arXiv:2603.04315v1 [stat.ME] 4 Mar 2026

一种用于确定网络中社区数量的谱推断方法

Yujia Wu*[1]、Xiucai Ding†[2]、Jingfei Zhang‡[3]、Wei Lan§[4] 和 Chih-Ling Tsai¶[5]

[1]香港理工大学数据科学与人工智能系[2]加州大学戴维斯分校统计系[3]埃默里大学戈伊苏埃塔商学院 [4]西南财经大学数据智能与金融经济创新实验室及统计研究中心[5]加州大学戴维斯分校管理研究生院

摘要

为刻画网络数据中的社区结构，研究人员开发了多种块模型，包括随机块模型、度校正随机块模型、混合成员块模型、度校正混合成员块模型等。有效应用这些模型的关键步骤之一，是确定网络中的社区数量。然而，据我们所知，现有的网络社区数量估计方法要么依赖显式的模型拟合，要么无法同时适用于网络稀疏性和社区数量不断增长的情形。本文提出一种基于特征值间隙比的无模型谱推断方法，以应对这些挑战。该推断程序计算简便，无需参数调优，而且无需估计网络分布参数，即可应用于广泛的块模型。此外，该方法对稠密和稀疏网络均有效，即使社区数量不断增长也是如此。在技术层面，我们证明，在原假设下，所提谱检验统计量通过艾里核收敛到 I 型 Tracy-Widom 分布的一个函数；在弱备择假设下，该检验具有渐近功效。针对稠密和稀疏网络开展的模拟研究表明，所提方法有效。我们还通过三个真实案例展示所提检验的实用价值。

关键词：网络数据、社区检测、谱推断、特征值间隙比检验、Tracy-Widom 分布

*电子邮箱：yujwu@polyu.edu.hk

†电子邮箱：xcading@ucdavis.edu。XCD 获得美国国家科学基金会 DMS-2515104 的部分资助。

‡电子邮箱：emma.zhang@emory.edu§
电子邮箱：lanwei@swufe.edu.cn
¶电子邮箱：cltsai@ucdavis.edu

1 引言

自 Lorrain 和 White（1971年）以及 Granovetter（1973年）识别网络中的社区结构以来，人们提出了各种块模型，用以刻画网络生成机制中潜在的块结构，并提高网络数据的有效利用率。举几个例子：随机块模型（SBM）（Holland 等人，1983年；Abbe，2018年）、度修正随机块模型（DCSBM）（Karrer 和 Newman，2011年；Jin 等人，2023年）、混合成员模型（MM）（Airoldi 等人，2008年；Ma 等人，2021年），以及度修正混合成员模型（DCMM）（Jin 等人，2017年；Han 等，2023年；Jin 等人，2024年）。

在这些模型中，网络数据被建模为不含自环的无向随机图。它们假设相应的邻接矩阵 $A = (A_{ij}) \in \mathbb{R}^{n\times n}$ 是对称的，且所有对角元素均为零。此外，假设网络由伯努利分布随机图模型生成；也就是说，

$$A_{ij} \sim \text{Bernoulli}(P_{ij}), \quad \text{for every } i < j, \tag{1.1}$$

其中 $P = (P_{ij}) \in \mathbb{R}^{n\times n}$ 是对称的边概率矩阵。具体而言，P_{ij} 表示节点 i 与 j之间存在连接的概率，其中 $i \neq j$。在上述块模型——SBM、DCSBM、MM 和 DCMM——中，主要差异在于对 P所施加的具体假设和形式。然而，一个共同且关键的特征是，概率矩阵 P 通常具有秩 K（Jin 等人，2023年），其中 K 也标识了社区数量。

然而在实践中，社区数量 K 通常是先验未知的，*a priori* 因此必须在分析具体块模型之前进行估计。本文旨在开发一种理论上可靠且实践中高效的序贯检验方法，该方法具有普适性，因为它不依赖于任何特定块模型对 P 所作的结构性和正则性假设，并且即使当 K随 n发散时仍然有效。下文将首先在第1.1节总结现有方法，然后在第 1.2节概述我们提出的方法。

1.1 现有文献综述

如上所述，现有关于估计 K 的文献大多依赖于拟合特定的块模型。为简化综述和阐述，我们聚焦于度修正混合成员模型（DCMM），它涵盖了上述其他常见的网络块模型。度修正混合成员模型可

表述如下。

设 $Q = (Q_{kl}) \in \mathbb{R}^{K\times K}$ 为对称社区交互矩阵。令 $\boldsymbol{\omega} = (\omega_1, \ldots\ldots, \omega_n)^\top$ 表示用于刻画度异质性的节点特定度参数向量，其中对所有 i 和 $\sum_{i=1}^{n}\omega_i = n$， $\omega_i > 0$ 。此外，每个节点 i 都关联一个 K 维概率质量向量

$$\boldsymbol{\pi}_i = (\pi_{i,1}, \cdots, \pi_{i,K})^\top, \quad \text{with } \pi_{i,k} \geqslant 0 \text{ and } \sum_{k=1}^{K}\pi_{i,k} = 1,$$

其中，$\pi_{i,\,k}$ 表示节点 i 属于社区 k的概率。DCMM 假设，对于 P_{ij}，式（1.1）满足

$$P_{ij} = \omega_i\omega_j\,\boldsymbol{\pi}_i^\top Q\boldsymbol{\pi}_j, \quad \text{for } 1 \leqslant i \leqslant j \leqslant n. \tag{1.2}$$

请注意，SBM、DCSBM 和 MM 都是上述统一公式的特例。具体而言，当 $\omega_i \equiv 1$ 且每个 $\boldsymbol{\pi}_i$ 均为退化分布时，DCMM 退化为 SBM，即 $\pi_{i,\,k} \in \{0,\ 1\}$；当每个 $\boldsymbol{\pi}_i$ 均为退化分布，而 ω_i 在不同节点间变化时，DCMM 退化为 DCSBM；当 $\omega_i \equiv 1$ 且 $\pi_{i,\,k} \in (0,\ 1)$ 允许重叠社区成员关系时，DCMM 退化为 MM。

近年来，在满足各种正则性条件的前提下，人们针对上述块模型开发了多种基于推断的方法来估计 K。相关文献大致可分为两类：序贯检验方法（参见 (1.3)）和非序贯检验方法。对于序贯检验方法，Lei（2016）基于标准化邻接矩阵的最大特征值，为 SBM 开发了一种拟合优度检验；而 Hu 等人（2021）则针对 SBM 和度修正随机块模型提出了逐元素偏差统计量，用于拟合优度检验。不过，这两项研究都聚焦于稠密情形，即假设对所有 $1 \leqslant i,\ j \leqslant n$，边概率 P_{ij} 均为常数阶，以确保其检验统计量中所用标准化邻接矩阵具有一定的普适性。在稀疏情形下，当 $\max_{i,j} P_{ij} \downarrow 0$ 时，Jin 等人 (2023) 提出了一种逐步拟合优度（StGoF）程序，用于在假设 K 固定时，对 SBM 和度修正随机块模型中的 K 进行最优估计。值得注意的是，上述所有方法都需要预先估计未知的网络参数 $\boldsymbol{\pi}_i$、Q 和 $\boldsymbol{\omega}$，以便在确定 K 之前计算检验统计量。对于可能发散的 K，当 $K = O(\log\log n)$ 时，Han 等人（2023）提出了一种通过残差子抽样进行秩推断的程序，用于检验 K，且适用于 (1.2)。然而，RIRS 程序

包含一个调优参数，必须谨慎选择，以确保在构造检验统计量时进行有效的子抽样。为方便读者，表格总结了这些基于序贯的推断方法的局限性1。

方法	稠密网络 含 $P_{ij} \asymp 1$	稀疏网络 含 $P_{ij} \ll 1$	发散数目 的社区	无参数 估计（无模型）	无调优 参数
所提方法	✓	✓	✓	✓	✓
Lei （2016）	✓	×	✓	×	✓
Hu 等人 (2021年)	✓	×	✓	×	✓
Han 等人 (2023年)	✓	✓	✓	✓	×
Jin 等人 (2023年)	✓	✓	×	×	✓

表1：下文所述的我们所提方法与其他序贯推断方法的比较。符号“✓”表示适用，“×”表示不适用。

需要指出的是，还存在非序贯检验方法，包括网络交叉验证方法（Chen 和 Lei，2018年）以及基于似然的方法（Daudin 等人，2008年；Wang 和 Bickel，2017年；Saldana 等人，2017年；Noroozi 等人，2021年；Hu 等人，2020年；Ma 等人，2021年）。值得注意的是，除了计算复杂度较高之外，这些方法还存在与表1所总结的类似局限性，尤其是基于似然的方法（Wang 和 Bickel，2017年；Noroozi 等人，2021年；Hu 等人，2020年；Ma 等人，2021年；Jin 等人，2023年）。

受上述挑战的启发，本文提出一种基于序贯谱推断的方法，该方法不依赖于特定的模型假设，允许网络为稠密或稀疏，并允许 K 随 n发散。所提方法的概述见第1.2节。

1.2 所提方法概述

本文基于序贯检验框架，通过考虑如下假设，建立一种用于确定 K 的谱推断程序：

$$\mathbf{H}_0 : \mathsf{K} = \mathsf{K}_0 \quad \text{v.s.} \quad \mathbf{H}_1 : \mathsf{K}_0 < \mathsf{K} \leqslant \mathsf{K}_{\max}, \tag{1.3}$$

其中，K 表示 P 的未知真实秩，K_0 表示 P 的假设秩，$\mathsf{K}_{\max}$ 表示 K 的一个可能上界。这里我们考虑单侧备择假设，因为一个具有 $\mathsf{K} > \mathsf{K}_0$ 个社区的网络，也可以通过拆分一个或多个真实社区，表示为具有 $\mathsf{K}_0 > \mathsf{K}$ 个社区的网络。Lei（2016）、Chen 和 Lei（2018）、Wang 和 Bickel（2017）以及 Hu 等人（2021）也考虑了这种单侧备择假设。注意，通过（1.3）可以得到 K 的一个自然的相合序贯检验估计，

即，

$$\widehat{\mathsf{K}} := \inf\{\mathsf{K}_0 \geqslant 0 : \ \mathbf{H}_0 \text{ is accepted}\}. \tag{1.4}$$

为应对第1.1节讨论的挑战，我们在算法1 中提出了一种基于邻接矩阵 A选定特征值的（1.3）推断程序。所提方法不依赖于关于 P 的任何特定模型假设，并允许K趋于无穷。该方法在理论上可靠、计算效率高，且无需使用调优参数。为检验（1.3），对于适当选取的$\mathsf{K}_{\max}$，所提出的算法采用以下检验统计量

$$\mathbb{T} = \frac{\lambda_{\mathsf{K}_0+1}(A) - \lambda_{\mathsf{K}_{\max}+1}(A)}{\lambda_{\mathsf{K}_{\max}+1}(A) - \lambda_{\mathsf{K}_{\max}+2}(A)}, \tag{1.5}$$

其中 $\lambda_1(A) \geqslant \lambda_2(A) \geqslant \cdots \geqslant \lambda_n(A)$ 表示 A 的特征值，按降序排列。正如第3节的理论结果将证明的，在 $\mathbb{T}$在$\mathbf{H}_0$ 下的分布，可通过艾里核由某种I型Tracy–Widom分布的函数形式刻画。在$\mathbf{H}_1$下， $\mathbb{T}$ 以 $\mathsf{O}(n^{2/3})$, 的速率发散，因此是检验（1.3）的有力检验统计量。

在实践中，构造 $\mathbb{T}$ 需要指定一个上界$\mathsf{K}_{\max}$。理论上，正如定理1所确立的，可将$\mathsf{K}_{\max} = \mathsf{K}_0 + \mathsf{C}$ 取为一个足够大的固定整数C。如我们的模拟和真实数据分析所示，所提方法对$\mathsf{K}_{\max}$的选择具有稳健性，只要其在合理范围内选取。为实现自动化实施，第2 节给出了一种实用的、数据驱动的程序，用于选取$\mathsf{K}_{\max}$，其依据为下述文献中讨论的置换法：Dobriban（2020年）。

我们指出，相似但却颇为不同的特征值比统计量已被用于在K被假定为固定时，借助大型样本协方差矩阵估计因子个数或信号个数；参见Onatski （2009）；Bao等人（2015年）；Ding 和 Yang （2022年）。关于该差异的更详细讨论，请读者参阅注记1 。

尽管 $\mathbb{T}$ 的极限分布可以在理论上确立，但 P 未知且可能复杂的结构使其无法给出显式而紧凑的公式，用以在$\mathbf{H}_0$下推导临界值。为解决这一问题，我们进一步证明，目标分布可通过校准被准确模拟，具体做法是从均值零、高斯元素方差为 $1/n$, 的 $n \times n$ 对称Wigner矩阵（即高斯正交系综（GOE）矩阵）中精心选取的特征值构造适当的量。因此，临界值可通过校准准确且高效地确定。与许多现有方法不同，我们强调，本文所提出的

算法不涉及调优参数的过程，因此具有计算效率。如第4和5节所示，无论在模拟研究还是真实数据分析中，我们的方法在准确性、功效和计算效率方面均优于现有方法。

如表1所示，我们所提出的方法适用于稠密和稀疏网络，并允许 K 随 n发散。一般而言，网络越稠密，所允许的社区数量就越多。为严格刻画这种权衡并建立 $\mathbb{T}$的渐近分布，我们提出一个同时涉及稀疏程度和社区数量的充分条件。更具体地说，我们要求 $n^{1/3}\max_{i,\ j} P_{ij}/K^2 \gg 1$，以此明确刻画二者之间的权衡；示例见补充材料中的图 S.2 。该条件还原了 $K \ll n^{1/6}$ 这一稠密随机块模型的上界，参见 Lei (2016)。此外，对于常用的块模型，在条件（1.2）下很容易满足这一条件。推导最优条件以及针对极稀疏网络发展相应的结果和方法，超出了本文的研究范围，将留待未来工作探讨。

本文其余部分安排如下。第2节介绍所提出的算法，第3 节建立其理论性质。第4 节展示模拟研究，其结果支持所提方法的优势，相关总结见表1。第5 节报告两项实证分析，第6节给出结论。所有技术细节、补充讨论和数值结果均置于补充文件中。

2 我们所提出的算法

在本节中，我们正式介绍所提出的算法。如第1.2节所述，该算法依赖于检验统计量 $\mathbb{T}$ 在（1.5）中的构造，这需要指定一个上界$K_{\max}$。从理论上讲（参见定理1），给定假设检验秩K_0，，只要$K_{\max} - K_0$ 保持有界，我们程序的有效性便得到保证。在实际实现中，我们采用基于置换的方法（也称为平行分析），如Dobriban（2020年）中所研究的那样；该方法已被广泛用于基于样本协方差矩阵的因子分析中的秩确定，并且在计算上可行且高效。

该流程概述如下，完整细节见补充材料中的算法 S.1，以供读者参考。简而言之，给定邻接矩阵A，通过独立打乱每个 A_τ 中的条目，构造置换后矩阵

列，并多次重复此置换程序。如果 A 的特征值超过从置换后矩阵 A中得到的相应特征值的指定百分位数（例如 95%），则选取该特征值。令K_{PA} 表示平行分析选取的特征值数量。根据Dobriban (2020年中的讨论，以高概率有$\mathsf{K}_{\mathrm{PA}} \geqslant \mathsf{K}$。受此结果启发，为了获得实用的上界，我们在模拟和真实数据分析中始终设定$\mathsf{K}_{\max} = \mathsf{K}_{\mathrm{PA}} + 5$。这种自动选择方法在实践中效果良好。

完成上述准备后，我们现在在算法1中提出所提方法，用于检验（1.3）。在步骤1中，我们使用通过算法$\mathsf{K}_{\max}$ 选出的S.1来构造检验统计量（1.5），具体讨论如上。在步骤2中，受定理1理论结果的启发，我们使用J 个相互独立的高斯正交系综（GOE）矩阵计算临界值。为方便起见，我们选择显著性水平 α 以及J，使得 $(1-\alpha)\mathsf{J}$ 为整数。

算法1： 所提谱推断程序

输入： 邻接矩阵 A，K_0, 个合成GOE矩阵，其数量为J（例如，$\mathsf{J} = 1000$），以及显著性水平 α。

1. 检验统计量的构造。

（i）。从算法$\mathsf{K}_{\max}$ 中计算S.1。

(ii). 构造检验统计量 $\mathbb{T}$，见（1.5）。

2. 临界值的校准。

(i). **对于** $\mathsf{j} = 1$ **到** J，**执行**

(a). 生成一个 GOE矩阵 W_{j}使得 $W_{\mathsf{j}} = W_{\mathsf{j}}^{\top}$，其中非对角元 $W_{\mathsf{j},\ k<l}$ 独立同分布 $\sim \mathcal{N}(0,\ n^{-1})$ 且对角元 $W_{\mathsf{j},\ kk}$ 独立同分布$\sim \mathcal{N}(0,\ 2/n)$。

(b). 计算校准统计量

$$\mathbb{T}_{W_{\mathsf{j}}} = \frac{\lambda_1(W_{\mathsf{j}}) - \lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+1}(W_{\mathsf{j}})}{\lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+1}(W_{\mathsf{j}}) - \lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+2}(W_{\mathsf{j}})}. \tag{2.1}$$

（ii）。将 c_α 计算为 J 个模拟值的 $(1-\alpha)$ 分位数$\{\mathbb{T}_{W_1},\ \cdots,\ \mathbb{T}_{W_{\mathsf{J}}}\}$。

输出： 拒绝原假设 $\mathbf{H}_0$，当其位于（1.3）中时 $\mathbb{T} > c_\alpha$。

接下来，我们对所提出的算法的有效性给出启发式讨论与评注。严格的理论分析推迟到第3节。假设真实秩为K。在$\mathbf{H}_0$ （见（1.3））下，我们证明出现在两者中的特征值间隙

分子与分母为同阶，这意味着统计量 $\mathbb{T}$在（1.5）中的阶为$O(1)$。相比之下，在备择假设 $K > K_0$, 下，分子显著大于分母，因此 $\mathbb{T}$ 以$O(n^{2/3})$的阶增长。这种分离使检验统计量具有功效。为选取适当的临界值并确保检验准确，我们建立 $\mathbb{T}$的渐近分布，该分布没有闭式表达式。尽管如此，只要特征值选取适当，该分布就可以由高斯正交系综（GOE）矩阵生成的统计量准确模拟，如（2.1）所述。

我们指出，我们的方法在计算上也是可行的。一方面，步骤2——计算临界值——可以离线完成，并将所得数值制成表格，以供检验程序使用。另一方面，该方法只需计算$K_{\max} + 2$ 个 A的最大特征值。对此已有许多高效的数值线性代数算法可供使用，因此能够快速完成计算，尤其适用于稀疏网络。关于计算方面的进一步讨论，以及不同方法计算时间的比较，请参阅补充材料第S.3.1 节。

注记1。我们指出，尽管基于特征值比的检验统计量已用于大样本协方差矩阵的推断和估计问题——例如，参见（*Onatski, 2009*; *Ding, 2021年*; *Ding 和 Yang, 2021年*; *Bao et al., 2015年*; *Ding 和 Yang, 2022*; *Ding 和 Xie, 2025*）——这些研究通常假设（K在*1.3*）中是固定的。相比之下，我们针对网络邻接矩阵提出的统计量和检验程序则截然不同。

首先，在上述研究中，所构造的比值通常基于相邻特征值之间的间距，而我们的统计量则考虑索引间隔更大的特征值间距。这种设计使得 K 通常可以发散。其次，许多现有方法由于其统计量的特定构造方式，要求总体矩阵的特征值互不相同，而我们的方法对矩阵 P没有此类限制。最后，尽管也可以使用合成矩阵来模拟这些研究中的渐近分布，但它们通常依赖矩形 *Wishart* 矩阵，并且需要选择一个长宽比作为调优参数。相比之下，我们的校准基于高斯正交系综（*GOE*）矩阵，无需进行此类调优。

3 理论论证

本节将建立所提方法的理论性质。在第3.1节中，我们推导原假设 H_0下式（1.5）中的所提检验统计量 $\mathbb{T}$ 的渐近分布。随后，第3.2节将研究检验的功效，并说明相应估计量在式（1.4）中的性能。本文始终采用以下约定。对于依赖于 n、 a_n =的序列 $\{a_n\}$ 和 $\{b_n\}$ ，(b_n)O $|a_n| \leqslant$ 表示对于某个常数 C > 0，有 $|b_n|$ C；而 a_n =o(b_n) 表示当 $n \to \infty$ 时，对于某个正序列 $c_n \downarrow 0$ ，有 $|a_n| \leqslant c_n|b_n|$ 。我们还用 $a_n \asymp b_n$ 同时表示 $a_n =$ O(b_n) 和$b_n =$O(a_n)。

3.1 $\mathbb{T}$的渐近分布

我们首先介绍随机矩阵理论中的一些著名结果。考虑一个高斯正交系综（GOE）矩阵 $W \in \mathbb{R}^{n\times n}$，其中 $W_{ij} = W_{ji}$，对于 $i \neq j$的非对角元W_{ij}相互独立，且分布为 $\mathcal{N}(0,$ $n^{-1})$，对角元 W_{ii}的分布为 $\mathcal{N}(0,$ $2/n)$。众所周知，经过适当归一化后， W的最大特征值，即 $n^{2/3}(\lambda_1(W) - 2),$，依分布收敛于Tracy和Widom（1994，1996）提出的I型Tracy–Widom分布。此外，对任意固定的C > 0， W的前C个最大特征值在归一化 $\{n^{2/3}(\lambda_k(W) - 2)\}_{1\leqslant k\leqslant}$C,后的联合分布具有由艾里核描述的联合密度，通常也称为Tracy–Widom 定律；例如，参见Forrester（1993）。

做好上述准备后，我们现在借助 GOE矩阵 W陈述 $\mathbb{T}$ 的渐近分布。在本节余下部分中，我们采用网络分析中的标准假设（Bickel 和 Chen，2009；Zhao 等，2012），即 max$_{i,j} P_{ij} \asymp$ min$_{i,j} P_{ij}$。本节的主要结果如下。

定理*1.* 假设 $n^{1/3}$ *max*$_{i,j} P_{ij}/K^2 \to \infty$ *as* $n \to \infty$，且对于某个常数 $c > 0$，*min* $_{1\leqslant k\leqslant}$K $|\lambda_k(P)| \geqslant cn$*max*$_{i,j} P_{ij}/K$。那么在原假设 H_0 *in (1.3)* 下，当 $K_{\max} - K_0$ 有限时，对于任意 $x \in \mathbb{R}$，有 *lim*$_{n\to\infty} F_{\mathbb{T}}(x) =$ *lim*$_{n\to\infty} F_{\mathbb{T}_W}(x)$，其中 $F_{\mathbb{T}}(x)$ 是检验统计量 $\mathbb{T}$的累积分布函数，而 $F_{\mathbb{T}_W}(x)$ 是 $\mathbb{T}_W$的累积分布函数，其定义为

$$\mathbb{T}_W = \frac{\lambda_1(W) - \lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+1}(W)}{\lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+1}(W) - \lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+2}(W)},$$

其中， W 是一个*GOE*矩阵。

一方面，定理1 表明，在某些条件下，检验统计量 $\mathbb{T}$ 的分布可以渐近地用 $\mathbb{T}_W$的分布来刻画。注意，由于 $\mathbb{T}_W$可以改写如下

$$\mathbb{T}_W = \frac{n^{2/3}(\lambda_1(W)-2) - n^{2/3}(\lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+1}(W)-2)}{n^{2/3}(\lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+1}(W)-2) - n^{2/3}(\lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+2}(W)-2)},$$

那么，根据本节开头的讨论，我们关注的量本质上是 I型Tracy–Widom分布的一个函数，该分布由艾里核刻画。另一方面，这种表示也使我们能够通过校准计算检验的临界值（1.3）。具体而言，对于给定的名义水平α，令 $c_\alpha = F^{-1}_{\mathbb{T}_W}(1-\alpha)$表示 $\mathbb{T}_W$的上 α 分位数。临界值 c_α随后通过数值近似 $F_{\mathbb{T}_W}$获得，例如采用算法1中的蒙特卡洛模拟，并在 $\mathbf{H}_0$ 满足 $\mathbb{T} > c_\alpha$时拒绝原假设。下面对定理的假设作几点说明。

注 *2.* 我们指出，条件 $n^{1/3}\max_{i,j} P_{ij}/K^2 \to \infty$ 明确体现了网络稀疏性（由 $\max_{i,j} P_{ij}$ 衡量）与 K 可能发散之间的权衡；示意见补充材料中的图 *S.2*。一方面，当 K 固定或缓慢增长时，该条件允许 $\max_{i,j} P_{ij} \downarrow 0$ ，因而适用于稀疏网络。因此，它比 *Lei*（*2016*）和胡等人（*2021*）的条件限制更少，后者假设网络稠密且 $P_{ij} \asymp 1$ 对所有 $1 \leqslant i, j \leqslant n$ 均成立。另一方面，当网络较稠密时，同一条件允许 $K \to \infty$。特别地，对于 *Lei*（*2016*）所考虑的稠密随机块模型，该条件涵盖了 $K \ll n^{1/6}$ 的情形；同时，它也比 *Jin* 等人 *(2023)* 中 K 固定的假设，以及 *Han* 等人（*2023*）中 K 极缓慢发散（例如 $O(\log \log n)$）的假设限制更少。

尽管如此，这一条件主要是出于技术上的考虑而设定的，旨在确保在$\mathbf{H}_0$ （*1.3*）中，统计量 $\mathbb{T}$ 的分布可以用 *Tracy–Widom* 定律描述；更多详情请参见我们补充材料中的注 *S.1* 。当这一条件不成立时——无论是在*Han* 等人（*2023*年）和*Jin* 等人（*2023*年）所研究的极稀疏网络中，还是像胡等人（*2021*年）那样，社区数量增长得足够快时——我们目前的方法可能会失效。其背后的技术原因与*Huang* 等人（*2020*年）讨论的情况类似：所提出的统计量可能不再服从 *Tracy–Widom* 分布，而是收敛到高斯极限，或高斯分量与 *Tracy–Widom* 分量的自由加性卷积。由于对这些情形进行严格分析超出了本文的范围，我们将其留待未来工作研究。

注3。 我们指出，本文还对 P的最小非零特征值作出了假设。从直观上看，这一假设可解释为信噪比条件，确保能够根据邻接矩阵P 的领先特征值，可靠地推断其谱行为（进而推断其秩） A。类似条件已在多种情形下采用；例如，参见（*Lei 和 Rinaldo，2015年*；*Jin 等人，2023年*）。我们还请读者参阅补充材料中的注*S.3*，了解更多详情。在温和条件下，某些块模型很容易满足这一假设。例如，在研究充分的同配随机块模型中，有$\min_{1\leqslant k\leqslant \mathsf{K}} |\lambda_k(P)| \asymp n(a-b)/\mathsf{K}$ （*Abbe，2018年*；*Amini 和 Levina，2018年*），其中 a 和 b分别表示社区内和社区间的连接概率。在这种情况下，我们的假设自动成立。当节点特定度参数为常数量级时，类似论证也适用于度修正随机块模型。

3.2 功效与序贯估计量的分析

本节研究所提出的统计量 $\mathbb{T}$ 的功效，并据此对估计量（1.4）进行渐近分析。如前所述，在$\mathbf{H}_1$ （1.3）成立时，定理 1 的条件不再成立，而 $\mathbb{T}$预期会取较大值，从而对备择假设表现出很强的判别能力。结果如下。

定理 2。 假设定理 *1* 的条件成立，且$\mathbf{H}_1$ （*1.3*）成立。那么，对于任意大的常数 $\mathsf{C}>0$，有

$$\mathbb{P}(\mathbb{T} > \mathsf{C}n^{2/3}) = 1 - \mathrm{o}(1).$$

该定理意味着，对于给定的名义水平 α，在备择假设 $\mathbf{H}_1$，下有 $\mathbb{P}(\mathbb{T} > c_\alpha) = 1 - \mathrm{o}(1)$，因此原假设 $\mathbf{H}_0$ 将以趋于 1 的概率被拒绝。特别地，定理 2 表明，在备择假设下，$\mathbb{T}$ 至少以 $\mathrm{O}(n^{2/3})$ 的速率发散，因此我们的检验比 Han 等人（2023年）和 Hu 等人（2021年）提出的方法更有功效。具体而言，在备择假设下，Han 等人（2023年）的检验统计量始终远离零，而 Hu 等人（2021年）的检验统计量仅以 $\mathrm{O}(\log n)$ 的阶发散。此外，已知 Hu 等人（2021年）的检验统计量在植入分区随机块模型下缺乏功效；该模型是一种社区规模相等、 Q_{ll} 相等且 Q_{lk} 相等（对于 $l \neq k$）的随机块模型。

借助定理 1 和 2，我们可以分析根据（1.4）生成的相关估计量 $\widehat{\mathsf{K}}$。在我们的检验框架下，这等价于定义

$$\widehat{\mathsf{K}} = \min\left\{\mathsf{K}_0 \in \{1, \ldots, \mathsf{K}_{\max}\} : \mathbb{T} \leqslant c_\alpha\right\}.$$

该估计量的性质总结如下。

推论 1. 在定理*1* 和 *2*的假设下，我们有

$$\mathbb{P}\left(\widehat{\mathsf{K}} = \mathsf{K}\right) = 1 - \alpha - \mathrm{o}(1).$$

该推论表明，K 在名义水平K α下能够有效地估计真实社区数。特别是，当 $\alpha \downarrow 0$（等价地，当 $c_\alpha = \mathsf{o}(n^{2/3})$ 趋于无穷）时，推论 1 意味着 $\mathbb{P}\left(\widehat{\mathsf{K}} = \mathsf{K}\right) = 1 - \mathsf{o}(1)$，从而得到$\mathsf{K}$的相合估计量。然而，在实际的序贯检验程序中，名义水平 α 通常保持固定，而不是随着 n趋于零，因为数值计算中令 $\alpha \downarrow 0$ 往往会导致经验水平表现不可靠。基于这一考虑，我们可以遵循Lei(2016）；Ding 和 Yang (2022) 来考虑估计量

$$\widehat{\mathsf{K}}^* = \min\left\{\mathsf{K}_0 \in \{1, \ldots, \mathsf{K}_{\max}\} : \mathbb{T} \leqslant t_n\right\},$$

其中 $t_n \to \infty$ 当 $n \to \infty$ 且与名义水平 α 独立。K^*的性质将在以下推论中给出。

推论 2. 在定理 *1* 和 *2*的假设下，当 t_n 满足 $t_n \asymp n^\epsilon$对某个 $\epsilon \in (0,\ 5/6)$, 我们有

$$\mathbb{P}\left(\widehat{\mathsf{K}}^* = \mathsf{K}\right) = 1 - \mathrm{o}(1).$$

4 数值模拟

为评估算法 1的有限样本表现，我们基于（1.2）开展模拟研究，以检验（1.3）中的假设，并在多种设定下同时考虑稠密和稀疏网络设置。

为了生成 $\boldsymbol{\pi}_i$，我们遵循 Han 等人 (2023年) 的方法，并从以下两组中的一组进行模拟：$\mathbf{PM}_{\mathsf{K}} = \{\boldsymbol{e}_1,\ \cdots,\ \boldsymbol{e}_{\mathsf{K}}\}$，其中 $\boldsymbol{e}_k$ 表示 k 在 $\mathbb{R}^{\mathsf{K}}$ 中的第 k 个标准基向量 $\mathbb{R}^{\mathsf{K}}$，而 **混合成员模型**$_{\mathsf{K}}$ =

$\{(0.2,\ 0.8,\ 0,\ \underbrace{\cdots\ 0}),\ ,\ (0.8,\ 0.2,\ 0,\ \underbrace{\cdots\ 0}),\ ,\ (\mathrm{K}\underbrace{^{-1}\ \cdots\ ^{-1}})\},\ ,\ \mathrm{K}$。

为了生成 ω_i，我们考虑两种 K–2 K–2K 设置：对所有 $1 \leqslant i \leqslant n$ 均采用 $\omega_i \equiv 1$，或者遵循 Zhao 等（2012）的方法，从期望为 1 的分布中独立抽取 $\omega_i' s$：

$$\omega_i = \begin{cases} \eta_i, & \text{w.p.} \quad 0.8, \\ 9/11, & \text{w.p.} \quad 0.1, \\ 13/11, & \text{w.p.} \quad 0.1, \end{cases} \tag{4.1}$$

其中 η_i 在区间 $[4/5, 6/5]$上服从均匀分布。 Q 矩阵的设置见第 4.1节和第 4.2节，分别刻画稠密和稀疏网络设置。

根据上述设置，为按照（1.2）生成具有块结构的网络，我们考虑三种常用块模型的不同变体。

1. 随机块模型（SBM）。我们设 $\boldsymbol{\pi}_i \in \mathbf{PM}_\mathrm{K}$ 以及 $\omega_i \equiv 1$. 对于每个社区 k，我们随机选择 n/K 个节点，并设置其对应的 $\boldsymbol{\pi}_i = \boldsymbol{e}_k$，使得 $\sum_{i=1}^n \pi_{i,\ k} = n/\mathrm{K}$。

2. 度修正随机块模型（DCSBM）。我们设 $\boldsymbol{\pi}_i \in \mathbf{PM}_\mathrm{K}$，其中 ω_i 根据（4.1）生成。对于每个社区 k，我们随机选择 n/K 个节点，并设置其对应的 $\boldsymbol{\pi}_i = \boldsymbol{e}_k$，使得 $\sum_{i=1}^n \pi_{i,\ k} = n/\mathrm{K}$。

3. 度修正混合成员模型（DCMM）。我们设置 $\boldsymbol{\pi}_i \in \mathrm{PM}_\mathrm{K} \cup \mathrm{MM}_\mathrm{K}$，其中 ω_i 按照（4.1）生成。对于每个社区 k，遵循 Han 等人（2023）的方法，我们随机选取 $n_0 := (\mathrm{K}^{-1} - 0.03)n$ 个节点，并将它们的 $\boldsymbol{\pi}_i$ 分配给 $\boldsymbol{e}_k$。由此得到总计 $n_0\mathrm{K}$ 个节点，其 $\boldsymbol{\pi}_i$ 从 PM_K 中抽取。对于其余节点，我们为 MM_K 中的每个向量随机分配 $(n - n_0\mathrm{K})/3$ 个节点，并相应地分配它们的 $\boldsymbol{\pi}_i$。

在所有模拟中，网络规模设为 $n = 3,000$，K 取值为 $\in \{3, 5, 10, 15, 20\}$。第4.1节研究稠密网络，第 4.2节分析稀疏网络。我们将所提出的统计量和算法1 中的方法（记为 $\mathbb{T}$）与另外五种方法进行比较：Lei （2016）中的两种方法，分别记为 $\mathbb{T}_{\mathrm{Lei}}$ 和 $\mathbb{T}_{\mathrm{Lei,boot}}$；Han 等人（2023年）中的方法，记为 $\mathbb{T}_{\mathrm{Han}}$；以及胡等人（2021年）中的两种方法，分别记为 $\mathbb{T}_{\mathrm{Hu}}$ 和 $\mathbb{T}^{\mathrm{aug}}{}_{\mathrm{Hu,boot}}$。其中，$\mathbb{T}_{\mathrm{Lei,boot}}$ 和 $\mathbb{T}^{\mathrm{aug}}{}_{\mathrm{Hu,boot}}$ 分别表示对 $\mathbb{T}_{\mathrm{Lei}}$ 和 $\mathbb{T}_{\mathrm{Hu}}$应用自助法校正和增强自助法程序得到的检验，

分别。由于 Lei(2016) 和 胡等人(2021年) 的方法不适用于度修正混合成员模型，因此我们在度修正混合成员模型的模拟研究中省略这些方法。

4.1 稠密网络的性能

本节比较各种方法在稠密网络上的表现，生成设置遵循 胡 等人 （2021年）；Han 等人（ 2023 年）中的设置。 Q 对于稠密 SBM 和稠密度校正随机块模型，我们设定 $Q_{kl} = 0.1(1 + 4 \times \mathbb{I}(k = l))$；对于稠密DCMM，当 $Q_{kl} = 0.1^{|k-l|}$时取 $k \neq l$，否则取 $Q_{kl} = (\mathsf{K} + 1 - k)/\mathsf{K}$ 。

表 2–4 报告了稠密SBM、稠密度校正随机块模型和稠密DCMM的经验水平和检验功效。总体而言，模拟研究表明，在稠密网络下的三个块模型中，我们提出的检验均表现良好，并优于其他方法。这些发现与我们的理论结果一致。下面将对这些发现进行更详细的阐述。

对于表 2 和3中呈现的稠密SBM和度修正随机块模型结果， $\mathbb{T}$、 $\mathbb{T}_{\text{Han}}$ 和 $\mathbb{T}^{\text{aug}}_{\text{胡, 启动}}$ 的经验水平接近真实的5%名义水平，尽管 $\mathbb{T}^{\text{aug}}_{\text{胡, 启动}}$的表现略逊于 $\mathbb{T}$ 和 $\mathbb{T}_{\text{Han}}$。对于其他方法，在SBM下，当$\mathsf{K} > 5$时， $\mathbb{T}_{\text{Hu}}$、 $\mathbb{T}_{\text{Lei}}$ 和$\mathbb{T}_{\text{Lei, boot}}$ 的经验水平明显偏离名义水平。在度修正随机块模型下，当$\mathsf{K} > 5$时， $\mathbb{T}_{\text{Hu}}$ 会出现尺寸扭曲，且可能远大于5%。此外，对于SBM和度修正随机块模型，当$\mathsf{K} = 3$较小时， $\mathbb{T}_{\text{Lei}}$ 和$\mathbb{T}_{\text{Lei, boot}}$ 的经验水平均显著偏离5%的名义水平。其次， $\mathbb{T}$和 $\mathbb{T}^{\text{aug}}_{\text{胡, 启动}}$ 的经验检验功效通常等于1，其中 $\mathbb{T}$ 的检验功效高于 $\mathbb{T}^{\text{aug}}_{\text{胡, 启动}}$，而 $\mathbb{T}_{\text{Hu}}$ 的检验功效较低。值得注意的是， $\mathbb{T}_{\text{Han}}$ 几乎没有检验功效。此外，在SBM下， $\mathbb{T}_{\text{Lei}}$和$\mathbb{T}_{\text{Lei, boot}}$ 的检验功效通常等于1。然而，在度修正随机块模型下，由于 $\mathbb{T}_{\text{Lei}}$ 和$\mathbb{T}_{\text{Lei, boot}}$存在尺寸扭曲，我们不再进一步分析它们的经验检验功效。

对于表4中报告的稠密DCMM结果，一方面， $\mathbb{T}$ 的经验水平接近名义上的5%水平。相比之下， $\mathbb{T}_{\text{Han}}$ 在$\mathsf{K} > 5$时表现出明显的水平失真，这表明该检验 $\mathbb{T}_{\text{Han}}$ 可能不适用于较大的K值。另一方面， $\mathbb{T}$ 的经验检验功效随着$\mathsf{K} - \mathsf{K}_0$ 的增大而提高。由于 $\mathbb{T}_{\text{Han}}$的经验水平存在失真，且其在$\mathsf{K} > 5$时的经验检验功效不稳定，因此我们不再进一步分析其经验检验功效表现。

	$\mathbb{T}$					$\mathbb{T}_{\text{Han}}$				
K	3	5	10	15	20	3	5	10	15	20
$K_0 = 3$	0.052	1	1	1	1	0.052	0.068	0.044	0.038	0.026
$K_0 = 5$	*	0.060	1	1	1	*	0.044	0.030	0.048	0.044
$K_0 = 10$	*	*	0.042	1	1	*	*	0.048	0.030	0.046
$K_0 = 15$	*	*	*	0.058	1	*	*	*	0.038	0.040
$K_0 = 20$	*	*	*	*	0.048	*	*	*	*	0.046
	$\mathbb{T}_{\text{Hu}}$					$\mathbb{T}^{\text{aug}}_{\text{Hu,boot}}$				
K	3	5	10	15	20	3	5	10	15	20
$K_0 = 3$	0.068	0.110	0.828	0.994	1	0.066	0.994	1	1	1
$K_0 = 5$	*	0.080	0.580	0.976	1	*	0.042	1	1	1
$K_0 = 10$	*	*	0.130	0.500	0.986	*	*	0.066	1	1
$K_0 = 15$	*	*	*	0.242	0.560	*	*	*	0.034	0.914
$K_0 = 20$	*	*	*	*	0.398	*	*	*	*	0.044
	$\mathbb{T}_{\text{Lei}}$					$\mathbb{T}_{\text{Lei,boot}}$				
K	3	5	10	15	20	3	5	10	15	20
$K_0 = 3$	0.064	1	1	1	1	0.052	1	1	1	1
$K_0 = 5$	*	0.072	1	1	1	*	0.096	1	1	1
$K_0 = 10$	*	*	0.182	1	1	*	*	0.138	1	1
$K_0 = 15$	*	*	*	0.614	1	*	*	*	0.730	1
$K_0 = 20$	*	*	*	*	0.846	*	*	*	*	0.930

表 2：Em　pirical水平与检验功效，针对名义水平 $\alpha = 0.05$.下的稠密随机块模型

4.2 稀疏网络的性能

本节比较各种方法在稀疏网络上的表现。与第4.1节类似，对于稀疏随机区块模型和DCSBM，我们设定 $Q_{kl} = 0.1n^{-2/9}(1 + 4 \times \mathbb{I}(k = l))$,；对于稀疏DCMM，当 $Q_{kl} = 0.5n^{-2/9}0.1^{|k-l|}$ 时取 $k \neq l$，并且当 $Q_{kl} = 0.5n^{-2/9}(\mathsf{K} + 1 - k)/\mathsf{K}$ 时取其他值。

表5–7 报告了稀疏随机区块模型、稀疏度修正随机块模型和稀疏DCMM的经验水平和检验功效。总体而言，模拟研究表明，在稀疏网络下，所提方法在这三种块模型中表现良好，且优于其他方法。我们将在下文进一步详细讨论这些结果。

对于表5 和6中报告的稀疏随机区块模型和度修正随机块模型结果， $\mathbb{T}$ 和 $\mathbb{T}_{\text{Han}}$ 的经验尺寸接近真实的5%名义水平。相比之下，尽管$\mathbb{T}_{\text{Lei,boot}}$ 和$\mathbb{T}^{\text{aug}}$ 胡,启动 在少数设定下取得了较为合理的经验尺寸，但其表现并不稳定，且其大多数经验尺寸呈现出明显失真。此外， $\mathbb{T}_{\text{Hu}}$ 和 $\mathbb{T}_{\text{Lei}}$显著偏离名义水平。就功效而言， $\mathbb{T}$的经验检验功效随着$\mathsf{K} - \mathsf{K}_0$ 的增大而稳步提高。相比之下， $\mathbb{T}_{\text{Han}}$ 几乎没有功效。由于

	$\mathbb{T}$					$\mathbb{T}_{\text{Han}}$				
K	3	5	10	15	20	3	5	10	15	20
$K_0 = 3$	0.036	1	1	1	1	0.046	0.054	0.044	0.036	0.050
$K_0 = 5$	*	0.056	1	1	1	*	0.064	0.050	0.052	0.046
$K_0 = 10$	*	*	0.048	1	1	*	*	0.054	0.050	0.050
$K_0 = 15$	*	*	*	0.034	1	*	*	*	0.050	0.054
$K_0 = 20$	*	*	*	*	0.044	*	*	*	*	0.052
	$\mathbb{T}_{\text{Hu}}$					$\mathbb{T}^{\text{aug}}_{\text{Hu,boot}}$				
K	3	5	10	15	20	3	5	10	15	20
$K_0 = 3$	0.036	0.276	0.840	0.988	1	0.028	0.854	1	1	1
$K_0 = 5$	*	0.042	0.466	0.842	1	*	0.052	0.890	1	1
$K_0 = 10$	*	*	0.182	0.766	1	*	*	0.072	0.900	1
$K_0 = 15$	*	*	*	0.304	1	*	*	*	0.062	0.990
$K_0 = 20$	*	*	*	*	0.712	*	*	*	*	0.040
	$\mathbb{T}_{\text{Lei}}$					$\mathbb{T}_{\text{Lei,boot}}$				
K	3	5	10	15	20	3	5	10	15	20
$K_0 = 3$	0.110	1	1	0.994	0.990	0.064	1	0.588	0.572	0.646
$K_0 = 5$	*	1	0.936	0.978	0.984	*	0.006	0	0.008	0.688
$K_0 = 10$	*	*	0.268	0.820	0.810	*	*	0.084	0.012	0.006
$K_0 = 15$	*	*	*	0.114	0.612	*	*	*	0.014	0
$K_0 = 20$	*	*	*	*	0.162	*	*	*	*	0.026

表3：实　　验尺寸和检验功效：名义水平为 $\alpha = 0.$ 时稠密度校正随机块模型的检验功效 $05.$

	$\mathbb{T}$					$\mathbb{T}_{\text{Han}}$				
K	3	5	10	15	20	3	5	10	15	20
$K_0 = 3$	0.058	0.344	0.610	0.864	0.966	0.034	1	1	0.420	0.802
$K_0 = 5$	*	0.056	0.330	0.440	0.646	*	0.040	1	0.408	0.768
$K_0 = 10$	*	*	0.038	0.184	0.390	*	*	0.112	0.414	0.766
$K_0 = 15$	*	*	*	0.040	0.186	*	*	*	0.342	0.740
$K_0 = 20$	*	*	*	*	0.036	*	*	*	*	0.688

表4：稠密DCMM在名义水平 $\alpha = 0.05.$下的经验尺寸和检验功效

$\mathbb{T}_{\text{Lei}}$、 $\mathbb{T}_{\text{Lei,boot}}$、 $\mathbb{T}_{\text{Hu}}$和 $\mathbb{T}^{\text{aug}}$ 胡,启动 的经验尺寸严重失真，我们不再进一步分析其经验检验功效。

表7中报告了稀疏DCMM的结果。一方面， $\mathbb{T}$ 的经验水平接近名义上的5%，而 $\mathbb{T}_{\text{Han}}$ 的经验水平则在$K > 5$时出现偏差。另一方面， $\mathbb{T}$ 的经验检验功效随着$K - K_0$的增大而提高。由于 $\mathbb{T}_{\text{Han}}$ 的水平存在偏差，且其在$K > 5$时的经验检验功效不稳定，因此我们不再进一步分析 $\mathbb{T}_{\text{Han}}$的检验功效。

	$\mathbb{T}$					$\mathbb{T}_{\text{Han}}$				
K	3	5	10	15	20	3	5	10	15	20
$K_0 = 3$	0.054	0.634	0.792	0.916	0.968	0.052	0.064	0.042	0.038	0.052
$K_0 = 5$	*	0.048	0.766	0.810	0.948	*	0.056	0.070	0.068	0.046
$K_0 = 10$	*	*	0.054	0.520	0.840	*	*	0.040	0.046	0.058
$K_0 = 15$	*	*	*	0.052	0.290	*	*	*	0.044	0.062
$K_0 = 20$	*	*	*	*	0.038	*	*	*	*	0.042
	$\mathbb{T}_{\text{Hu}}$					$\mathbb{T}^{\text{aug}}_{\text{Hu,boot}}$				
K	3	5	10	15	20	3	5	10	15	20
$K_0 = 3$	0.258	0.830	0.990	1	1	0.060	0.596	0.992	1	1
$K_0 = 5$	*	0.380	1	1	1	*	0.040	0.840	0.972	1
$K_0 = 10$	*	*	0.992	1	1	*	*	0.016	0.848	1
$K_0 = 15$	*	*	*	1	1	*	*	*	0.026	0.982
$K_0 = 20$	*	*	*	*	1	*	*	*	*	0.088
	$\mathbb{T}_{\text{Lei}}$					$\mathbb{T}_{\text{Lei,boot}}$				
K	3	5	10	15	20	3	5	10	15	20
$K_0 = 3$	0.990	0.994	1	1	1	0.080	0.976	0.850	0.982	1
$K_0 = 5$	*	1	1	1	1	*	0.162	0.636	0.840	1
$K_0 = 10$	*	*	1	1	1	*	*	0.016	0.932	1
$K_0 = 15$	*	*	*	1	1	*	*	*	0.074	1
$K_0 = 20$	*	*	*	*	1	*	*	*	*	0

表5：实　　证结果：在名义水平 $\alpha = 0.05$ 下，稀疏随机区块模型的检验规模与检验功效　　.

5 真实数据分析

本节通过分析两个真实世界网络，展示所提方法的实用价值：第5.1节介绍的政治博客网络 Adamic and Glance (2005年) 以及第5.2节介绍的新浪微博网络 Wu 等人 (2022)，并将其性能与现有方法进行比较。补充材料的S.3节还介绍了对西蒙斯学院的一个脸书网络的另一项应用。

5.1 政治博客网络

第一个数据集记录了2004年美国总统选举前两个月内互联网博客之间的链接，最初由 Adamic 和 Glance (2005年) 进行研究。节点代表博客，边表示博客之间的网页链接。遵循以往研究中的惯例（Lei，2016；Hu 等人，2021年；Han 等人，2023年），我们聚焦于该网络的最大连通分量，其中包含1,222个节点，边密度约为2.24%。

由于所有博客 根据政治立场分为“保守派”或“自由派”社区

	$\mathbb{T}$					$\mathbb{T}_{\text{Han}}$				
K	3	5	10	15	20	3	5	10	15	20
$K_0 = 3$	0.040	0.410	0.544	0.890	1	0.040	0.052	0.058	0.042	0.050
$K_0 = 5$	*	0.054	0.250	0.694	0.900	*	0.064	0.060	0.056	0.048
$K_0 = 10$	*	*	0.046	0.512	0.620	*	*	0.048	0.038	0.040
$K_0 = 15$	*	*	*	0.042	0.510	*	*	*	0.066	0.052
$K_0 = 20$	*	*	*	*	0.036	*	*	*	*	0.046
	$\mathbb{T}_{\text{Hu}}$					$\mathbb{T}^{\text{aug}}_{\text{Hu,boot}}$				
K	3	5	10	15	20	3	5	10	15	20
$K_0 = 3$	0.020	0.830	0.932	0.756	0.650	0.060	0.484	0.802	0.910	1
$K_0 = 5$	*	0.130	1	1	1	*	0.168	0.750	0.626	0.942
$K_0 = 10$	*	*	0.650	1	1	*	*	0.058	0.542	1
$K_0 = 15$	*	*	*	0.808	1	*	*	*	0.250	0.854
$K_0 = 20$	*	*	*	*	1	*	*	*	*	0.026
	$\mathbb{T}_{\text{Lei}}$					$\mathbb{T}_{\text{Lei,boot}}$				
K	3	5	10	15	20	3	5	10	15	20
$K_0 = 3$	0.230	1	1	1	1	0.034	0.862	0.644	0.850	0.456
$K_0 = 5$	*	0.168	1	1	1	*	0.180	0.590	0.248	0.832
$K_0 = 10$	*	*	0.642	1	1	*	*	0.076	0.340	0.250
$K_0 = 15$	*	*	*	0.136	1	*	*	*	0.034	1
$K_0 = 20$	*	*	*	*	0	*	*	*	*	0

表6：经验　　尺寸和稀疏度修正随机块模型在名义水平 $\alpha = 0.05$. 下的检验功效

	$\mathbb{T}$					$\mathbb{T}_{\text{Han}}$				
K	3	5	10	15	20	3	5	10	15	20
$K_0 = 3$	0.044	0.286	0.460	0.638	0.872	0.048	1	1	0.910	0.956
$K_0 = 5$	*	0.038	0.250	0.482	0.600	*	0.040	1	0.900	0.952
$K_0 = 10$	*	*	0.052	0.206	0.390	*	*	0.410	0.898	0.944
$K_0 = 15$	*	*	*	0.038	0.154	*	*	*	0.786	0.938
$K_0 = 20$	*	*	*	*	0.040	*	*	*	*	0.900

表 7：在名义水平 $\alpha = 0.05$. 下稀疏DCMM 的经验尺寸和检验功效

立场，该网络中的真实社区数量可视为$K = 2$。我们采用检验统计量 $\mathbb{T}$，并结合竞争方法$\mathbb{T}_{\text{Lei、启动}}$、 $\mathbb{T}^{\text{aug}}{}_{\text{胡、启动}}$以及 $\mathbb{T}_{\text{Han}}$，检验$K_0 = 1$ 和 2。检验结果汇总于表8，表明 $\mathbb{T}$、$\mathbb{T}_{\text{Han}}$ 和$\mathbb{T}^{\text{aug}}{}_{\text{胡、启动}}$拒绝$\mathbf{H}_0$：$K = 1$，但不拒绝$\mathbf{H}_0$：$K = 2$。相比之下，$\mathbb{T}_{\text{Lei、启动}}$拒绝所有原假设。因此，$\mathbb{T}$的结果与 $\mathbb{T}_{\text{Han}}$ 和 $\mathbb{T}^{\text{aug}}{}_{\text{胡、启动}}$的结果一致，并支持真实情况：$K = 2$。

K_0	$\mathbb{T}$	$\mathbb{T}_{\text{Han}}$	$\mathbb{T}_{\text{Lei, 启动}}$	$\mathbb{T}^{\text{aug}}_{\text{Hu, boot}}$
1	拒绝	拒绝	拒绝	拒绝
2	**接受**	**接受**	拒绝	**接受**

表 8：政治博客网络的检验结果。

5.2 新浪微博数据

第二个数据集由Wu 等人(2022)进行分析，包含 2,580 名新浪微博用户之间的好友关系链接，边密度约为 0.41%。在Wu 等人(2022)中，邻接矩阵 A 定义于有向网络，其中 $A_{ij} = 1$ 若用户 i 关注用户j。通过分割程序，Wu 等人(2022)表明，可以根据两个有向节点影响力指标将用户划分为四个象限："内向影响力"（对影响的接受程度）和"外向影响力"（对他人施加的影响）。

为适配我们的无向网络框架，我们重新定义 $A_{ij} = 1$ 如果用户 i 并且 j 彼此关注。在该网络中，每对节点的两个方向影响力指标自然合并为一个无向相互影响力指标。应用Wu 等人(2022的分割方法得到两个社区。因此，该网络的真实社区数取为$K = 2$。

我们报告检验 $K_0 = 1$ 和 2 的结果，因为所有 $K_0 > 2$ 检验都得出相同结论。如表 9 所示，$\mathbb{T}$ 拒绝 $H_0 : K = 1$，但不拒绝 $H_0 : K = 2$，从而正确识别出至少两个社区。相比之下，$\mathbb{T}_{\text{Lei, boot}}$、$\mathbb{T}^{\text{aug}}_{\text{胡, 启动}}$和 $\mathbb{T}_{\text{Han}}$拒绝所有原假设。

因此，$\mathbb{T}$ 的检验结果与真实情况 $K = 2$一致，表明我们的方法在稀疏网络中有效，而其他方法则无法正确识别社区结构。

K_0	$\mathbb{T}$	$\mathbb{T}_{\text{Han}}$	$\mathbb{T}_{\text{Lei, 启动}}$	$\mathbb{T}^{\text{aug}}_{\text{Hu, boot}}$
1	拒绝	拒绝	拒绝	拒绝
2	**接受**	拒绝	拒绝	拒绝

表 9：新浪微博网络的检验结果。

6 结论

本文提出了一种无模型谱方法，用于检验网络中的社区数量。该方法计算效率高，且易于实现。与现有方法相比，我们的检验无需显式进行模型拟合，在适用于稀疏网络和社区数量不断增长方面具有明显优势。具体而言，我们建立了网络稀疏性与社区数量之间的显式权衡关系，即 $n^{1/3}\max_{i,\ j} P_{ij}/\mathrm{K}^2 \ \to \ \infty$. 我们证明，在适当的正则性条件下，所提出的检验统计量在原假设下收敛至一个由艾里核刻画的 I 型 Tracy–Widom 分布的泛函。我们还推导了该检验的渐近功效，并证明了所提出的序贯估计量具有一致性。

未来研究仍有三个值得探索的方向。首先，将我们的方法扩展到极稀疏网络（Han 等人，2023年）在理论上颇具意义。其次，将所提框架推广至非参数图模型（Wolfe 和 Olhede，2013）是另一个重要方向。第三，将相关二元网络数据纳入考量，例如由潜在空间模型（Hoff 等人，2002）生成的数据，也值得进一步研究。我们相信，这些进展将进一步提升所提方法的适用性。

补充材料

在本补充材料中，我们给出详细的技术证明、辅助引理、补充说明以及更多数值结果。在本文余下部分，我们将始终采用以下约定。对于任意矩阵 B，用 $B_{m\times n}$ 表示由 B 的最后 m 行和 n 列组成的子矩阵。此外，令 $\lambda_s(B)$ 表示 B的第 s大$\tilde{}$特征值，并将 $\lambda_s(B)$ 定义为 B 按绝对值计算的第 s大特征值。此外，对于矩阵 B，定义 $\|B\| = \lambda_1(B^\top B)^{1/2},$ ；对于向量 $\boldsymbol{b}$，定义$|\boldsymbol{b}| = (\boldsymbol{b}^\top \boldsymbol{b})^{1/2}$。回顾一下，对于依赖于 n的序列 $\{a_n\}$ 和 $\{b_n\}$ ， a_n =O(b_n) 表示$|a_n| \leqslant \mathrm{C}|b_n|$ ，其中 C > 0为某个常数；而 $a_n =$ o (b_n) 表示 $|a_n| \leqslant c_n|b_n|$ ，其中 $c_n \downarrow 0$ 是某个正序列，且当 $n \to \infty$时成立。我们还用 $a_n \asymp b_n$ 同时表示 $a_n =$ O(b_n) 和$b_n =$ O(a_n)。对于依赖于 n的两个随机变量序列 $\{a_n\}$ 和 $\{b_n\}$ ，若 $a_n =$O(b_n) 以概率 $1 -$ o(1)成立，则记作 $a_n =$ O$_{\mathbb{P}}(b_n)$ 。类似地，若 $a_n =$ o(b_n) 以概率 $1 -$o(1)成立，则可定义$a_n =$ o$_{\mathbb{P}}(b_n)$ 。

S.1 技术证明

本节将给出第3节中结果的技术证明。回顾邻接矩阵 $A = (A_{ij}) \in \mathbb{R}^{n\times n}$ 并考虑

$$\widetilde{A} = A - \mathbb{E}(A), \tag{S.1}$$

其中， $\mathbb{E}(A)$ 是 A的期望值。因此， A 具有零均值元素和零对角元素。显然， A 的元素具有有限方差。给定 $\mathbb{E}(A) = P - \mathrm{diag}(P)$， （S.1）可改写为

$$A = P + \widetilde{A} - \mathrm{diag}(P), \tag{S.2}$$

其中，$\mathsf{diag}(P)$ 表示由 $(P_{11}、 \cdots、 P_{nn})^\top$构成的对角矩阵。

S.1.1 定理1的证明

定理1 的证明可分为以下五步。为简化符号，记

$$p_{\max} = \max_{i,j} P_{ij}.$$

回顾（S.1）和（S.2）。步骤 I 表明，对于 $1 \leqslant s \leqslant \mathsf{K}_{\max} - \mathsf{K}_0 + 2$,

$$\frac{1}{np_{\max}}\lambda_{\mathsf{K}_0+s}(AA^\top) = \frac{1}{np_{\max}}\lambda_s(\breve{A}_2\breve{A}_2^\top) + \mathrm{o}_{\mathbb{P}}\left(n^{-2/3}\right),$$

其中 $\breve{A}_2$ 是一个超经过适当变换后的 A 子矩阵。步骤 II 表明

$$\frac{1}{\sqrt{np_{\max}}}\lambda_{\mathsf{K}_0+s}(A) = \frac{\xi_s}{\sqrt{np_{\max}}}\lambda_s(\widetilde{A}_{G_I}) + \mathrm{o}_{\mathbb{P}}\left(n^{-2/3}\right), \tag{S.3}$$

其中， $\xi_s > 0$ 是某个常数， A_{G_I} 是 A 的一个子矩阵，并且 $\lambda_s(A_{G_I}) \asymp \lambda_s(A_2)$以 $1 - \mathsf{o}(1)$的概率成立。步骤 III 证明

$$c_{i^*j^*}\widetilde{\Omega}_{G_I,i^*j^*}\lambda_{\mathsf{K}_0+s}(A) = \lambda_s(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) + \mathrm{o}_{\mathbb{P}}\left(n^{-2/3}\right),$$

其中， $c_{i^*j^*}$和 $\widetilde{\Omega}_{G_I,\ i^*j^*}$是满足 $c_{i^*j^*}\widetilde{\Omega}_{G_I,\ i^*j^*} \asymp \frac{1}{\sqrt{np_{\max}}}$、 的某些确定性标量，而 $A_{G_I} \circ \Omega_{G_I}$ 表示 A_{G_I}的方差归一化版本。第四步表明，对于

$1 \leqslant s \leqslant \mathsf{K}_{\max} - \mathsf{K}_0 + 2$ 以及任意 $x_s \in \mathbb{R}$,

$$\begin{aligned}&\lim_{n\to\infty} \mathbb{P}\big(n^{2/3}(\lambda_1(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \mathsf{L}) \leqslant x_1, \cdots, n^{2/3}(\lambda_s(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \mathsf{L}) \leqslant x_s\big)\\ =&\lim_{n\to\infty} \mathbb{P}\big(n^{2/3}(\lambda_1(W) - 2) \leqslant x_1, \cdots, n^{2/3}(\lambda_s(W) - 2) \leqslant x_s\big),\end{aligned}$$

其中， W 是一个GOE矩阵，L是某个常数。步骤 V推导得出，对任意 $x \in \mathbb{R}$,

$$\lim_{n\to\infty} F_{\mathbb{T}}(x) = \lim_{n\to\infty} F_{\mathbb{T}_W}(x),$$

其中， $F_{\mathbb{T}}(x)$ 和 $F_{\mathbb{T}_W}(x)$ 分别是 $\mathbb{T}$ 和 $\mathbb{T}_W$的累积分布函数。接下来，我们按照上述五个步骤给出证明的详细过程。

定理1的证明。步骤 I。为了将 P 的对角元控制在 $\mathsf{O}(n^{-1/3})$ 以内以推导式（S.3），同时保持 $\mathbb{E}(A) = P - \operatorname{diag}(P)$ 不变，以便于技术分析，我们对 P进行对角置换。具体而言，令

$$P' = P + \Delta,$$

其中， $\Delta = -\left(1 - n^{-1/3}\right) \operatorname{diag}(P)$ 是一个对角扰动矩阵。根据韦尔不等式，我们有

$$|\lambda_i(P') - \lambda_i(P)| \leqslant \|\Delta\| = \max_i \left(1 - n^{-1/3}\right) |P_{ii}|. \tag{S.4}$$

在条件 $n^{1/3}p_{\max}/\mathsf{K}^2 \to \infty$ 且 $\min_{1\leqslant k\leqslant \mathsf{K}} |\lambda_k(P)| \geqslant cnp_{\max}/\mathsf{K}$成立时，对于某个常数$c > 0$, 我们可以推导出 $\min_{1\leqslant k\leqslant \mathsf{K}} |\lambda_k(P)| \geqslant cn^{2/3}$。因此,

$$\min_{1\leqslant k\leqslant \mathsf{K}} |\lambda_k(P)| \gg \|\Delta\|.$$

这意味着 $\min_{1\leqslant k\leqslant \mathsf{K}} |\lambda_k(P')| \gg \|\Delta\|$，且 P' 的任何特征值都不穿过零。因此，这种对角扰动对 P 的领先特征值只引入可忽略的变化，从而在 n 足够大时使 P' 的秩等于 P 的秩。

基于这一扰动，不难验证 $\mathbb{E}(A) = P' - \operatorname{diag}(P')$。~回顾（S.1）， $A = A - \mathbb{E}(A)$。我们可以将（S.2）改写为

$$A = P' + \widetilde{A} - \operatorname{diag}(P'), \tag{S.5}$$

~其中， A、 P'和 $A - \operatorname{diag}(P')$均为对称矩阵。在$\mathbf{H}_0$,下， P'的秩为 K_0。令 $P'_* = HP'$，其中 H是用于执行正交变换的正交矩阵

在 P'满足以下条件：前 K_0列 P'_*非零，其余 $n-\mathsf{K}_0$列$\tilde{}$ $\tilde{}$ P'_*为零。记 $A_* = HA$和 $A_* = H(A-\mathrm{diag}(P'))$。由式（S.5）可得

$$A_* = P'_* + \tilde{A}_*. \tag{S.6}$$

根据引理S.1， $A_*A_*^\top$、 $P'_*P'^\top_*$和 $\tilde{A}_*\tilde{A}_*^\top$的特征值分别与 $AA^\top$、$P'P'^\top$和 $(\tilde{A}-\mathrm{diag}(P'))(\tilde{A}-\mathrm{diag}(P'))^\top$的特征值相同。

将与非零特征值相关的K_0 列组成的矩阵 P'_* 记为$P'_{*,\ 1}\in\mathbb{R}^{n\times\mathsf{K}_0}$将$\mathsf{K}_0$ 列组成的矩阵 $\tilde{A}_*$ 记为$\tilde{A}_{*,\ 1}\in\mathbb{R}^{n\times\mathsf{K}_0}$ 将 n –K_0 列组成的矩阵 $\tilde{A}_*$ 记为$\tilde{A}_{*,\ 2}\in\mathbb{R}^{n\times(n-\mathsf{K}_0)}$。那么，根据（S.6），我们可以将 $A_*A_*^\top$ 分解为以下两个项之和：

$$A_*A_*^\top = (P'_{*,1}+\tilde{A}_{*,1})(P'_{*,1}+\tilde{A}_{*,1})^\top + \tilde{A}_{*,2}\tilde{A}_{*,2}^\top. \tag{S.7}$$

设$G^\top\Lambda_{P'_{*,1}\tilde{A}_{*,1}}G$ 是第一项$(P'_{*,1}+\tilde{A}_{*,1})(P'_{*,1}+\tilde{A}_{*,1})^\top$ 的谱分解，其中 G 是正交矩阵。由于$(P'_{*,1}+\tilde{A}_{*,1})(P'_{*,1}+\tilde{A}_{*,1})^\top$ 的秩为 K_0,，对角矩阵 $\Lambda_{P'_{*,1}\tilde{A}_{*,1}}$可选取为使前 K_0 个对角元素非零，其余 n –K_0 个对角元素为零，即

$$\Lambda_{P'_{*,1}\tilde{A}_{*,1}} = \mathrm{diag}\left\{\lambda_1\big((P'_{*,1}+\tilde{A}_{*,1})(P'_{*,1}+\tilde{A}_{*,1})^\top\big),\cdots,\lambda_{\mathsf{K}_0}\big((P'_{*,1}+\tilde{A}_{*,1})(P'_{*,1}+\tilde{A}_{*,1})^\top\big),0,\cdots,0\right\}.$$

$\breve{}$ $\tilde{}$此外，设 $A=GA_{*,2}$，并将由 K_0 行和最后 n –K_0 行组成的矩阵 $\breve{A}$ 分别记为 $\breve{A}_1\in\mathbb{R}^{\mathsf{K}_0\times(n-\mathsf{K}^0)}$ 和 $\breve{A}_2\in\mathbb{R}^{(n-\mathsf{K}_0)\times(n-\mathsf{K}^0)}$。那么，根据（S.7），可得

$$\frac{1}{np_{\max}}GA_*A_*^\top G^\top = \frac{1}{np_{\max}}\Lambda_{P'_{*,1}\tilde{A}_{*,1}} + \frac{1}{np_{\max}}\breve{A}\breve{A}^\top.$$

定义 $\tilde{d} = \frac{\lambda_{\mathsf{K}_0}\left((P'_{*,1}+\tilde{A}_{*,\ 1})(P'_{*\ 1}+\tilde{A}_{*\ 1})^\top\right)_{,,\circ}}{2np_{\max}}$ 根据引理 S.2，对于有限的 $1\leqslant s\leqslant\mathsf{K}_{\max}$ –K_0+2，我们有

$$\left|\lambda_{\mathsf{K}_0+s}\left(\frac{1}{np_{\max}}GA_*A_*^\top G^\top\right)-\frac{1}{np_{\max}}\lambda_s(\breve{A}_2\breve{A}_2^\top)\right| \leqslant 3\tilde{d}\frac{n^{-2}p_{\max}^{-2}\|\breve{A}_2\breve{A}_2^\top\|^2}{(\tilde{d}-n^{-1}p_{\max}^{-1}\|\breve{A}_2\breve{A}_2^\top\|)^2}. \tag{S.8}$$

接下来，我们对（S.8）的右端进行控制。令

$$G_I = \begin{pmatrix} 0_{\mathsf{K}_0\times\mathsf{K}_0} & 0_{\mathsf{K}_0\times(n-\mathsf{K}_0)} \\ 0_{(n-\mathsf{K}_0)\times\mathsf{K}_0} & I_{(n-\mathsf{K}_0)} \end{pmatrix},$$

˘其中， I_m 表示 m 维单位矩阵。于是， A_2 可重新表示为

$$\breve{A}_2 = \{G_I G H(\widetilde{A} - \mathrm{diag}(P'))G_I\}_{(n-\mathsf{K}_0)\times(n-\mathsf{K}_0)}. \tag{S.9}$$

在条件 $n^{1/3}p_{\max}/\mathsf{K}^2 \to \infty$下，我们有 $\lim_{n\to\infty} np_{\max}\log^{-1}(n) \to \infty$。因此，根据引理S.3，存在某个常数 $\psi_{\widetilde{A}} > 0$，使得以概率 $1 - \mathrm{o}(1)$

$$\|\widetilde{A}\| \equiv \widetilde{\lambda}_1(\widetilde{A}) \leqslant \psi_{\widetilde{A}}\sqrt{np_{\max}}. \tag{S.10}$$

该结果结合（S.9）以及 H 和 G 均为正交的事实，可推出

$$\begin{aligned}
\|\breve{A}_2\breve{A}_2^\top\| &\leqslant \|G_I\|\|(\widetilde{A} - \mathrm{diag}(P'))(\widetilde{A} - \mathrm{diag}(P'))^\top\| \\
&\leqslant \|\widetilde{A}\widetilde{A}^\top\| + 2\|\widetilde{A}\|\|\mathrm{diag}(P')\| + \|\mathrm{diag}^2(P')\| \\
&\leqslant \widetilde{\lambda}_1^2(\widetilde{A}) + 2\widetilde{\lambda}_1(\widetilde{A})p_{\max} + p_{\max}^2 \\
&= \mathrm{O}_{\mathbb{P}}(np_{\max}).
\end{aligned} \tag{S.11}$$

这意味着存在某个常数 $\psi_{\breve{A}_2} > 0$，使得以概率 $1 - \mathsf{o}(1)$，$\|\breve{A}_2\breve{A}_2^\top\| \leqslant \psi_{\breve{A}_2} np_{\max}$。

此外，令

$$c = \widetilde{d}/\psi_{\breve{A}_2}^2 - 2/\psi_{\breve{A}_2} + 1/\widetilde{d}.$$

经过直接计算，我们得到（S.8）右侧小于 $3c^{-1}$，其成立概率为 $1 - \mathrm{o}(1)$。接下来我们证明 $c^{-1} = \mathrm{o}_{\mathbb{P}}(n^{-2/3})$。采用与推导（S.11）时类似的技巧，对于某个常数$\psi_{\widetilde{A}_1} > 0$，我们可以得到其成立概率为 $1 - \mathrm{o}(1)$

$$\|\widetilde{A}_{*,1}\widetilde{A}_{*,1}^\top\| \leqslant \|(\widetilde{A} - \mathrm{diag}(P'))(\widetilde{A} - \mathrm{diag}(P'))^\top\| \leqslant \psi_{\widetilde{A}_1} np_{\max}. \tag{S.12}$$

注意， $P'_{ij} \leqslant p_{\max}$根据定义，我们有 $\widetilde{\lambda}_1^2(P') \leqslant \mathrm{tr}(P'^\top P') = \sum_{i,\ j} P'^2_{ij} \leqslant n^2 p_{\max}^2$，˜由此可得 $\lambda_1(P') \leqslant np_{\max}$。记 $\boldsymbol{\alpha}_{\mathsf{K}_0}$ 为对应于K_0 个最大特征值中第($P'_{*\text{个的单位特征向量}1} + \widetilde{A}_{*,\ 1})(P'_{*,\ 1} + \widetilde{A}_{*,\ 1})^\top$。结合上述结果、式（S.4）、式（S.10）、式（S.12）以及假设 $\min_{1\leqslant k\leqslant \mathsf{K}} |\lambda_k(P)| \geqslant cnp_{\max}/\mathsf{K}$，可知在

概率 $1-\mathsf{o}(1)$，我们有

$$
\begin{aligned}
&\|\lambda_{\mathsf{K}_0}\big((P'_{*,1}+\widetilde{A}_{*,1})(P'_{*,1}+\widetilde{A}_{*,1})^\top\big)\boldsymbol{\alpha}_{\mathsf{K}_0}\| \\
&\geqslant \|P'_{*,1}P'^\top_{*,1}\boldsymbol{\alpha}_{K_0}\| - \|P'_{*,1}\widetilde{A}^\top_{*,1}\|\|\boldsymbol{\alpha}_{K_0}\| - \|\widetilde{A}_{*,1}P'^\top_{*,1}\|\|\boldsymbol{\alpha}_{\mathsf{K}_0}\| - \|\widetilde{A}_{*,1}\widetilde{A}^\top_{*,1}\|\|\boldsymbol{\alpha}_{\mathsf{K}_0}\| \\
&\geqslant \|P'_{*,1}P'^\top_{*,1}\boldsymbol{\alpha}_{\mathsf{K}_0}\| - 2\|P'\|\|\widetilde{A}\| - 2\|P'\|\|\mathrm{diag}(P')\| - \psi_{\widetilde{A}_1}np_{\max} \\
&\geqslant \|P'_{*,1}P'^\top_{*,1}\boldsymbol{\alpha}_{\mathsf{K}_0}\| - 2\widetilde{\lambda}_1(P')\widetilde{\lambda}_1(\widetilde{A}) - 2\widetilde{\lambda}_1(P')\max_i P'_{ii} - \psi_{\widetilde{A}_1}np_{\max} \\
&\geqslant c^*_{\mathsf{K}_0}\widetilde{\lambda}^2_{\mathsf{K}_0}(P') - 2\psi_{\widetilde{A}}np_{\max}\sqrt{np_{\max}} - 2c^*_4np^2_{\max} - \psi_{\widetilde{A}_1}np_{\max} \\
&\geqslant c^*_{\mathsf{K}_0}n^2p^2_{\max}\mathsf{K}_0^{-2} - 2\psi_{\widetilde{A}}n^{3/2}p^{3/2}_{\max} - 2c^*_4np^2_{\max} - \psi_{\widetilde{A}_1}np_{\max} \\
&\geqslant \widetilde{c}^*_{\mathsf{K}_0}n^2p^2_{\max}\mathsf{K}_0^{-2},
\end{aligned}
\tag{S.13}
$$

其中 $\widetilde{c}^*_{\mathsf{K}_0}$、 $c^*_{\mathsf{K}_0}$、 c^*_4、 $\psi_{\widetilde{A}}$和 $\psi_{\widetilde{A}_1}$是一些正常数。因此，以概率1 $-\mathsf{o}(1)$,

$$
\widetilde{d} = \frac{\lambda_{\mathsf{K}_0}\big((P'_{*,1}+\widetilde{A}_{*,1})(P'_{*,1}+\widetilde{A}_{*,1})^\top\big)}{2np_{\max}} \geqslant \frac{\widetilde{c}^*_{\mathsf{K}_0}np_{\max}\mathsf{K}_0^{-2}}{2}.
$$

回顾$c = \widetilde{d}/\psi^2_{\check{A}_2} - 2/\psi_{\check{A}_2} + 1/\widetilde{d}$。由此可得 $c^{-1} = \mathrm{O}_{\mathbb{P}}\left(n^{-1}p^{-1}_{\max}\mathsf{K}_0^2\right)$。结合条件 $n^{1/3}p_{\max}/\mathsf{K}_0^2 \to \infty$，我们有 $c^{-1} = \mathrm{o}_{\mathbb{P}}(n^{-2/3})$。回顾（S.8）的右端小于 $3c^{-1}$。于是由（S.8）可得

$$
\frac{1}{np_{\max}}\lambda_{\mathsf{K}_0+s}(GA_*A^\top_*G^\top) = \frac{1}{np_{\max}}\lambda_s(\check{A}_2\check{A}^\top_2) + \mathrm{o}_{\mathbb{P}}(n^{-2/3}).
$$

此外，由于 G 和 H 是正交矩阵，由引理S.1，我们有

$$
\frac{1}{np_{\max}}\lambda_{\mathsf{K}_0+s}(AA^\top) = \frac{1}{np_{\max}}\lambda_s(\check{A}_2\check{A}^\top_2) + \mathrm{o}_{\mathbb{P}}\left(n^{-2/3}\right), \tag{S.14}
$$

至此完成了证明的第一步。

步骤 II. 由（S.9）回顾可知 $A_2 = \{G_IGH(A-\mathbf{diag}(P'))G_I\}_{(n-\mathsf{K}_0)\times(n-\mathsf{K}_0)}$。为验证本步骤的结果，我们引入一个辅助矩阵

$$
\widetilde{A}_2 = \{G_I(\widetilde{A} - \mathrm{diag}(P'))G_I\}_{(n-\mathsf{K}_0)\times(n-\mathsf{K}_0)}. \tag{S.15}
$$

由于 H和 G正交，矩阵 $GH(\widetilde{A}-\mathbf{diag}(P'))G_I(\widetilde{A}-\mathbf{diag}(P'))H^\top G^\top$和 $\sim$ $\sim(A-\mathbf{diag}(P'))G_I(A-\mathbf{diag}(P'))$具有相同的特征值 $\lambda_{\widetilde{A},\,1}$，$\cdots$，$\lambda_{\widetilde{A},\,n}$。因此，它们对应的谱分解为

$$
GH(\widetilde{A} - \mathrm{diag}(P'))G_I(\widetilde{A} - \mathrm{diag}(P'))H^\top G^\top = S_1^\top\Lambda_{\widetilde{A}}S_1,
$$

$$(\widetilde{A} - \mathrm{diag}(P'))G_I(\widetilde{A} - \mathrm{diag}(P')) = S_2^\top \Lambda_{\widetilde{A}} S_2,$$

其中 $\Lambda_{\widetilde{A}} = \mathbf{diag}\{\lambda_{\widetilde{A},\,1},\ \cdots,\ \lambda_{\widetilde{A},\,n}\}$，而 S_1 和 S_2 是与这两个分解相关联的正交矩阵。由此可得

$$\breve{A}_2\breve{A}_2^\top = \{G_I S_1^\top \Lambda_{\widetilde{A}} S_1 G_I\}_{(n-\mathsf{K}_0)\times(n-\mathsf{K}_0)} \ \text{ and } \ \widetilde{A}_2\widetilde{A}_2^\top = \{G_I S_2^\top \Lambda_{\widetilde{A}} S_2 G_I\}_{(n-\mathsf{K}_0)\times(n-\mathsf{K}_0)}. \quad \text{(S.16)}$$

利用 S_1 与 S_2 正交这一事实，方程（S.16）表明，对于 $1 \leqslant s \leqslant \mathsf{K}_{\max} - \mathsf{K}_0 + 2$，存在一些常数 $\xi_s > 0$ 使得

$$\frac{1}{np_{\max}}\lambda_s(\breve{A}_2\breve{A}_2^\top) = \frac{\xi_s^2}{np_{\max}}\lambda_s(\widetilde{A}_2\widetilde{A}_2^\top).$$

结合步骤 I 的结果（S.14），可得

$$\frac{1}{np_{\max}}\lambda_{\mathsf{K}_0+s}(AA^\top) = \frac{\xi_s^2}{np_{\max}}\lambda_s(\widetilde{A}_2\widetilde{A}_2^\top) + \mathrm{o}_{\mathbb{P}}\left(n^{-2/3}\right).$$

~回顾一下，对于任意方阵 B，我们将 $\lambda_s(B)$ 记为 B~按绝对值从大到小排列的第 s 个特征值。由于 A 和 A_2 均为对称矩阵，其奇异值与特征值的绝对值相同。因此，我们有

$$\frac{1}{np_{\max}}\widetilde{\lambda}^2_{\mathsf{K}_0+s}(A) = \frac{\xi_s^2}{np_{\max}}\widetilde{\lambda}^2_s(\widetilde{A}_2) + \mathrm{o}_{\mathbb{P}}\left(n^{-2/3}\right).$$

在以下条件下：$n^{1/3}p_{\max}/\mathsf{K}^2 \to \infty$，$\mathbf{min}_{1\leqslant k\leqslant \mathsf{K}}\,|\lambda_k(P)| \geqslant cnp_{\max}/\mathsf{K}$ 对某个常数 $c > 0$，假设 $\mathbf{max}_{ij}\,P_{ij} \asymp \min_{ij} P_{ij}$，且 $\mathsf{K}_{\max} - \mathsf{K}_0$ 有限，引理 S.7 表明 A 至少有 $\mathsf{K}_{\max} + 2$ 个远离零的特征值，因此可得

$$\frac{1}{\sqrt{np_{\max}}}\widetilde{\lambda}_{\mathsf{K}_0+s}(A) = \frac{\xi_s}{\sqrt{np_{\max}}}\widetilde{\lambda}_s(\widetilde{A}_2) + \mathrm{o}_{\mathbb{P}}\left(n^{-2/3}\right).$$

这意味着，存在某个 $\psi \in \{1,\ -1\}$ 使得

$$\frac{1}{\sqrt{np_{\max}}}\lambda_{\mathsf{K}_0+s}(A) = \psi\frac{\xi_s}{\sqrt{np_{\max}}}\lambda_s(\widetilde{A}_2) + \mathrm{o}_{\mathbb{P}}\left(n^{-2/3}\right).$$

由于选择 ψ 会导出相同的后续渐近论证，我们取 $\psi = 1$

不失一般性，可得

$$\frac{1}{\sqrt{np_{\max}}}\lambda_{\mathsf{K}_0+s}(A) = \frac{\xi_s}{\sqrt{np_{\max}}}\lambda_s(\widetilde{A}_2) + \mathrm{o}_{\mathbb{P}}\left(n^{-2/3}\right). \tag{S.17}$$

回顾（S.15）可知，$A_2 = \{G_I(A - \mathsf{diag}(P'))G_I\}_{(n-\mathsf{K}_0)\times(n-\mathsf{K}_0)}$。根据$P'$的定义，我们有

$$G_I\widetilde{A}G_I - \max_i P'_{ii}G_I \leqslant G_I\widetilde{A}G_I - G_I\mathrm{diag}(P')G_I \leqslant G_I\widetilde{A}G_I - \min_i P'_{ii}G_I. \tag{S.18}$$

随后，我们定义

$$\widetilde{A}_{G_I} = \{G_I\widetilde{A}G_I\}_{(n-\mathsf{K}_0)\times(n-\mathsf{K}_0)}. \tag{S.19}$$

利用（S.18），并注意到所有项 P'_{ii} 都是同阶的，我们可以得到

$$\lambda_s(\widetilde{A}_2) = \lambda_s(\widetilde{A}_{G_I}) - \xi_s^* \max_i P'_{ii}, \tag{S.20}$$

其中，$\xi_s^* > 0$ 是 $1 \leqslant s \leqslant \mathsf{K}_{\max} - \mathsf{K}_0 + 2$的一些常数。由于 $P'_{ii} = \mathrm{O}(n^{-1/3})$ 根据定义，在条件 $n^{1/3}p_{\max}/\mathsf{K}^2 \to \infty$下，我们有$\max_i P'_{ii}/\sqrt{np_{\max}} = \mathrm{o}\left(n^{-2/3}\right)$。结合（S.17）和（S.20），可得

$$\frac{1}{\sqrt{np_{\max}}}\lambda_{\mathsf{K}_0+s}(A) = \frac{\xi_s}{\sqrt{np_{\max}}}\lambda_s(\widetilde{A}_{G_I}) + \mathrm{o}_{\mathbb{P}}\left(n^{-2/3}\right), \tag{S.21}$$

由此完成证明的第二步。

步骤 III。定义一个对称矩阵 $\Omega = (\Omega_{ij}) \in \mathbb{R}^{n\times n}$ 其中 $\Omega_{ij} = \{P_{ij}(1 - P_{ij})\}^{-1/2}$ 对于$1 \leqslant i, j \leqslant n$。此外，定义

$$\widetilde{\Omega}_{G_I} = (n - \mathsf{K}_0 - 1)^{-1/2}\{G_I\Omega G_I\}_{(n-\mathsf{K}_0)\times(n-\mathsf{K}_0)}. \tag{S.22}$$

由于所有元素 P_{ij}均为同阶量，这意味着 $\widetilde{\Omega}_{G_I,\,ij} \asymp \frac{1}{\sqrt{np_{\max}}}$对于$1 \leqslant i, j \leqslant n - \mathsf{K}_0$。结合（S.21），可得

$$\widetilde{\Omega}_{G_I,ij}\lambda_{\mathsf{K}_0+s}(A) = \xi_s\widetilde{\Omega}_{G_I,ij}\lambda_s(\widetilde{A}_{G_I}) + \mathrm{o}_{\mathbb{P}}\left(n^{-2/3}\right). \tag{S.23}$$

令 $\circ$ d 记哈达玛（逐项）积。应用引理 S.4，可得

$$\lambda_s(\widetilde{\Omega}_{G_I} \circ \widetilde{A}_{G_I}\widetilde{A}_{G_I}^\top \circ \widetilde{\Omega}_{G_I}) \geqslant \min_i \widetilde{\Omega}_{G_I,ii}^2\lambda_s(\widetilde{A}_{G_I}\widetilde{A}_{G_I}^\top),$$

$$\lambda_s(\widetilde{\Omega}_{G_I} \circ \widetilde{A}_{G_I}\widetilde{A}_{G_I}^\top \circ \widetilde{\Omega}_{G_I}) \leqslant \max_i \widetilde{\Omega}^2_{G_I,ii}\lambda_s(\widetilde{A}_{G_I}\widetilde{A}_{G_I}^\top).$$

根据定义，Ω_{G_I} 的所有元素均为同阶，因此，对于任意固定的配对 $(i^*,\ j^*)$，都存在一些常数 $0 < m_1 \leqslant m_2 < \infty$，它们与 s 和 n 无关，使得

$$m_1^2\xi_s^2\widetilde{\Omega}^2_{G_I,i^*j^*}\lambda_s(\widetilde{A}_{G_I}\widetilde{A}_{G_I}^\top) \leqslant \lambda_s\Big(\widetilde{\Omega}_{G_I} \circ \widetilde{A}_{G_I}\widetilde{A}_{G_I}^\top \circ \widetilde{\Omega}_{G_I}\Big) \leqslant m_2^2\xi_s^2\widetilde{\Omega}^2_{G_I,i^*j^*}\lambda_s(\widetilde{A}_{G_I}\widetilde{A}_{G_I}^\top),$$

其中 $\xi_s > 0$ 是（S.17）中的某个常数，而常数 m_1 和 m_2 确保$\widetilde{\Omega}$ ξ_s的$_{G_I}$对角元 $1 \leqslant i$在 $j \leqslant n-\mathsf{K}_0$且有限的 $1 \leqslant s \leqslant \mathsf{K}_{\max}-\mathsf{K}_0+2$下是一致有界的。通过类似于推导（S.17）时所用的论证，可得

$$m_1\xi_s\widetilde{\Omega}_{G_I,i^*j^*}\lambda_s(\widetilde{A}_{G_I}) \leqslant \lambda_s\Big(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}\Big) \leqslant m_2\xi_s\widetilde{\Omega}_{G_I,i^*j^*}\lambda_s(\widetilde{A}_{G_I}).$$

因此，对于给定的 $(i^*,\ j^*)$，存在一个与 s 和 n 无关的常数 $m_1 \leqslant c_{i^*j^*} \leqslant m_2$，使得

$$c_{i^*j^*}\xi_s\widetilde{\Omega}_{G_I,i^*j^*}\lambda_s(\widetilde{A}_{G_I}) = \lambda_s\Big(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}\Big).$$

结合（S.23）即可得到

$$c_{i^*j^*}\widetilde{\Omega}_{G_I,i^*j^*}\lambda_{\mathsf{K}_0+s}(A) = \lambda_s\Big(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}\Big) + \mathrm{o}_{\mathbb{P}}\left(n^{-2/3}\right). \tag{S.24}$$

根据（S.1）、（S.19）和（S.22），不难验证 $A_{G_I} \circ \Omega_{G_I}$ 是$\widetilde{\ }$的方差归一化版本 A_{G_I}。至此，证明的第三步完成。

第四步。记一个GOE矩阵为 $W \in \mathbb{R}^{n\times n}$，其中 $W_{ij} = W_{ji}$，非对角元W_{ij}相互独立且服从 $\mathcal{N}(0,\ n^{-1})$，对角元 W_{ii}服从 $\mathcal{N}(0,\ 2/n)$。众所周知，$n^{2/3}(\lambda_1(W) - 2)$依分布收敛于I型Tracy–Widom分布（Tracy 和 Widom，1994，1996）。此外，对于任意固定的 $s > 0$，$\{n^{2/3}(\lambda_k(W) - 2)\}_{1\leqslant k\leqslant s}$, 的联合分布具有由艾里核描述的联合密度，这通常也称为Tracy–Widom 定律（Forrester，1993）。

考虑对称矩阵 $A_{G_I} \circ \Omega_{G_I}$。在随机图模型（1.1）下，对于$1 \leqslant i + j \leqslant n - \mathsf{K}_0$, ，我们可以得到

$$\mathbb{E}(\widetilde{A}_{G_I,ij} \circ \widetilde{\Omega}_{G_I,ij}) = 0 \text{ and } \textstyle\sum_i \mathbb{E}|\widetilde{A}_{G_I,ij} \circ \widetilde{\Omega}_{G_I,ij}|^2 = 1.$$

经过一些直接的计算，我们还可以验证，对于任意 $m \geqslant 2$,

$$\mathbb{E}|\widetilde{A}_{G_I,ij} \circ \widetilde{\Omega}_{G_I,ij}|^m \leqslant \tfrac{(cm)^{cm}}{nq^{m-2}},$$

其中 $c > 0$ 是某个常数，且 $q = n^{-1/2}(\sum_{i,j} P_{ij})^{1/2}$。在这些矩条件下，并结合如下假设： $n^{1/3}p_{\max}/\mathsf{K}^2 \to \infty$ 以及 $\max_{i,\ j} P_{ij} \asymp \max_{i,\ j} P_{ij}$，引理 S.5 表明，对于任意 $x_1 \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\big(n^{2/3}(\lambda_1(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \mathsf{L}) \leqslant x_1\big) = \lim_{n\to\infty} \mathbb{P}\big(n^{2/3}(\lambda_1(W) - 2) \leqslant x_1\big),$$

其中 L 是一个满足以下条件的确定性标量：$\mathsf{L} - 2 = \mathrm{o}(1)$ 如引理 S.5中所述。此外，根据引理 S.6，这一收敛可以加强为联合收敛。具体而言，对于有限的 $1 \leqslant s \leqslant \mathsf{K}_{\max} - \mathsf{K}_0 + 2$ 以及任意 $x_s \in \mathbb{R}$,

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{P}\big(n^{2/3}(\lambda_1(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \mathsf{L}) \leqslant x_1, \cdots, n^{2/3}(\lambda_s(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \mathsf{L}) \leqslant x_s\big) \\ &= \lim_{n\to\infty} \mathbb{P}\big(n^{2/3}(\lambda_1(W) - 2) \leqslant x_1, \cdots, n^{2/3}(\lambda_s(W) - 2) \leqslant x_s\big). \end{aligned} \tag{S.25}$$

因此， $n^{2/3}(\lambda_s(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \mathsf{L})$ 的联合分布具有由艾里核给出的 Tracy–Widom 定律所描述的联合密度。这就完成了证明的第四步。**步骤 V**。根据（S.24），我们有

$$\begin{aligned} \mathbb{T} &= \frac{\lambda_{\mathsf{K}_0+1}(A) - \lambda_{\mathsf{K}_{\max}+1}(A)}{\lambda_{\mathsf{K}_{\max}+1}(A) - \lambda_{\mathsf{K}_{\max}+2}(A)} \\ &= \frac{c_{i^*j^*}\widetilde{\Omega}_{G_I,i^*j^*}\lambda_{\mathsf{K}_0+1}(A) - c_{i^*j^*}\widetilde{\Omega}_{G_I,i^*j^*}\lambda_{\mathsf{K}_{\max}+1}(A)}{c_{i^*j^*}\widetilde{\Omega}_{G_I,i^*j^*}\lambda_{\mathsf{K}_{\max}+1}(A) - c_{i^*j^*}\widetilde{\Omega}_{G_I,i^*j^*}\lambda_{\mathsf{K}_{\max}+2}(A)} \\ &= \frac{\lambda_1(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+1}(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) + \mathrm{o}_{\mathbb{P}}\left(n^{-2/3}\right)}{\lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+1}(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+2}(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) + \mathrm{o}_{\mathbb{P}}\left(n^{-2/3}\right)} \\ &= \frac{n^{2/3}\{\lambda_1(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \mathsf{L}\} - n^{2/3}\{\lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+1}(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \mathsf{L}\} + \mathrm{o}_{\mathbb{P}}(1)}{n^{2/3}\{\lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+1}(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \mathsf{L}\} - n^{2/3}\{\lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+2}(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \mathsf{L}\} + \mathrm{o}_{\mathbb{P}}(1)}. \end{aligned}$$

回顾一下，

$$\begin{aligned} \mathbb{T}_W &= \frac{\lambda_1(W) - \lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+1}(W)}{\lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+1}(W) - \lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+2}(W)} \\ &= \frac{n^{2/3}\{\lambda_1(W) - 2\} - n^{2/3}\{\lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+1}(W) - 2\}}{n^{2/3}\{\lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+1}(W) - 2\} - n^{2/3}\{\lambda_{\mathsf{K}_{\max}-\mathsf{K}_0+2}(W) - 2\}}. \end{aligned}$$

因此，利用（S.25），我们得到 $\lim_{n\to\infty} F_{\mathbb{T}}(x) = \lim_{n\to\infty} F_{\mathbb{T}_W}(x)$ 对任意 $x \in \mathbb{R}$，其中 $F_{\mathbb{T}}(x)$ 和 $F_{\mathbb{T}_W}(x)$ 分别是 $\mathbb{T}$ 和 $\mathbb{T}_W$的累积分布函数。至此，定理1的证明完成。

□

S.1.2 定理2的证明

在备择假设 $\mathbf{H}_1$： $\mathsf{K}_0 < \mathsf{K} \leqslant \mathsf{K}_{\max}$下，我们有 $\lambda_{\mathsf{K}_0+1}(A) \geqslant \lambda_{\mathsf{K}}(A)$。结合（S.24）中的论证，可得

$$\begin{aligned}
\mathbb{T} &= \frac{\lambda_{\mathsf{K}_0+1}(A) - \lambda_{\mathsf{K}_{\max}+1}(A)}{\lambda_{\mathsf{K}_{\max}+1}(A) - \lambda_{\mathsf{K}_{\max}+2}(A)} \\
&\geqslant \frac{\lambda_{\mathsf{K}}(A) - \lambda_{\mathsf{K}_{\max}+1}(A)}{\lambda_{\mathsf{K}_{\max}+1}(A) - \lambda_{\mathsf{K}_{\max}+2}(A)} \\
&= \frac{c_{i^*j^*}\widetilde{\Omega}_{G_I,i^*j^*}\lambda_{\mathsf{K}}(A) - c_{i^*j^*}\widetilde{\Omega}_{G_I,i^*j^*}\lambda_{\mathsf{K}_{\max}+1}(A)}{c_{i^*j^*}\widetilde{\Omega}_{G_I,i^*j^*}\lambda_{\mathsf{K}_{\max}+1}(A) - c_{i^*j^*}\widetilde{\Omega}_{G_I,i^*j^*}\lambda_{\mathsf{K}_{\max}+2}(A)} \\
&= \frac{c_{i^*j^*}\widetilde{\Omega}_{G_I,i^*j^*}\lambda_{\mathsf{K}}(A) - \lambda_{\mathsf{K}_{\max}-\mathsf{K}+1}(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) + \mathrm{o}_{\mathbb{P}}\left(n^{-2/3}\right)}{\lambda_{\mathsf{K}_{\max}-\mathsf{K}+1}(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \lambda_{\mathsf{K}_{\max}-\mathsf{K}+2}(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) + \mathrm{o}_{\mathbb{P}}\left(n^{-2/3}\right)}.
\end{aligned} \tag{S.26}$$

在条件 $n^{1/3}p_{\max}/\mathsf{K}^2 \to \infty$ 并应用引理 S.5 的结果 (i)，可得

$$\begin{aligned}
n^{2/3}\left\{\lambda_{\mathsf{K}_{\max}-\mathsf{K}+1}(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \mathsf{L}\right\} &= \mathrm{O}_{\mathbb{P}}(n^{\phi}) \quad \text{and} \\
n^{2/3}\left\{\lambda_{\mathsf{K}_{\max}-\mathsf{K}+2}(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \mathsf{L}\right\} &= \mathrm{O}_{\mathbb{P}}(n^{\phi}),
\end{aligned} \tag{S.27}$$

对于任意小的 $\phi > 0$，其中 $\mathsf{L} \asymp 1$是引理S.5中指定的确定性标量。结合 (S.26) 和 (S.27)，可得

$$\begin{aligned}
\mathbb{T} &\geqslant \frac{\left|c_{i^*j^*}\widetilde{\Omega}_{G_I,i^*j^*}\lambda_{\mathsf{K}}(A) - \mathsf{L}\right| - \left|\lambda_{\mathsf{K}_{\max}-\mathsf{K}+1}(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \mathsf{L}\right| + \mathrm{o}_{\mathbb{P}}\left(n^{-2/3}\right)}{\left\{\lambda_{\mathsf{K}_{\max}-\mathsf{K}+1}(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \mathsf{L}\right\} - \left\{\lambda_{\mathsf{K}_{\max}-\mathsf{K}+2}(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I} - \mathsf{L}\right\} + \mathrm{o}_{\mathbb{P}}\left(n^{-2/3}\right)} \\
&= \frac{n^{2/3}\left|c_{i^*j^*}\widetilde{\Omega}_{G_I,i^*j^*}\lambda_{\mathsf{K}}(A) - \mathsf{L}\right| - n^{2/3}\left|\lambda_{\mathsf{K}_{\max}-\mathsf{K}+1}(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \mathsf{L}\right| + \mathrm{o}_{\mathbb{P}}(1)}{n^{2/3}\left\{\lambda_{\mathsf{K}_{\max}-\mathsf{K}+1}(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I}) - \mathsf{L}\right\} - n^{2/3}\left\{\lambda_{\mathsf{K}_{\max}-\mathsf{K}+2}(\widetilde{A}_{G_I} \circ \widetilde{\Omega}_{G_I} - \mathsf{L}\right\} + \mathrm{o}_{\mathbb{P}}(1)}, \\
&= \frac{n^{2/3}\left|c_{i^*j^*}\widetilde{\Omega}_{G_I,i^*j^*}\lambda_{\mathsf{K}}(A) - \mathsf{L}\right| - \mathrm{O}_{\mathbb{P}}(n^{\phi}) + \mathrm{o}_{\mathbb{P}}(1)}{\mathrm{O}_{\mathbb{P}}(n^{\phi}) + \mathrm{o}_{\mathbb{P}}(1)}.
\end{aligned} \tag{S.28}$$

根据（S.7）、（S.8）和（S.11），K 个最大特征值 $\frac{1}{np_{\max}}AA^\top$ 渐近地由以下各项的特征值所支配：$\frac{1}{np_{\max}}(P'_{*,\,1} + \widetilde{A}_{*,\,1})(P'_{*,\,1} + \widetilde{A}_{*,\,1})^\top$。根据（S.13），在条件 $n^{1/3}p_{\max}/\mathsf{K}^2 \to \infty$ 下，以概率 $1 - \mathrm{o}(1)$，

$$\frac{1}{np_{\max}}\lambda_{\mathsf{K}}\left((P'_{*,1} + \widetilde{A}_{*,1})(P'_{*,1} + \widetilde{A}_{*,1})^\top\right) \geqslant \tilde{c}^*_{\mathsf{K}} np_{\max}/\mathsf{K}^2 \gg n^{2/3},$$

其中 $\tilde{c}^*_{\mathsf{K}} > 0$ 是一个常数。因此，我们在（ $\phi > 0$ 中选择S.27），使得以概率 $1 - \mathrm{o}(1)$,

$$\frac{1}{np_{\max}}\lambda_{\mathsf{K}}(AA^\top) \gg n^{2\phi}.$$

结合 $c_{i^*j^*}\tilde{\Omega}_{G_I,\ i^*j^*} \asymp \frac{1}{\sqrt{np_{\max}}}$ 这一事实，~~我们有~~，以概率 $1 - \mathsf{o}(1)$,

$$c_{i^*j^*}\widetilde{\Omega}_{G_I,i^*j^*}\widetilde{\lambda}_{\mathsf{K}}(A) \gg n^{\phi}.$$

因此，以概率 $1 - \mathsf{o}(1)$,

$$\left|c_{i^*j^*}\widetilde{\Omega}_{G_I,i^*j^*}\lambda_{\mathsf{K}}(A) - \mathsf{L}\right| \gg n^{\phi}. \tag{S.29}$$

结合 (S.28) 和 (S.29)，对于任意大的常数 $\mathsf{C} > 0$，我们有 $\mathbb{P}(\mathbb{T} > \mathsf{C}n^{2/3}) = 1 - \mathrm{o}(1)$ 当 $n \to \infty$ 时，证明得证。

S.1.3 推论1的证明

考虑估计量 $\widehat{\mathsf{K}} = \min\{\mathsf{K}_0 \in \{1, \ldots, \mathsf{K}_{\max}\} : \mathbb{T} \leqslant c_\alpha\}$, 其中 c_α 是 $\mathbb{T}_W$ 的上α分位数。

对于 $\mathsf{K}_0 < \mathsf{K}$，定理 2 意味着 $\mathbb{P}(\mathbb{T} > c_\alpha) = 1 - \mathrm{o}(1)$ 当 $n \to \infty$ 时。因此，所有 $\mathsf{K}_0 < \mathsf{K}$ 都以趋于 1 的概率被拒绝。

对于 $\mathsf{K}_0 \geqslant \mathsf{K}$，定理 1 表明 $\mathbb{T}$ 依分布收敛到 $\mathbb{T}_W$，因此 $\mathbb{P}(\mathbb{T} \leqslant c_\alpha) = 1 - \alpha - \mathrm{o}(1)$ 当 $n \to \infty$ 时。因此，所有 $\mathsf{K}_0 \geqslant \mathsf{K}$ 被拒绝的概率趋于 $1 - \alpha$。

因此，由于 K 被定义为最小的未被拒绝的 K_0，结合上述两种情况可得

$$\mathbb{P}(\widehat{\mathsf{K}} = \mathsf{K}) = 1 - \alpha - \mathrm{o}(1) \quad \text{as } n \to \infty,$$

由此完成证明。

S.1.4 推论 2 的证明

考虑估计量 $\widehat{\mathsf{K}}^* = \min\{\mathsf{K}_0 \in \{1, \ldots\ldots, \mathsf{K}_{\max}\} : \mathbb{T} \leqslant t_n\}$, 其中 $t_n \asymp n^\epsilon$ 对于某个$\epsilon \in (0, 5/6)$。我们通过考虑两种情况来证明该推论，即低估和高估。

(i) 低估。 对任意 $\mathsf{K}_0 < \mathsf{K}$，不失一般性，我们假设

$\lambda_{\mathsf{K}_0+1}(A) \geqslant \lambda_{\mathsf{K}}(A) \geqslant \lambda_{\mathsf{K}_{\max}+1}(A) \geqslant \lambda_{\mathsf{K}_{\max}+2}(A) > 0$，因为另一种情形 $\lambda_{\mathsf{K}_{\max}+2}(A) \leqslant \lambda_{\mathsf{K}_{\max}+1}(A) \leqslant \lambda_{\mathsf{K}}(A) \leqslant \lambda_{\mathsf{K}_0+1}(A) \leqslant 0$ 会得出相同的结论。

根据（S.24）和（S.27），对于确定性标量 $c_{i^*j^*}\widetilde{\Omega}_{G_I,\, i^*j^*}$ 以及任意小的 $\phi > 0$，有

$$c_{i^*j^*}\widetilde{\Omega}_{G_I,i^*j^*}\lambda_{\mathsf{K}_{\max}+1}(A) = \mathrm{O}_{\mathbb{P}}(1),$$

$$c_{i^*j^*}\widetilde{\Omega}_{G_I,i^*j^*}\big\{\lambda_{\mathsf{K}_{\max}+1}(A) - \lambda_{\mathsf{K}_{\max}+2}(A)\big\} = \mathrm{O}_{\mathbb{P}}(n^{-2/3+\phi}).$$

根据定义 $c_{i^*j^*}\widetilde{\Omega}_{G_I,\, i^*j^*} \asymp n^{-1/2}p_{\max}^{-1/2}$，可得

$$\lambda_{\mathsf{K}_{\max}+1}(A) = \mathrm{O}_{\mathbb{P}}(n^{1/2}p_{\max}^{1/2}),$$

$$\lambda_{\mathsf{K}_{\max}+1}(A) - \lambda_{\mathsf{K}_{\max}+2}(A) = \mathrm{O}_{\mathbb{P}}(n^{-1/6+\phi}p_{\max}^{1/2}). \tag{S.30}$$

此外，根据上述讨论（S.29），$\frac{1}{np_{\max}}\lambda_{\mathsf{K}}(AA^\top) \geqslant \tilde{c}_{\mathsf{K}}^* np_{\max}/\mathsf{K}^2$以概率 $1-\mathrm{o}(1)$成立，其中 $\tilde{c}_{\mathsf{K}}^* > 0$ 是某个常数。结合 $n^{1/3}p_{\max}/\mathsf{K}^2 \to \infty$这一假设，可推出 $\widetilde{\lambda}_{\mathsf{K}}(A) \gg n^{2/3}$ 以概率$1-\mathrm{o}(1)$成立。因此，在假设 $\lambda_{\mathsf{K}}(A) > 0$下，以概率 $1-\mathrm{o}(1)$，有

$$\lambda_{\mathsf{K}}(A) \gg n^{2/3}.$$

结合（S.30）和 $\lambda_{\mathsf{K}_0+1}(A) \geqslant \lambda_{\mathsf{K}}(A)$，根据 $n \to \infty$，可得

$$\begin{aligned}
\mathbb{P}(\mathbb{T} \leqslant t_n) &= \mathbb{P}\Big(\frac{\lambda_{\mathsf{K}_0+1}(A) - \lambda_{\mathsf{K}_{\max}+1}(A)}{\lambda_{\mathsf{K}_{\max}+1}(A) - \lambda_{\mathsf{K}_{\max}+2}(A)} \leqslant t_n\Big) \\
&= \mathbb{P}\Big[\lambda_{\mathsf{K}_0+1}(A) \leqslant t_n\big\{\lambda_{\mathsf{K}_{\max}+1}(A) - \lambda_{\mathsf{K}_{\max}+2}(A)\big\} + \lambda_{\mathsf{K}_{\max}+1}(A)\Big] \\
&= \mathbb{P}\Big(\lambda_{\mathsf{K}_0+1}(A) \leqslant \mathsf{C}_1 n^{-1/6+\phi}p_{\max}^{1/2}t_n + \mathsf{C}_2 n^{1/2}p_{\max}^{1/2}\Big) \\
&\leqslant \mathbb{P}\Big(\lambda_{\mathsf{K}}(A) \leqslant \mathsf{C}_1 n^{-1/6+\phi}p_{\max}^{1/2}t_n + \mathsf{C}_2 n^{1/2}p_{\max}^{1/2}\Big) \\
&= \mathrm{o}(1),
\end{aligned}$$

其中 C_1,$\mathsf{C}_2 > 0$ 是一些常数。最后一个等式来自以下事实：$\lambda_{\mathsf{K}}(A) \gg n^{2/3}$ 以概率 $1-\mathsf{o}(1)$成立，结合假设 $t_n \ll n^{5/6}$。后者意味着存在 $\phi > 0$ 使得 $t_n \leqslant n^{5/6-\phi}$。因此，$\mathsf{C}_1 n^{-1/6+\phi}p_{\max}^{1/2}t_n + \mathsf{C}_2 n^{1/2}p_{\max}^{1/2} = \mathsf{O}(n^{2/3})$。于是，对于预先给定的 $\mathsf{K}_{\max}$，我们有

$$\mathbb{P}(\widehat{\mathsf{K}}^* < \mathsf{K}) \leqslant \sum_{\mathsf{K}_0=1}^{\mathsf{K}-1} \mathbb{P}(\mathbb{T} \leqslant t_n) \leqslant \mathsf{K}_{\max}\mathbb{P}(\mathbb{T} \leqslant t_n) = \mathrm{o}(1).$$

(ii) 高估。 在 $\mathbf{H}_0$ 下：对于预先给定的 $\mathsf{K}_{\max}$，$\mathsf{K} = \mathsf{K}_0$，，我们有

$$\begin{aligned}\mathbb{P}(\mathbb{T} > t_n) &= \mathbb{P}\Big(\frac{\lambda_{\mathsf{K}+1}(A) - \lambda_{\mathsf{K}_{\max}+1}(A)}{\lambda_{\mathsf{K}_{\max}+1}(A) - \lambda_{\mathsf{K}_{\max}+2}(A)} > t_n\Big) \\ &= \mathbb{P}\Big[\lambda_{\mathsf{K}+1}(A) > t_n\big\{\lambda_{\mathsf{K}_{\max}+1}(A) - \lambda_{\mathsf{K}_{\max}+2}(A)\big\} + \lambda_{\mathsf{K}_{\max}+1}(A)\Big] \\ &= \mathbb{P}\Big(\lambda_{\mathsf{K}+1}(A) > \mathsf{C}_1 n^{-1/6+\phi} p_{\max}^{1/2} t_n + \mathsf{C}_2 n^{1/2} p_{\max}^{1/2}\Big) \\ &= \mathrm{o}(1).\end{aligned}$$

最后一个等式源于如下事实：类似于（S.30）的推导，$\lambda_{\mathsf{K}+1}(A) = \mathrm{O}_{\mathbb{P}}(n^{1/2} p_{\max}^{1/2})$, 连同条件 $t_n \asymp n^{\epsilon}$ 对某个 $\epsilon \in (2/3 - \phi,\ 5/6)$。因此，当 $n \to \infty$时，我们有

$$\mathbb{P}(\widehat{\mathsf{K}}^* > \mathsf{K}) \leqslant \mathbb{P}\Big(\mathbb{T} > t_n\Big) = \mathrm{o}(1).$$

(i) 和 (ii) 的结果表明，当 $n \to \infty$时，$\mathbb{P}\left(\widehat{\mathsf{K}}^* = \mathsf{K}\right) = 1 - \mathsf{o}(1)$，从而完成了证明。

S.1.5 辅助引理

在本节中，我们给出若干引理，这些引理将用于主要结果的证明。

引理 S.1.对任意矩阵 $B \in \mathbb{R}^{n\times n}$，其奇异值在正交变换下保持不变，该正交变换记为 B_*。具体而言，$B_* B_*^{\top}$ 的特征值与 $BB^{\top}$的特征值相同。

证明。 将 $BB^{\top}$ 的谱分解记为

$$BB^{\top} = U \Lambda_B U^{\top},$$

其中 U 是正交矩阵，Λ_B 是对角矩阵，其对角元由 $BB^{\top}$ 的特征值组成。对于 B 的任意正交变换，记为 B_*，都存在一个正交矩阵 S，使得 $B_* = SB$。因此，我们有

$$B_* B_*^{\top} = (SB)(SB)^{\top} = SBB^{\top}S^{\top} = S(U\Lambda_B U^{\top})S^{\top} = (SU)\Lambda_B (SU)^{\top}.$$

其中矩阵 x SU 也正交。因此，$B_* B_*^{\top}$ 的特征值与

$BB^\top$的特征值相同，即为 Λ_B，至此证明完成。

□

引理 S.2. 设 $D^{(1)} \in \mathbb{R}^{n\times n}$ 为一个对称非负定矩阵，并令 $D = \boldsymbol{diag}(d_1, \cdots, d_{\mathsf{K}}, 0, \cdots, 0) \in \mathbb{R}^{n\times n}$，其中 $d_1 \geqslant \cdots \geqslant d_{\mathsf{K}} > 0$。记 M_0 为与特征值 $\boldsymbol{0}$ 对应的 D 的不变子空间，并令 M_0 为到 M_0上的正交投影。令 $\lambda_s(\widetilde{M}_0 D^{(1)} \widetilde{M}_0 | M_0)$ 表示将 $\widetilde{M}_0 D^{(1)} \widetilde{M}_0$ 限制在子空间 M_0上所得矩阵的第 s 大特征值。定义 $D(\eta) = D + \eta D^{(1)}$ 和 $\widetilde{d} = d_{\mathsf{K}}/2,$，其中 η 为满足 $0 < \eta < \widetilde{d}/\|D^{(1)}\|$的实数。则有

$$\left|\lambda_{\mathsf{K}+1}(D(\eta)) - \eta\lambda_1(\widetilde{M}_0 D^{(1)} \widetilde{M}_0 | M_0)\right| \leqslant 3\widetilde{d}\frac{|\eta|^2\|D^{(1)}\|^2}{(\widetilde{d} - |\eta|\|D^{(1)}\|)^2}.$$

证明。 参见Onatski(2009）中的引理 1。 □

引理 S.3。 设 $A = (A_{ij})_{n\times n}$ 表示一个没有自环的无向网络的邻接矩阵，$P = (P_{ij})_{n\times n}$ 表示边概率矩阵。对于 $1 \leqslant i < j \leqslant n$，其元素 A_{ij} 相互独立地服从 A_{ij} ~伯努利(P_{ij})分布。假设n最大值$_{i,\ j} P_{ij} \geqslant c_0$对数 n，其中 $c_0 > 0$为某个常数。那么，存在一个常数C > 0，使得其概率为 $1 - \mathrm{o}(1)$

$$\|A - \mathbb{E}(A)\| \leqslant \mathsf{C}\sqrt{n \max_{i,j} P_{ij}}.$$

证明。 参见Lei 和 Rinaldo (2015年）中的定理 5.2，以及Benaych-Georges 等人 (2020年）中的定理 3.2。 □

引理 S.4。 设 $\circ$ 表示哈达玛（逐项）积，$\lambda_s(\cdot)$表示矩阵的第 s大特征值。对于半正定矩阵 M、$B \in \mathbb{R}^{n\times n}$，有

$$\lambda_1(M \circ B) \leqslant \max_i M_{ii}\lambda_1(B), \quad and \quad \lambda_n(M \circ B) \geqslant \min_i M_{ii}\lambda_n(B).$$

证明。参见Horn 和 Johnson（1994）中的第 316 页定理 5.3.4。 □

引理 S.5. 设 $A = (A_{ij})_{n\times n}$ 为具有 K 个均衡社区的对称邻接矩阵。对角元满足 $A_{ii} = 0$，且非对角元 $\{A_{ij} : i < j\}$相互独立。令 q 表示期望平均度的平方根。假设对 $m \geqslant 2$ 及某个常数 $c > 0$，$\mathbb{E}(A_{ij}) = 0$、$\sum_i \mathbb{E}|A_{ij}|^2 = 1$和 $\mathbb{E}|A_{ij}|^m \leqslant \frac{(cm)^{cm}}{nq^{m-2}}$ 成立。用 $\lambda_s(\cdot)$ 表示一个矩阵的第 s 大特征值。 ——

(i). 若对所有较小的 $\phi > 0$、较大的 $\mathsf{C} > 0$ 以及固定的 i 均有 $1 \ll q \leqslant n^{1/2},$，则当 $n \to \infty$ 时，我们有

$$\mathbb{P}\big\{\big|\lambda_i(A) - \mathsf{L}\big| > n^{\phi}(q^{-4} + n^{-2/3})\big\} \leqslant n^{-\mathsf{C}},$$

其中 $\mathsf{L} = 2 + n\kappa_4 + \mathrm{O}(q^{-4})$，且 κ_4 表示满足 $|\kappa_4| = \mathrm{O}(n^{-1}q^{-2})$ 的 A_{ij} 4 阶累积量。

(ii)。此外，若对任意 $x_1 \in \mathbb{R}$ 都有 $n^{1/6} \ll q \leqslant n^{1/2},$，则有

$$\lim_{n\to\infty} \mathbb{P}\big(n^{2/3}(\lambda_1(A) - \mathsf{L}) \leqslant x_1\big) = \lim_{n\to\infty} \mathbb{P}\big(n^{2/3}(\lambda_1(W) - 2) \leqslant x_1\big),$$

其中 $W \in \mathbb{R}^{n\times n}$ 是一个 *GOE*矩阵。

证明。 （i）可由 Hwang等人 (2020年) 的定理 2.12 以及 Lee和Schnelli (2018年) 的引理 5.2 得出。（ii）见 Hwang等人(2020年) 的定理2.13。

□

引理 S.6。 在与引理*S.5*结果（*ii*）相同的条件下，对任意有限的$s > 0$ 以及任意 $x_s \in \mathbb{R},$

$$\begin{aligned}&\lim_{n\to\infty} \mathbb{P}\big(n^{2/3}(\lambda_1(A) - \mathsf{L}) \leqslant x_1, \cdots, n^{2/3}(\lambda_s(A) - \mathsf{L}) \leqslant x_s\big)\\ =&\lim_{n\to\infty} \mathbb{P}\big(n^{2/3}(\lambda_1(W) - 2) \leqslant x_1, \cdots, n^{2/3}(\lambda_s(W) - 2) \leqslant x_s\big).\end{aligned}$$

证明。 见 Lee和Yin (2014年) 的注 3.4、Lee和Schnelli (2018年) 的注 2.11，以及 Erd˝os 等人 (2012) 的方程 (2.42)。 □

引理 S.7. 设 $A = (A_{ij})_{n\times n}$ 为具有 K 个平衡社区的对称邻接矩阵。对角元满足 $A_{ii} = 0,$ 非对角元 $\{A_{ij} : i < j\}$独立地服从 $A_{ij} \sim$ 伯努利(P_{ij})分布，其中 $P = (P_{ij})_{n\times n}$ 表示边概率矩阵。记 $\lambda_s(\cdot)$ 为按绝对值计的第 s大特征值。假设 $\min_{i,j} P_{ij} \gg n^{-1}$，$\max_{i,j} P_{ij} = \mathrm{O}(1)$ 且 $\widetilde{\lambda}_{\mathsf{K}}(P) \gg n^{1/2}$。则对某个充分小的常数 $\delta > 0$ 以及充分大的正整数 C，以下结论以概率 $1 - \mathrm{o}(1)$成立，

$$\#\big\{i : \widetilde{\lambda}_i(A) \geqslant \delta\big\} \geqslant \mathsf{K} + \mathsf{C},$$

其中 $\#\{\cdot\}$表示该集合的基数。

证明。 证明分两种情况进行： $i \leqslant \mathsf{K}$ 和 $i > \mathsf{K}$。对于 $i \leqslant \mathsf{K}$，我们证明，以概率 $1 - \mathrm{o}(1)$，$\lambda_i(A)$ 发散，因此必然超过任意正

常数 δ。对于 $i > \mathsf{K}$，引理 S.5 的结果 (i) 意味着，对于每个固定的 i，存在一个充分小的常数 $\delta > 0$ 使得 $\lambda_i(A) \geqslant \delta$ 以概率 $1 - \mathrm{o}(1)$成立。因此，存在常数 δ 和 C，使得以概率 $1 - \mathrm{o}(1)$，至少有$\mathsf{K} + \mathsf{C}$ 个 A 的特征值远离 δ。

(i) **对于** $i \leqslant \mathsf{K}$。令 $A = A - \mathbb{E}(A)$ 为中心化噪声矩阵。应用引理S.3，可得

$$\|\widetilde{A}\| = \mathrm{O}_{\mathbb{P}}\big(n^{1/2}\big). \tag{S.31}$$

给定 $\mathbb{E}(A) = P - \mathsf{diag}(P)$，可写为 $A = P + A - \mathsf{diag}(P)$。记 $\lambda_s(\cdot)$ 为矩阵的第s大特征值。由韦尔不等式，得

$$|\lambda_i(A) - \lambda_i(P)| \leqslant \|\widetilde{A} - \mathrm{diag}(P)\|.$$

结合假设 $\widetilde{\lambda}_{\mathsf{K}}(P) \gg n^{1/2}$ 以及（S.31），以概率 $1 - \mathrm{o}(1)$,

$$\widetilde{\lambda}_{\mathsf{K}}(A) \gg n^{1/2}.$$

因此，对于任意常数 $\delta > 0$, $\lambda_i(A) \geqslant \delta$ 以概率 $1 - \mathsf{o}(1)$ 成立。

(ii) **对于** $i > \mathsf{K}$。回顾 $A = P + A - \mathsf{diag}(P)$。由韦尔不等式,

$$|\lambda_i(A) - \lambda_i(\widetilde{A} + P)| \leqslant \|\mathrm{diag}(P)\| = \max_i P_{ii}.$$

因此,

$$\lambda_i(A) = \lambda_i(\widetilde{A} + P) + \mathrm{O}(\max_i P_{ii}).$$

再次应用韦尔不等式可得 $\lambda_i(A + P) \geqslant \lambda_i(A) + \lambda_n(P)$。由此得到

$$\lambda_i(A) \geqslant \lambda_i(\widetilde{A}) + \lambda_n(P) + \mathrm{O}(\max_i P_{ii}).$$

由于 P的秩为 K，因此可得 $\lambda_n(P) = 0$。所以,

$$\lambda_i(A) \geqslant \lambda_i(\widetilde{A}) + \mathrm{O}(\max_i P_{ii}). \tag{S.32}$$

接下来，我们介绍辅助矩阵 $A \circ \mathbf{\Omega}$，其中 $\circ$ 表示 Hadamard 积，且 $\mathbf{\Omega} = (\Omega_{ij})_{n \times n}$ 定义为

$$\widetilde{\Omega}_{ij} = \{(n-1)P_{ij}(1 - P_{ij})\}^{-1/2}.$$

设 $q = n^{-1/2}(\sum_{i,j} P_{ij})^{1/2}$。在 $A_{ij} \sim$ 伯努利(P_{ij})假设下，可以验证对于 $1 \leqslant i \neq j \leqslant n$,

$$\mathbb{E}(\widetilde{A}_{ij} \circ \widetilde{\Omega}_{ij}) = 0, \quad \sum_i \mathbb{E}|\widetilde{A}_{ij} \circ \widetilde{\Omega}_{ij}|^2 = 1,$$

且对所有 $m \geqslant 2$

$$\mathbb{E}|\widetilde{A}_{ij} \circ \widetilde{\Omega}_{ij}|^m \leqslant \frac{(cm)^{cm}}{nq^{m-2}},$$

其中 $c > 0$ 是某个常数。在这些矩条件以及以下假设下：$\min_{i,j} P_{ij} \gg n^{-1}$，且网络被划分为若干规模均衡的~社区，引理 S.5 的结果 (i) 表明，对于固定的 i, $\lambda_i(A \circ \Omega) \to$ L，以~概率 $1-$o(1)成立，其中 L$-2 =$ o(1)。由于 $\Omega_{ij} \to 0$ 在 $1 \leqslant i, j \leqslant n$, $A \circ$中一致成立，~Ω是 A的一个趋于零的缩放。因此，$\lambda_i(A)$ 在渐近意义下的绝对值大于~的 $\lambda_i(A \circ \Omega)$。因此，对于固定的 i，存在某个常数 $0 < \delta <$ L，使得~ $\lambda_i(A) \geqslant \delta$ 以概率 $1-$o(1)成立。根据 (S.32)，可以选择 δ，使得$\lambda_i(A) \geqslant \delta$ 以概率 $1-$o(1)成立。

综合情形 (i) 与 (ii) 的讨论，我们得出结论：以概率$\gtrsim 1 -$ o(1)，存在常数 $\delta > 0$，使得 $\lambda_i(A) \geqslant \delta$对 $i = 1, \cdots,$ K $+$ C 成立，其中 C > 0 为固定整数。至此证明完毕。

□

S.2 用于确定K_{max}的算法

在本节中，我们给出计算K_{PA} 的详细步骤，见算法S.1，遵循Dobriban （2020年）。据此，在（1.5）中，我们设定$K_{max} = K_{PA} + 5$。

我们指出，为方便起见，在算法 S.1中，我们选择百分位数水平 q 和置换次数 B，使得 qB 为整数。还应注意的是，在针对大规模网络实际实现算法 S.1时，可以通过从 K_{PA} 中选取前 C 个最大特征值 A，而不是计算全部 n 特征值来加速计算，其中C 是一个足够大的常数，例如C $= \lfloor\sqrt{n}\rfloor$。

–

算法 S.1： 用于选择 K_{PA} 的平行分析

输入： 邻接矩阵 $A \in \mathbb{R}^{n\times n}$，置换次数B，百分位数水平 q。**1. 计算 A的特征值：** $\lambda_1(A) \geqslant \lambda_2(A) \geqslant \cdots \geqslant \lambda_n(A)$。**2. 基于置换的参考分布**。(i). **对于 $b = 1$ 到B，执行**

(a). 生成一个置换数组

$$\boldsymbol{\tau}^{(b)} = (\tau_1^{(b)}, \tau_2^{(b)}, \ldots, \tau_n^{(b)}),$$

其中每个 $\tau_j^{(b)}$ 都是 $\{1、\ldots\ldots、\ n\}$ 的独立随机置换，用于置换 A的第 j列。定义置换矩阵 $A_{\boldsymbol{\tau}}^{(b)}$，其元素为

$$(A_{\boldsymbol{\tau}}^{(b)})_{ij} = A_{\tau_j^{(b)}(i),j}, \qquad 1 \leqslant i, j \leqslant n.$$

(b). 计算 $A_{\boldsymbol{\tau}}^{(b)}$的特征值： $\lambda_1(A_{\boldsymbol{\tau}}^{(b)}) \geqslant \cdots \geqslant \lambda_n(A_{\boldsymbol{\tau}}^{(b)})$。

(ii). 对于每个 $j = 1,\ \cdots,\ n$，令 q_j为 $\{\lambda_j(A_{\boldsymbol{\tau}}^{(b)})$ 的经验 q分位数： $b = 1,\ \ldots\ldots,\ \mathsf{B}\}$。

输出： $\mathsf{K}_{\mathrm{PA}} =$ 最大值$\{j \in \{1,\ \cdots,\ n\} :\ \lambda_j(A) > q_j\}$。

S.3 补充数值结果

S.3.1 计算效率分析

本节报告 $\mathbb{T}$ 以及 $\mathbb{T}_{\mathrm{Lei}}$、 $\mathbb{T}_{\mathrm{Lei,\ boot}}$、 $\mathbb{T}_{\mathrm{Han}}$、 $\mathbb{T}_{\mathrm{Hu}}$ 和$\mathbb{T}^{\mathrm{aug}}{}_{\mathrm{Hu,boot}}$的计算时间。值得注意的是，求出 A 的全部 n 特征值的复杂度大致为 $\mathsf{O}(n^3)$，这在计算上可能代价高昂。相比之下，对于给定的 $\mathsf{K}_{\max}$，我们的检验统计量 $\mathbb{T}$只需计算 A的前 $\mathsf{K}_{\max}+2$ 个特征值。因此，计算复杂度可降低至 $\mathsf{O}(n^2\mathsf{K}_{\max})$；如果网络是稀疏的，甚至可降至 $\mathsf{O}(\mathsf{K}_{\max})$。因此，对于 $\mathbb{T}$，计算成本主要取决于网络规模、稀疏性和 $\mathsf{K}_{\max}$的取值，但与具体的模型结构无关。

我们在与第4节所用相同的设置下，对稠密和稀疏随机块模型进行模拟，不同之处在于我们考虑的网络规模为 $n = 3,000, 6,000$和$9,000$。我们将$\mathsf{K}_0 = 5$固定，并选取$\mathsf{K}_{\max} = \mathsf{K}_{\mathrm{PA}} + \mathsf{C}$，其中$\mathsf{C} = 2、5、10$。计算时间（以秒为单位）是在配备M3 Pro 芯片和 18 GB RAM的MacBook上获得的。我们注意到，检验统计量 $\mathbb{T}$的临界值是针对$\mathsf{K}_0, \mathsf{K}_{\max}$与 n的各种组合分别预先计算的。因此，在我们的数值分析中， $\mathbb{T}$ 直接与相应的预先计算临界值进行比较，所报告的计算时间不包括

计算临界值的开销。

根据图S.1中的结果，我们观察到，计算成本主要由网络规模 n决定，对 K_{max} 和网络密度的增加仅表现出轻微敏感性。在所有配置下，所提方法 $\mathbb{T}$始终具有最低的计算时间，而竞争方法的计算成本则显著更高。具体而言，$\mathbb{T}_{\text{Lei, boot}}$ 和 $\mathbb{T}^{\text{aug}}_{\text{胡, 启动}}$的运行速度慢了几个数量级，计算时间达到数万秒。相比之下，即使在稠密和稀疏网络中，n 增加到 9，000 时，$\mathbb{T}$ 的运行时间仍保持在亚秒级至个位数秒的水平。因此，我们所提方法在大规模网络分析中展现出更强的稳健性和实用可扩展性。

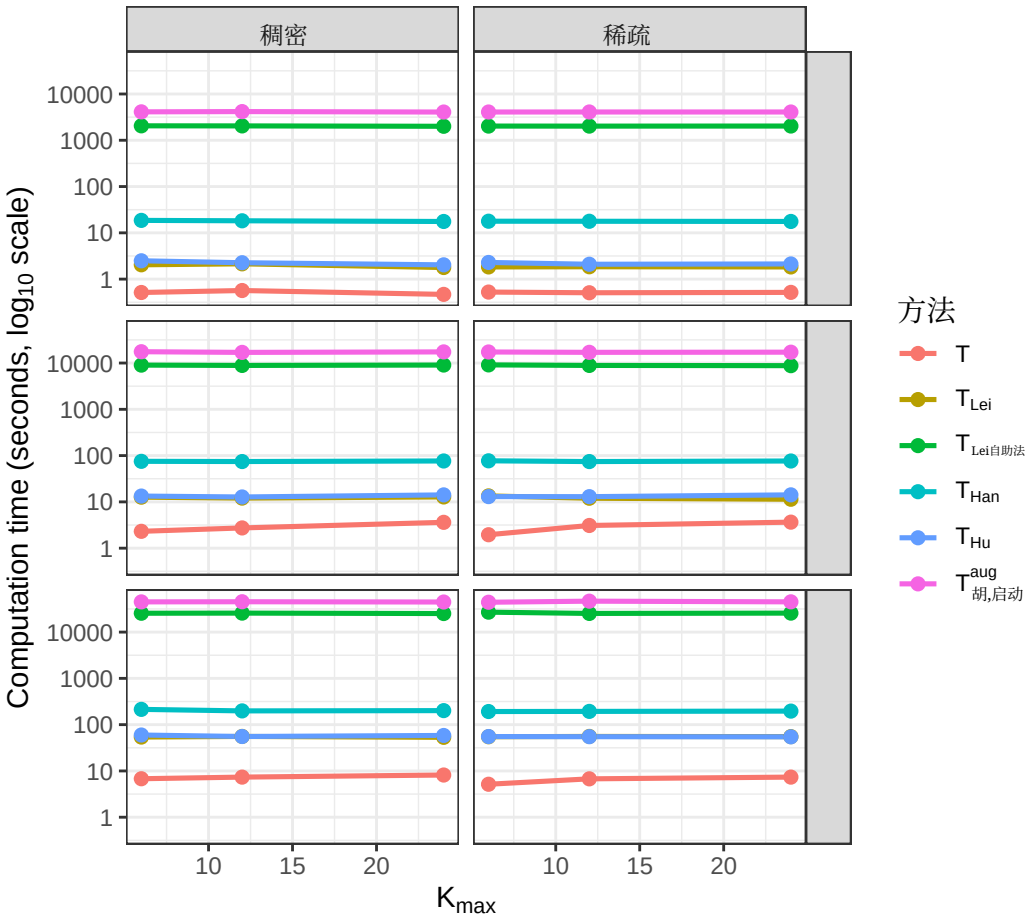

图 S.1：稠密和稀疏随机块模型的计算时间（秒）。网络中的社区规模相同，社区数为 $n \in \{3,000, 6,000, 9,000\}$，$K_0 = 5$，$C = 2, 5, 10$。

S.3.2 西蒙斯学院网络数据分析

本节中，我们考察另一个真实网络，该网络包含西蒙斯学院脸书用户之间的所有好友关系链接，记录于 2005年9月的某一天（Traud 等人，2012）。该网络由 1,137 个节点构成，边密度约为 3.76%。Traud 等人（2012）指出，该网络的社区结构与毕业年份高度相关，这意味着同一年毕业的学生更有可能成为朋友。因此，该网络中真实的簇数可视为$K = 2$。Jin 等人（2022）指出，西蒙斯学院网络的社区结构非常弱，社区检测算法通常会产生较高的误分类错误。

我们使用检验统计量 $\mathbb{T}$，以及$\mathbb{T}_{\text{Lei, 启动}}$、 $\mathbb{T}^{\text{aug}}{}_{\text{胡, 启动}}$ 和 $\mathbb{T}_{\text{Han}}$，检验$K_0 = 1,2$，因为所有涉及$K_0 > 2$ 的检验均得出相同结论。表S.1汇总的检验结果表明， $\mathbb{T}$ 和 $\mathbb{T}_{\text{Han}}$ 均拒绝 $H_0 : K = 1$，但不拒绝$H_0 : K = 2$，从而识别出至少两个社区。相比之下，$\mathbb{T}_{\text{Lei, 启动}}$ 和 $\mathbb{T}^{\text{aug}}{}_{\text{胡, 启动}}$ 拒绝所有原假设。因此，尽管西蒙斯学院网络中的社区结构较弱，我们的方法在识别K方面仍然适用且有效。

K_0	$\mathbb{T}$	$\mathbb{T}_{\text{Han}}$	$\mathbb{T}_{\text{Lei, 启动}}$	$\mathbb{T}^{\text{aug}}_{\text{Hu, boot}}$
1	拒绝	拒绝	拒绝	拒绝
2	**接受**	**接受**	拒绝	拒绝

表 S.1：西蒙斯学院网络的检验结果。

S.4 一些补充说明

注 *S.1.* 值得注意的是，在 *S.1.1* 节证明定理*1*时用到的两个关键结果，分别是 *Onatski (2009)* 的引理 *1* 和 *Hwang* 等人（*2020*）的定理*2.13*。具体而言，我们首先应用 *Onatski (2009)* 的引理 *1*，仔细界定任意 $k \geqslant K + 1$ 下 $\tilde{}$ $\tilde{}$ $\lambda_k(A)$ 与 $\lambda_{k-K}(A_{G_I} \circ \Omega_{G_I})$ 之间的差异，如式 *(S.24)* 所示。这里，$\tilde{}$ $\tilde{}$ 和 Ω_{G_I,i^*j^*} 是缩放因子，而 $A_{G_I} \circ \Omega_{G_I}$ 表示 A_{G_I} 的方差归一化版本，它是噪声矩阵 A 的一个 $n - K_0$ 主子矩阵。然而，*Onatski (2009)* 假设 K 有限，且 A 和 A_{G_I} 均为稠密连续矩阵。因此，要将 *Onatski (2009)* 的引理 *1* 推广到本文所研究的二元稀疏矩阵 A 和 A_{G_I}，尤其是在 K 发散时，会面临挑战。接下来，

我们利用*Hwang*等人*(2020*年）通过艾里核证明$n^{2/3}(\lambda_{k-\mathsf{K}}(\tilde{A}_{G_I} \circ \tilde{\Omega}_{G_I}) - \mathsf{L})$ 的分布收敛于 *I* 型 *Tracy-Widom* 分布 ˜ ˜，方法是为 $A_{G_I} \circ \Omega_{G_I}$ 建立高阶矩条件，其中 *L* 是一个与 k无关的位置参数。这反过来对上一步中˜的构造 A_{G_I} 提出了高阶矩要求。

注 S.2。 尽管现有研究（*Lei*, *2016*; *Hu* 等人，*2021*年; *Han* 等人，*2023*年）也涉及网络稀疏性与K发散速率之间的权衡，但我们的工作基于 $n^{1/3}\max_{i,\ j} P_{ij}/\mathsf{K}^2 \to \infty$提供了一个明确且有理论依据的判据。与现有方法相比，该判据允许网络更加稀疏，或允许K以更高的速率增长。值得注意的是，要确保 $n^{2/3}\big(\lambda_{k-\mathsf{K}}(\tilde{A}_{G_I} \circ \tilde{\Omega}_{G_I}) - \mathsf{L}\big)$ 的分布通过艾里核收敛至 *I*型*Tracy–Widom*分布（这是实施假设检验 （1.3）的基础），该判据是必要的。下文图 *S.2*对此作了说明。

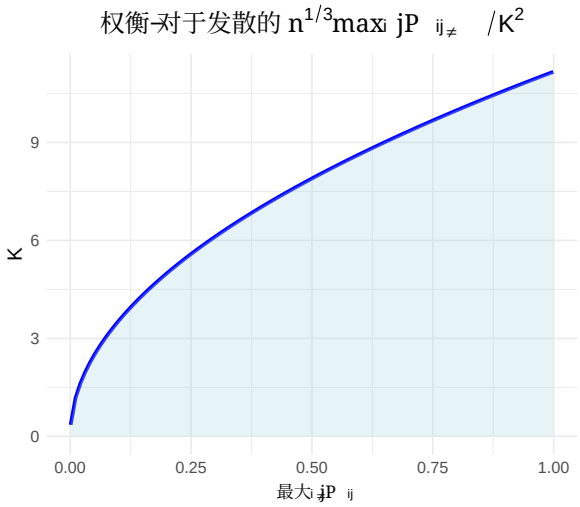

图 S.2： $n^{1/3}\max_{i,\ j} P_{ij}/\mathsf{K}^2 = 3n^{\epsilon_1}$ 的可行区域，其中 $n = 3,\ 000$ 且 $\epsilon_1 = 0.005$。蓝色阴影区域表示对于每个 $\max_{i,\ j} P_{ij}$，K 的可行取值，而蓝色实线表示 K 可取的最大值。

注 S.3. 这里有两点技术性说明。首先，假设 最小值$_{1\leqslant k\leqslant \mathsf{K}} |\lambda_k(P)| \geqslant cn$最大值$_{i,\ j} P_{ij}/\mathsf{K}$ 源于定理 *1*背后的两个核心要求：*(i)* K 个主导特征值 A 由 P所支配，而其余特征值则遵循˜中心化噪声矩阵 A的特征值；*(ii)* 其余$_{\max} - \mathsf{K} + 2$ 个特征值 A 通过艾里核依分布收敛于 *Tracy–Widom* 定律。要求 *(i)* 必须满足˜ 最小值$_{1\leqslant k\leqslant \mathsf{K}} |\lambda_k(P)|$ 大于 A的算子范数。要求 *(ii)* 还要求

这些$\mathsf{K}_{\max} - \mathsf{K} + 2$个特征值收敛至中心化噪声谱的速率（已证明由最小值$_{1\leqslant k\leqslant \mathsf{K}} |\lambda_k(P)|$决定）必须快于 *Tracy–Widom*分布的涨落尺度，即$\mathrm{O}(n^{-2/3})$。

~其次，为了应用*Hwang等人（2020年）*的定理*2.13*，并证明 $\lambda_{k-\mathsf{K}}(A_{G_I} \circ \tilde{\Omega}_{G_I})$ 通过艾里核渐近服从 *Tracy–Widom* 分布，从而能够使用 *Tracy–Widom* 定律刻画 $\mathbb{T}$的极限分布，我们对网络在各社区间相对均衡这一点提出一个温和的要求。具体而言，对于 $1 \leqslant u \leqslant \mathsf{K}$，令 $n_k = \sum_{i=1}^{n} \pi_{i,k}$ 表示可能位于社区k中的节点数。我们要求 $c_1^* \leqslant \min_{k,l} n_k/n_l \leqslant \max_{k,l} n_k/n_l \leqslant c_2^*$，其中 c_1^* 和 c_2^*为正的常数。

参考文献

Abbe, E.（2018年）。社区检测与随机块模型：最新进展。《机器学习研究期刊》*18*(177), 1–86。

Adamic, L. A. 和 N. Glance（2005年）。政治博客圈与2004年美国大选：他们各自写博客，彼此分裂。载于第*3*届链接发现国际研讨会论文集，第36–43页。ACM，纽约。

Airoldi, E. M., D. Blei, S. Fienberg 和 E. Xing（2008年）。混合成员随机块模型。《机器学习研究期刊》*9*(65), 1981–2014。

Amini, A. A. 和 E. Levina（2018年）。关于块模型的半定松弛。《统计年鉴》*46*(1), 149–179。

Bao, Z., G. Pan 和 W. Zhou（2015年）。具有一般总体的样本协方差矩阵最大特征值的普适性。《统计年鉴》*43*(1), 382–421。

Benaych-Georges, F., C. Bordenave 和 A. Knowles（2020年）。稀疏随机矩阵的谱半径。*Annales de l'Institut Henri Poincar´e, Probabilit´es et Statistiques 56*(3), 2141–2161。

Bickel, P. J. 和 A. Chen（2009）。从非参数视角看网络模型、Newman–Girvan 模块度及其他模块度。《美国国家科学院院刊》106(50), 21068–21073。

Chen, K. 和 J. Lei（2018）。用于确定网络数据中社区数量的网络交叉验证。《美国统计学会会刊》113(521), 241–251。Daudin, J.-J., F. Picard 和 S. Robin（2008）。随机图的混合模型。《统计与计算》18(2), 173–183。Ding, X.（2021）。具有多个可能总体成分的尖峰样本协方差矩阵。《随机矩阵：理论与应用》10(01), 2150014。Ding, X. 和 J. Xie（2025）。椭圆模型边缘特征值的 Tracy–Widom 分布。《信息与推断：IMA期刊》14(2), iaaf004。Ding, X. 和 F. Yang（2021）。尖峰可分离协方差矩阵与主成分。《统计年鉴》49(2), 1113–1138。Ding, X. 和 F. Yang（2022）。异质Gram矩阵的 Tracy–Widom 分布及其在信号检测中的应用。《IEEE信息论汇刊》68(10), 6682–6715。Dobriban, E.（2020）。用于因子分析和主成分分析的置换方法。《统计年鉴》48(5), 2824–2847。Erd˝os, L., A. Knowles, H.-T. Yau 和 J. Yin（2012）。Erd˝os-R´enyi 图的谱统计 II：特征值间距与极端特征值。《数学物理通讯》314(3), 587–640。Forrester, P. J.（1993）。随机矩阵系综的谱边缘。《核物理B》402(3), 709–728。Granovetter, M. S.（1973）。弱联系的力量。《美国社会学杂志》78(6), 1360–1380。Han, X., Q. Yang 和 Y. Fan（2023）。通过残差子抽样进行普适秩推断及其在大型网络中的应用。《统计年鉴》51(3), 1109–1133。Hoff, P. D., A. E. Raftery 和 M. S. Handcock（2002）。社会网络分析的潜在空间方法。《美国统计学会会刊》97(460), 1090–1098。Holland, P. W., K. B. Laskey 和 S. Leinhardt（1983）。随机块模型：初步研究。《社交网络》5(2), 109–137。

Horn, R. A. 和 C. R. Johnson（1994）。矩阵分析专题。剑桥大学出版社。Hu, J.、H. Qin、T. Yan 和 Y. Zhao（2020）。随机块模型的修正贝叶斯信息准则。《美国统计学会会刊》*115*(532), 1771–1783。Hu, J.、J. Zhang、H. Qin、T. Yan 和 J. Zhu（2021）。利用最大逐项偏差检验随机块模型的拟合优度。《美国统计学会会刊》*116*(535), 1373–1382。Huang, J.、B. Landon 和 H.-T. Yau（2020）。稀疏 Erdős–Rényi 图极端特征值从特雷西–维多姆到高斯涨落的转变。《统计概率年鉴》*48*(2), 916–962。Hwang, J. Y.、J. O. Lee 和 W. Yang（2020）。稀疏随机块模型的局部定律与特雷西–维多姆极限。伯努利 *26*(3), 2400–2435。Jin, J.、Z. T. Ke 和 S. Luo（2017）。通过单纯形顶点搜索估计网络成员归属。*arXiv* 预印本 *arXiv:1708.07852*。Jin, J.、Z. T. Ke 和 S. Luo（2022）。得分的改进，尤其是针对弱信号的改进。*Sankhya A 84*(1), 127–162。Jin, J.、Z. T. Ke 和 S. Luo（2024）。社交网络的混合成员估计。《计量经济学杂志》*239*(2), 105369。Jin, J.、Z. T. Ke、S. Luo 和 M. Wang（2023）。网络社区数量的最优估计。《美国统计学会会刊》*118*(543), 2101–2116。Karrer, B. 和 M. E. Newman（2011）。随机块模型与网络中的社区结构。《物理评论*E*：统计、非线性与软物质物理》*83*(1), 016107。Lee, J. O. 和 K. Schnelli（2018）。稀疏随机矩阵的局部定律与特雷西–维多姆极限。《概率论及相关领域》*171*, 543–616。Lee, J. O. 和 J. Yin（2014）。维格纳矩阵边缘普适性的充要条件。《杜克数学杂志》*163*(1), 117–173。

Lei, J.（2016）。随机块模型的拟合优度检验。《统计年鉴》*44*(1), 401–424。Lei, J. 和 A. Rinaldo （ 2015 ）。随机块模型中谱聚类的一致性。《统计年鉴》 *43*(1), 215–237。Lorrain, F. 和 H. C. White（1971）。社交网络中个体的结构等价。《数学社会学杂志》*1*(1), 49–80。Ma, S.、L. Su 和 Y. Zhang（2021）。确定度修正随机块模型中的社区数量。《机器学习研究期刊》*22*(1), 3217–3279。Noroozi, M.、R. Rimal 和 M. Pensky（2021）。流行度调整块模型中的估计与聚类。《英国皇家统计学会杂志*B*辑：统计方法学》*83*(2), 293–317。Onatski, A.（2009）。近似因子模型中因子个数的正式统计检验。《计量经济学》*77*(5), 1447–1480。Saldana, D. F.、Y. Yu 和 Y. Feng（2017）。有多少个社区？《计算与图形统计学杂志》*26*(1), 171–181。Tracy, C. A. 和 H. Widom（1994）。能级间距分布与艾里核。《数学物理通讯》*159*, 151–174。Tracy, C. A. 和 H. Widom（1996）。关于正交与辛矩阵系综。《数学物理通讯》*177*, 727–754。Traud, A. L.、P. J. Mucha 和 M. A. Porter（2012）。脸书网络的社会结构。《物理学*A*：统计力学及其应用》*391*(16), 4165–4180。Wang, Y. R. 和 P. J. Bickel（2017）。随机块模型的基于似然的模型选择。《统计年鉴》*45*(2), 500–528。Wolfe, P. J. 和 S. C. Olhede（2013）。非参数图子估计。*arXiv* 预印本 *arXiv:1309.5936* 。Wu, Y.、W. Lan、T. Zou 和 C.-L. Tsai（2022）。网络内向与外向影响分析。《商业杂志》& 《经济统计》*40*(4), 1617–1628。

Zhao, Y., E. Levina 和 J. Zhu（2012）。度修正随机块模型下网络社区检测的一致性。《统计年鉴》*40*(4)，2266– 2292。