathbb{P}(X_n > 0) \to 1,$$

于是结论成立. □

备注 3.3. 如果把 $\gg$ 的条件进一步弱化为对任意 $\varepsilon > 0$, 若 $p_n \ge (1+\varepsilon)\frac{\log n}{n}$, 则直接计算得

$$\mathbb{P}(G\text{存在孤立点}) \to 0$$

仍然成立; 类似地, 若 $p_n \le (1-\varepsilon)\frac{\log n}{n}$, 则

$$\mathbb{P}(G\text{存在孤立点}) \to 1$$

也成立, 可以看出, 这个阈值更加"锐利". 于是对一般的阈值函数, 我们也可以定义其是否锐利. 设 $G_n \sim \mathbb{G}_{n,p_n}$, $n \in \mathbb{N}^*$, 对于图的某种性质 P, 若函数 $r(n)$ 满足对任意 $\varepsilon > 0$, 都有

$$\lim_{n\to\infty} \mathbb{P}_{n,p_n}(G\text{有性质}P) = \begin{cases} 0,\ p_n \le (1-\varepsilon)r(n), \\ 1,\ p_n \ge (1+\varepsilon)r(n), \end{cases}$$

则称 r 是图性质 P 的**锐利阈值函数** *(sharp threshold function)*, 从图 1中就可看出锐利阈值函数和一般阈值函数的对比.

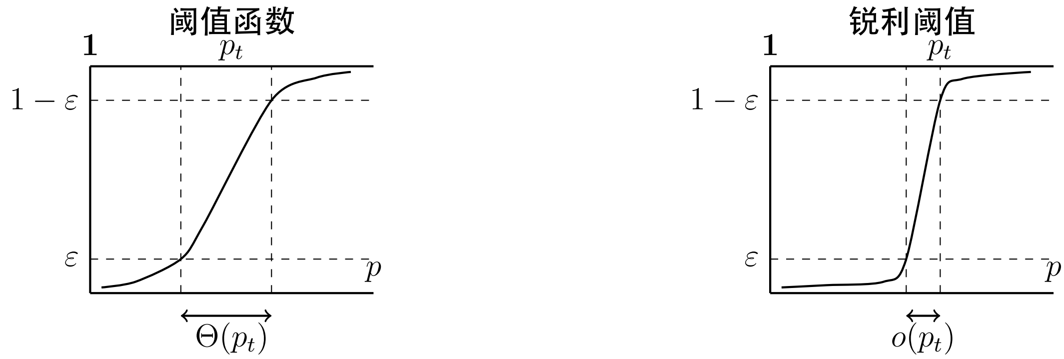


图 1: 阈值函数与锐利阈值函数的对比

研究了孤立点, 我们来看它的连通性. 注意不存在孤立点要比连通性条件弱, 但是它们的阈值函数居然相同. 我们有:

定理 3.4. 性质"图连通"的一个阈值函数为 $\frac{\log n}{n}$.

证明. 由定理 3.2, 当 $p_n \ll \frac{\log n}{n}$, G_n 几乎必然会出现孤立点, 于是其几乎必然不连通.

下面考虑另一侧. 当 $p_n \gg \frac{\log n}{n}$, 记事件 D_n 为 G_n 不连通. 设 Y_k 为与其他 $n-k$ 个顶点不连通的 $k-$ 顶点子图数目, 且 $k \in \{1,2,\cdots,\lfloor\frac{n}{2}\rfloor\}$($G_n$ 可以分为若干个连通分支, 其中最小的连通分支的规模一定不超过 $\lfloor\frac{n}{2}\rfloor$), 那么由一阶矩方法,

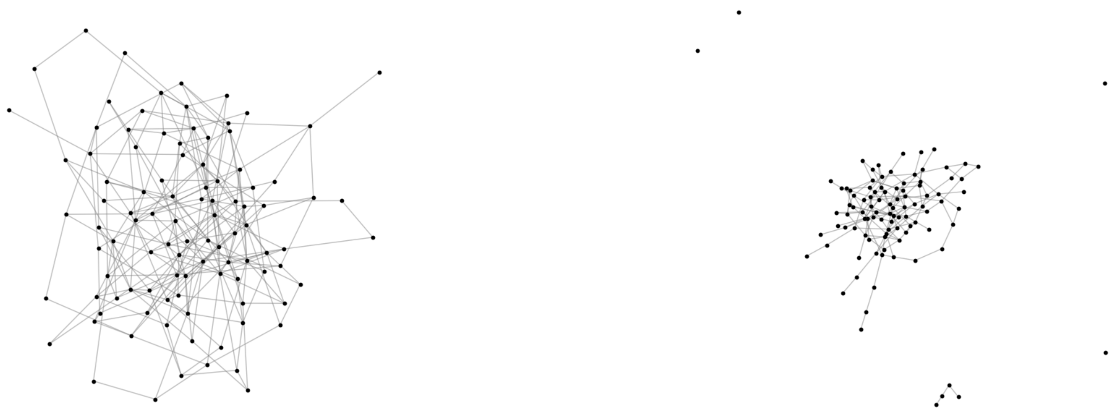
$$\mathbb{P}(D_n) = \mathbb{P}\left(\sum_{k=1}^{\lfloor\frac{n}{2}\rfloor} Y_k > 0\right) \le \sum_{k=1}^{\lfloor\frac{n}{2}\rfloor} \mathbb{E}[Y_k].$$

注意到

$$\begin{aligned}\sum_{k=1}^{\lfloor\frac{n}{2}\rfloor}\mathbb{E}[Y_k]&=\sum_{k=1}^{\lfloor\frac{n}{2}\rfloor}\binom{n}{k}(1-p_n)^{k(n-k)}\\&\leq\sum_{k=1}^{\lfloor\frac{n}{2}\rfloor}n^k(1-p_n)^{\frac{n}{2}k}\\&\leq\sum_{k=1}^{\infty}n^k(1-p_n)^{\frac{n}{2}k}\\&=\frac{n(1-p_n)^{\frac{n}{2}}}{1-n(1-p_n)^{\frac{n}{2}}}\\&=o(1),\qquad n\to\infty,\end{aligned}$$

于是 $\mathbb{P}(D_n)\to 0$, 当 $n\to\infty$, 结论得证. □

备注 3.5. 实际上我们也可以证明这个阈值是锐利的. 图 2展示了图的连通性.



(a) $p_n \gg \frac{\log n}{n}$, 很可能连通 (b) $p_n \ll \frac{\log n}{n}$, 很可能存在孤立点

图 2: 随机图的连通性变化

3.3 环的产生

定理 3.6. *性质"存在环"的一个阈值函数为* $\frac{1}{n}$.

证明. 一侧的证明是标准的, 仍然是一阶矩方法的应用. 记 X_n 为 G_n 环的个数, 则对于圈长为 k 的环, 如果选定顶点, 有 $\frac{(k-1)!}{2}$ 种成环方法, 于是

$$\mathbb{E}[X_n]=\sum_{k=3}^{n}\binom{n}{k}\frac{(k-1)!}{2}p_n^k\leq\sum_{k=3}^{n}(np_n)^k\leq\frac{(np_n)^3}{1-np_n}\to 0,$$

当 $p_n \ll \frac{1}{n}$, 进而 $\mathbb{P}(G_n\text{无环}) \to 0$.

对于另一侧, 我们可以验证当 $p_n \gg \frac{1}{n}$ 时, $\mathbb{P}(G_n\text{中含三角形}) \to 1$, 具体方法与定理 3.1类似, 留作习题. $\square$

由此可见, 当 $p_n \ll \frac{1}{n}$ 时, 随着 n 变大, G_n 几乎必然会变成森林, 如图 3, 然而对于 $c < 1$, 如果令 $p_n = \frac{c}{n}$, 却没有这样的性质, 所以这个阈值不是锐利的.

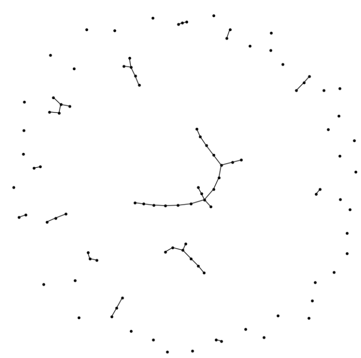


图 3: 当 $p_n \ll \frac{1}{n}$, 图很可能变成森林

结束之前, 我们还可以列几个阈值函数, 供读者练习:

- 性质"图的直径 ≤ 2(即对每组顶点 $x \neq y$, 存在 z, 使得 $\{x,z\},\{y,z\} \in V$)"的锐利阈值函数为 $\sqrt{\frac{2\log n}{n}}$.
- 性质"图仅含一个巨大连通分支和若干孤立点"的锐利阈值函数为 $\frac{\log n}{2n}$.

最后, 让我们以一个优雅而奇妙的定理结束我们对阈值的探讨.

对于图性质 P, 如果对任意满足 P 的图 G 任意加一条边, 仍然满足性质 P, 称 P 是单调的; 如果对任意充分大的 n, 总存在 n 个顶点的图 G 满足 P, 也存在 n 个顶点的图 G 不满足 P, 则称 P 是非平凡的, 那么我们有:

定理 3.7 (Bollobás and Thomason, 1987)**.** *每个非平凡单调图性质存在阈值函数.*

限于篇幅以及涉及到耦合技巧, 此定理的证明不在本文范围之内.

3.4 最大度估计

最后我们展示 Chernoff-Cramér 方法的一个应用, 我们通常喜欢将随机变量分解为若干示性随机变量之和, 示性随机变量是一致有界的 (都在 $[0,1]$ 内), 所以我们可以利用此特点, 结合有界版本的 Bernstein 不等式很好地估计尾概率, 我们以最大度的估计为例.

命题 3.8. 对于图 $G_n \sim \mathbb{G}_{n,p_n}$, 记 D_n 为 G_N 顶点度数最大值. 若 $np_n = \omega(\log n)$, 则对任意 $\varepsilon > 0$, 当 $n \to \infty$ 时,

$$\mathbb{P}(|D_n - (n-1)p_n| \geq 2\sqrt{(1+\varepsilon)np_n \log n}) \to 0.$$

证明. 对任意一个顶点 v, 其度数 $\delta(v) = S_{n-1} \sim B(n-1, p_n)$, 于是设

$$S_{n-1} = \sum_{k=1}^{n-1} X_k,$$

其中 $X_1, \cdots, X_{n-1}$ 独立同参数为 p_n 的伯努利分布, 于是在 Bernstein 不等式中取 $\mu_i = p_n$, $\sigma_i = p_n(1-p_n)$, 及 $c = 1$. 由定理 2.12, 若记 $\nu = (n-1)p_n(1-p_n)$, 则

$$\mathbb{P}(S_{n-1} - (n-1)p_n \geq \beta) \leq \begin{cases} \exp\left(-\dfrac{\beta^2}{4\nu}\right), & 0 < \beta \leq \nu, \\ \exp\left(-\dfrac{\beta}{4}\right), & \beta > \nu, \end{cases}$$

由假设, 取

$$\beta = 2\sqrt{(n-1)p_n(1-p_n)(1+\varepsilon)\log n} = o(\nu),$$

代入上面有

$$\mathbb{P}(S_{n-1} \geq (n-1)p_n + 2\sqrt{(n-1)p_n(1-p_n)(1+\varepsilon)\log n}) \leq n^{-1-\varepsilon},$$

于是

$$\mathbb{P}(D_{n-1} \geq (n-1)p_n + 2\sqrt{(n-1)p_n(1-p_n)(1+\varepsilon)\log n}) \leq n \cdot n^{-1-\varepsilon} \to 0$$

当 $n \to \infty$, 同理可得另一边估计. □

备注 3.9. 如果我们用 Hoeffding 不等式来估计, 我们只能取 $\beta = \sqrt{(1+\varepsilon)n \log n}$, 我们只能估计到 $O(1)$ 的数量级.

4 回顾与展望

虽然 Erdős–Rényi 模型为我们提供了一个简单而有效的工具来理解随机图的基本特性, 但它也有一些局限性. ER 模型假设所有节点之间的连接概率是相同的, 这使得它在描述现实世界中的许多网络时, 显得过于简化. 比如, 社交网络中并非每个人都与其他人有同等的联系; 而且我们通常会看到一些节点的连接非常密集, 而有些节点几乎没有连接.

为了克服这些问题, 研究者们提出了许多更复杂的随机图模型. 例如，渗流模型就更加关注网络中节点或边的“占据”情况, 特别是当某些部分被断开时, 网络的连通性如何发生变化; 小世界网络模型则很好地模拟了现实世界中的“六度分隔”现象, 网络中的节点通过少数几个

中介就能连接起来; 而无尺度网络模型则揭示了许多真实网络中, 少数几个“超级节点”拥有大量的连接, 这种现象在互联网和社交网络中非常普遍.

尽管 ER 模型简单易懂, 但随着研究的深入, 我们逐渐认识到, 现实世界中的网络远比我们想象的复杂. 随机图的世界远远不止于此, 未来我们将看到更多新模型的出现, 以及它们如何帮助我们理解社交网络、互联网、数据科学等领域的复杂性. 所以, 本文提供了随机图“入门”级的领略, 其真正深奥且吸引人的地方, 留待读者自己挖掘.

致谢

衷心感谢宗语轩学长在本文写作过程中给予的宝贵帮助, 特别感谢学长提出了不少深入的意见, 这些意见帮助我在文章的框架、逻辑以及细节处理上做出了重要的改进. 学长耐心细致的指导, 使我能够顺利完成这篇文章.

附录: 符号说明

为了方便读者理解本文中的数学公式和符号, 以下是本文中用到的主要符号的说明:

- $G = (V, E)$: 图, 其中 V 是节点集, E 是边集
- $[n]$: 代表集合 $\{1, 2, \cdots, n\}$
- $\sim$: 代表“分布”或表示等价量.
- O: 称 $f(n) = O(g(n))$, 若存在 $C > 0$, 使得任意充分大的 n, $|f(n)| \leq C|g(n)|$
- Ω: 称 $f(n) = \Omega(g(n))$, 若存在 $c > 0$, 使得任意充分大的 n, $|f(n)| \geq c|g(n)|$
- Θ: 称 $f(n) = \Theta(g(n))$, 若 $f(n) = O(g(n))$, $f(n) = \Omega g(n)$
- o: 称 $f(n) = o(g(n))$, 若当 $n \to \infty$, $f(n)/g(n) \to 0$
- ω: 称 $f(n) = \omega(g(n))$, 若 $g(n) = o(f(n))$
- $\ll$: 称 $f(n) \ll g(n)$, 若 $f(n) = o(g(n))$
- $\gg$: 称 $f(n) \gg g(n)$ 若 $g(n) = o(f(n))$

参考文献

[1] Sebastien Roch. *Modern discrete probability: An essential toolkit.* Cambridge University Press, 2024.

[2] Yufei Zhao. Lecture notes on probabilistic methods in combinatorics. PDF document, 2022. Massachusetts Institute of Technology, `http://yufeizhao.com/pm/`.