arXiv:2304.00739v1 [stat.AP] 3 Apr 2023

稀疏随机块模型的双样本检验

Qianyong Wu 和 Jiang Hu*

东北师范大学数学与& 统计学院，中国长春

摘要

本文讨论了一个统计问题，即检验两个具有社区结构的稀疏网络之间是否存在差异。社区边概率矩阵的元素为 $O(n^{-1}/\log\ n)$ 量级，其中 n表示网络规模。作者提出了一种检验统计量，该统计量结合了 Wu 等人 [39] 提出的方法和重抽样过程。他们推导了该检验统计量的渐近零分布，并保证了其对备择假设具有渐近功效。为评估所提出检验统计量的性能，作者进行了模拟并提供了真实数据示例。结果表明，该检验统计量在实践中表现良好。

关键词：
随机区块模型、假设检验、稀疏网络、Gumbel 分布。

1. 引言

网络已广泛应用于描述不同领域中个体之间的关系或复杂系统各单元之间的相互作用，

*通讯作者

提交至 *Elsevier* 的预印本

例如生物学、计算机科学、社会学等诸多领域 [17, 10, 12, 31, 38, 4]。随着数据收集的发展以及社交媒体的爆发式增长，网络分析已成为数据分析中日益重要的组成部分。

给定一个大型网络，一项重要任务是找出其社区结构。例如，识别客户网络中的社区有助于推广某个产品品牌 [21]。在蛋白质相互作用网络中，社区是具有特定细胞功能的蛋白质群体；发现这些社区有助于我们了解细胞功能的组织方式和动态变化 [6]。在统计学家的合著与引文网络中，社区检测有助于我们发现统计学家的研究习惯、趋势和拓扑模式 [17, 18]。随机区块模型 [14] 是网络数据中一种非常流行的社区结构模型，在计算机科学和统计学文献中得到了广泛研究。随机区块模型将 n 个节点分配到 K个互不相交的区块中。给定节点的隶属关系后，观测到的网络由对称邻接矩阵 A 表示。对称邻接矩阵 A 的各项 A_{ij}（$i > j$）是相互独立的伯努利随机变量，其参数 E（A_{ij}）仅取决于节点 i 和 j 的隶属关系。

社区检测是指在观察到 A的一次观测结果后，恢复其隶属关系。这一问题受到了不同研究领域的广泛关注，已有许多方法被提出，例如谱聚类 [19, 27, 25], 似然方法[2, 4, 29] 和模块度最大化[28], ，综述请参见 [1] 。然而，这些方法大多假设我们知道簇的数量，而实际上我们并不知道。例如，谱聚类及其变体会估计社区隶属关系

通过对由 A的K_0 个领先奇异向量构成的矩阵的各行应用 k-均值聚类。其中， K_0 是假设的社区数量，必须预先给定。大多数此类方法将簇的数量 K 视为先验已知，但在实践中， K 通常是未知的。为了解决这一问题，已有大量方法被提出，包括基于似然的方法 [37, 34], 、基于随机矩阵理论的方法 [24, 5, 22, 9], 、网络交叉验证 [26, 7, 15], 以及最大逐元素偏差法 [16], ，这里只列举其中几种。

与单网络问题类似，随机网络的双样本假设检验自然地出现在神经科学、社交网络和机器学习中 [3, 41, 6]。已有多项工作研究这一重要问题。Tang 等人通过邻接谱嵌入提出了一种用于比较两个随机点积图的统计量 [35, 36]。Ghoshdasti- dar 等人在 [12] 提出了两个检验统计量，分别基于两个样本的连边概率矩阵之间的弗罗贝尼乌斯范数和谱范数的估计，并清晰刻画了极小极大分离阈值。Ghoshdastidar 和 von Luxburg [13], Chen 等人[8]提出了基于缩放并中心化后的矩阵的极端特征值的检验统计量，并证明在原假设下，这些统计量近似服从特雷西-维多姆分布。

然而，大多数文献关注的都是样本量较大的情形，这限制了推断的适用范围。例如，组学数据通常会产生一个大型相互作用网络，也就是说，任一总体的样本量都是一。因此，我们需要针对小样本情形设计新的双样本检验。Hu 等人 [16] 提出了一种基于

随机区块模型的邻接矩阵最大元素的拟合优度检验。受 Hu 等人 [16], 工作的启发，Wu 等人 [39] 提出了一种基于 SBM 的方法，用于检验两个稠密随机网络，其中样本量为一。

早期研究关注稠密网络，但当网络变得稀疏时，其性能通常会变得不尽如人意。其中一个原因是，期望度较小的稀疏网络无法集中于其期望值附近。这是因为此类网络中存在度异常高或异常低的顶点 [23]。然而，稀疏网络在现实场景中十分常见，例如社交网络和基因共表达网络；尽管其中包含数十万个对象，其边可能仍不易被观测到。值得注意的是，稀疏网络已被证明对机器学习十分有用，并且在各种实际应用中展现出了有效性 [30, 32, 2, 20, 40]。在本项目中，我们将提出一种针对稀疏随机块模型进行双样本检验的新方法。

本文结构安排如下。第 2 节说明 Wu 等人提出的方法为何在稀疏网络中失效。第 3 节给出新的检验统计量及其渐近零分布和渐近功效。为展示该检验统计量的表现，第 4 节和第 5 节分别提供模拟结果和真实数据示例。最后，第 6 节总结全文。

2. 稠密网络上的双样本检验问题

假设给定两个具有相同顶点的随机网络，它们的邻接矩阵由对称连接概率矩阵 P_1 生成，其中 $A_{1,\ ij} \sim Bernoulli(P_{1,\ ij})$、$i$，$j = 1$，2、...、$n$ 以及另一组邻接矩阵，$_{ij}$ 由连接概率矩阵 P_2生成。这两个网络之间是否存在差异？这个问题可以表述为以下假设检验框架：

$$H_0 : P_1 = P_2 \quad vs \quad H_1 : P_1 \neq P_2. \tag{1}$$

假设 A_1 和 A_2 具有社区结构，因此可以用随机块模型建模。考虑一个包含 n 个节点和 K 个社区的随机区块模型，其邻接矩阵 A_ℓ, $\ell = 1$，2，的概率质量函数为

$$P(A_\ell) = \prod_{1 \leq i < j \leq n} B_{\ell, g_i g_j}^{A_{\ell,ij}} (1 - B_{\ell, g_i g_j})^{1 - A_{\ell,ij}}.$$

其中，$g \in \{1, 2, \ldots, K\}^n$ 是成员向量，而 $B_\ell \in [0, 1]^{K \times K}$,$\ell = 1$，2 是对称的社区级边概率矩阵。因此，$A_1$ 和 A_2 可分别由 (g_1，B_1) 和 (g_2，B_2) 参数化。在此情形下，假设检验 (1) 变为

$$H_0 : B_1 = B_2 \quad vs \quad H_1 : B_1 \neq B_2. \tag{2}$$

在假设 $g := g_1 = g_2$ 的条件下。

给定假设的社区数量 K，设 $\hat{}g$ 为估计的社区成员向量。定义 $\hat{}n_k = \#(\hat{}V_k)$，以及 $\hat{}\ V_k = \hat{}g^{-1}(k)$，$1 \leq k \leq K$。我们考虑 B_ℓ、$\ell = 1$，2 的代入估计量，这些估计量由

在 [24] 中
提出:

$$\hat{B}_{\ell,kl} = \begin{cases} \frac{\sum_{i\in\hat{V}_k, j\in\hat{V}_l} A_{\ell,ij}}{\hat{n}_k\hat{n}_l}, & k \neq l, \\ \frac{\sum_{i,j\in\hat{V}_k, i\leq j} A_{\ell,ij}}{\hat{n}_k(\hat{n}_k-1)/2}, & k = l. \end{cases} \tag{3}$$

为了使用发散的 K 检验原假设 (1)，Wu 等人[39] 提出了逐项偏差,

$$\hat{\rho}_{ik} = \frac{1}{\sqrt{\#(\hat{g}^{-1}(k)\setminus\{i\})}} \sum_{j\in\hat{g}^{-1}(k)\setminus\{i\}} \frac{A_{1,ij} - A_{2,ij}}{\sqrt{\hat{B}_{1,\hat{g}_i\hat{g}_j}(1-\hat{B}_{1,\hat{g}_i\hat{g}_j}) + \hat{B}_{2,\hat{g}_i\hat{g}_j}(1-\hat{B}_{2,\hat{g}_i\hat{g}_j})}},$$

其中 $\hat{g}^{-1}(k) = \{i : 1 \leq i \leq n, \hat{g}_i = k\}$，而 $\#(\hat{g}^{-1}(k)\setminus\{i\})$ 表示集合 $\hat{g}^{-1}(k)\setminus\{i\}$ 的基数。

记 $L_n = \max_{1\leq i\leq n,\ 1\leq k\leq K} |\hat{\rho}_{ik}|$，在其论文中的稠密网络假设下，Wu 等人 [39] 建立了 L_n的理论性质。他们提出的检验不适用于稀疏网络，正如他们在论文中所讨论的那样。事实上，当社区间边概率 $\max_{1\leq i\leq K,\ 1\leq k\leq K} B_{ij} = O(\frac{\log n}{n})$ 时，$\hat{\rho}_{ik}$可以表示为一个标准正态分布 $N(0, 1)$ 加上一个额外项 $O_p(\frac{1}{\sqrt{\log n}})$。在这种情况下，检验统计量 $L_n^2 - 2\log(2Kn) + \log\log(2Kn)$ 收敛于 I 型极值分布加上一个阶为 $O_p(1)$ 的项。因此，在稀疏网络情形下，$L_n^2 - 2\log(2Kn) + \log\log(2Kn)$ 不会收敛于 Gumbel 分布；正如 Wu 等人 [39] 在稠密网络情形下所证明的那样。

3. 重采样方法

已有多种方法被提出，以应对网络的稀疏性。Jing 等人 [20] 提出了一种利用对称化拉普拉斯逆矩阵（SLIM）划分稀疏网络的方法。Gao 等人 [11]

证明了归一化谱聚类有助于改善稀疏网络中的聚类结果。Wu 等人 [40]开发了一种重采样方法，用于检验随机块模型能否充分拟合具有 K 个社区的稀疏网络。

本文将通过对 $\hat{\rho}_{ik}$ 进行重抽样，提出一种新颖的检验统计量，并建立其理论性质。首先，根据 Wu 等人 [39] 计算 $\hat{\rho}_{ik}$ ($1 \leq i \leq n$，$1 \leq k \leq K$)。其次，对于任意固定的社区 $k \in \{1, 2, \cdots, K\}$，我们从 $\{\hat{\rho}_{1k}, \cdots \hat{\rho}_{nk}\}$ 在第 m 次实现 ($m = 1, \cdots, M$) 中随机抽取 S 个样本，并将它们记为 $\{\hat{\rho}_{m_1 k}, \cdots \hat{\rho}_{m_S k}\}$。随后，计算

$$\hat{\gamma}_{mk} = \frac{\hat{\rho}_{m_1 k} + \hat{\rho}_{m_2 k} + \hat{\rho}_{m_S k}}{\sqrt{S}}.$$

在这种情况下，$\hat{\gamma}_{mk}$ 可以表示为标准正态分布 $N(0, 1)$ 加上一个额外项 $O_p(1/\sqrt{S \log n})$。可以看出，重抽样过程会加快额外项收敛至 0 的速度。记为 $F_n = \max_{1 \leq i \leq n,\ 1 \leq k \leq K} |\hat{\gamma}_{mk}|$，直观上，$F_n$ 取较大值为拒绝式 (2) 中的 H_0 提供了证据。为了建立理论结果，我们需要三个假设。

假设1。 存在两个正常数 c_1 和 c_2 ，使得 $\frac{c_1 n}{K} \leq$ 最小值$_{1 \leq k \leq K} n_k \leq$ 最大值$_{1 \leq k \leq K} n_k \leq K^{\frac{c_2 n^2}{\text{对数}\, 2n}}$ 对所有 n 均成立。

假设2。 对于 $\ell = 1$，2 以及 $1 \leq k$，$l \leq K$，$B_{\ell,\ kl} \in (0,\ 1)$ 和 最小值$_{k,\ l} n B_{\ell,\ kl} \to \infty$，且 B_ℓ 没有相同的行，对于 $\ell = 1$，2。

假设3. 假设 $S =$ 最大值$\{O(q^{-1/2})\ o((n/\text{对数}\ n)^{c_3})\}$，其中 $q =$最大值$_{k,l} B_{\ell,\ kl}$ 和 $K = o((n/\text{对数}\ n)^{c_3})$，对于 $0 < c_3 < 0.5$。$MK = o(n)$ 且最小值 $\{n, M, S\} \to \infty$。

令 $T := T_n = L_n^2 - 2\log(MK) + \log\log(MK)$。那么，检验统计量 T 有如下渐近结果。

定理1. 假设假设*1*、*2*和*3*成立。那么在原假设 $H_0: B_1 = B_2$下，当 $n \to \infty$ 时，我们有

$$\lim_{n\to\infty} P(T \le y) = \exp\left\{-\frac{1}{\sqrt{\pi}} e^{-y/2}\right\}. \tag{4}$$

定理1的证明见补充材料。

(4)式右侧是位置参数为$\mu = -2\log(\sqrt{\pi})$、尺度参数为 $\beta = 2$ 的 Gumbel 分布的累积分布函数。因此，对于检验问题 (2)，当 $T \ge t_{1-\alpha}$ 时，我们拒绝 H_0，其中 t_α 是 Gumbel 分布 $Gumbel(\mu, \beta)$ 的第 α 个分位数。

接下来，我们考虑基于 T 的检验之间的比较，并将其总结在以下定理中。我们的工作考虑最大期望节点度的量级小至 $\log n$ 的稀疏网络。

定理2. 在备择假设 $H_1: B_1 \ne B_2$下，假设假设*1*、*2*和*3*成立，且 $\max_{1\le i\le n, 1\le j\le n} |B_{1,g_i g_j} - B_{2,g_i g_j}| = \Omega(\sqrt{\log_n^{MK}})$，则当 $n \to \infty$ 时，我们有

$$P(T \ge c\log(MK)) \to 1,$$

对于任意正数 c。

定理2的证明汇集在补充材料中。

4. 数值实验

本节通过模拟数据集展示我们提出的检验统计量的性能。请注意，在以下设置中，我们使用对称化拉普拉斯逆矩阵（SLIM）算法，在 [20] 中估计标签 $\hat{g}$。

4.1. 零分布

在本次模拟中，我们考察检验统计量 T 的有限样本零分布，并验证定理1中的结果。假设有两个社区，其社区级边概率矩阵为 $B_1 = B_2 = B$，其中 $B_{11} = B_{22} = 7\frac{\log n}{n}$ 且 $B_{12} = B_{21} = 3\log_n{}^{n}$。我们考虑一个规模较小的网络，其中 $n = 200$。成员向量 g 通过对每个分量独立抽样生成，各分量以相同概率取自 $\{1, 2\}$。

图1展示了原假设下进行1,000次独立重复试验所得的所提检验统计量直方图。图中还绘出了理论极限分布（红线）作为参考。该图直观地验证了定理1中的结果。

4.2. 稠密网络下的第一类错误和第二类错误

在本次模拟中，我们将检验统计量 T 与 Wu 等人在[39] 中提出的最大逐元素偏差统计量，以及 Ghoshdastidar 等人在[13] 和 Chen 等人在 [8], 中提出的最大特征值统计量进行比较；他们的检验统计量分别记为 T^{+} 和 TW。

我们采用与 [39] 相同的参数设置。社区 k 和 l 之间的边概率为 $B_{1,kl} = 0.05r + 0.05r \times I\{k = l\}$,

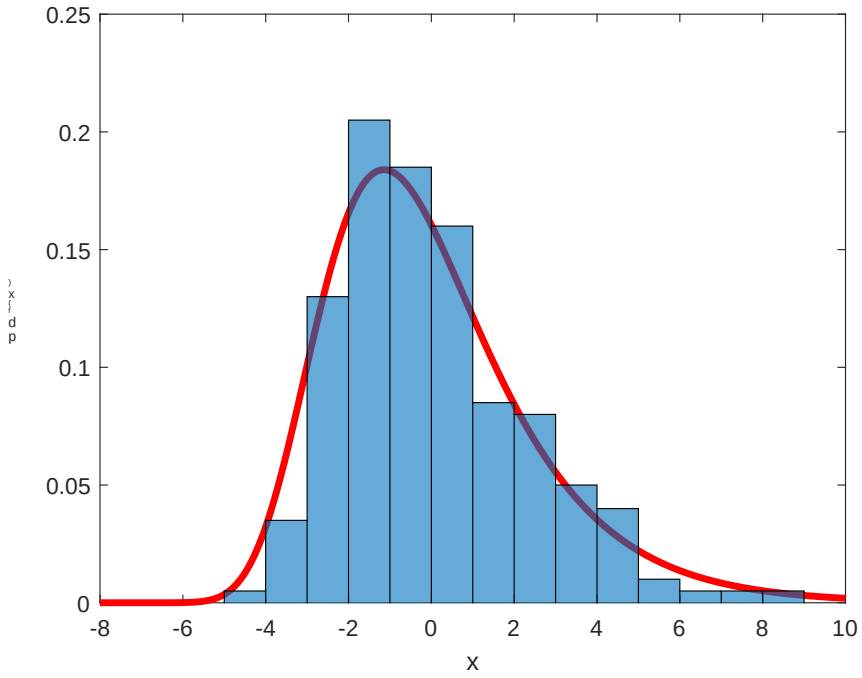


图 1：检验统计量 T 的直方图，以及参数为 $\mu = -2\log(\sqrt{\pi})$ 和 $\beta = 2$的 Gumbel 分布密度函数。

$B_{2,\ kl} = B_{1,\ kl} + \epsilon r \times I\{k = l\}$，其中 r 控制网络的稀疏性。在原假设 $\epsilon = 0$ 下，显然有 $B_1 = B_2$。在备择假设下，我们设定 $\epsilon = 0.04$，这与 [39] 中的设定相同。成员向量 g 的生成方式是：从 $\{1, 2\}$ 中以相同概率独立抽取每个分量。首先，我们固定 $r = 1$，并令 n 从 200 增至 1000。在进行 1,000 次独立重复试验后，名义显著性水平为 0.05 时的拒绝比例如图 2 所示。原假设情形对应的虚线表示 5% 的显著性水平。这表明，我们的检验统计量 T 始终优于 T^+

以及 TW。接下来我们考虑 $r = 0.5$，其余参数均相同

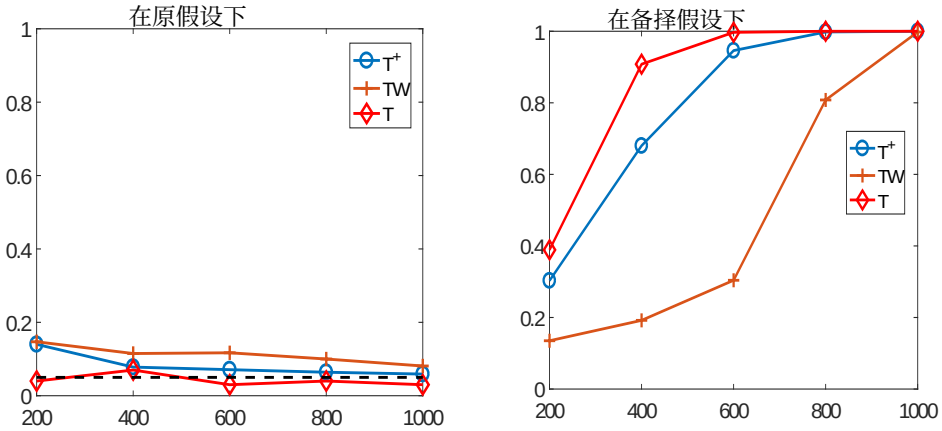


图 2：顶点数不断增加时各项检验的比较 n 和 $r =$ 1.

与设定 $r = 1$相同。从图 3 可以看出，我们的检验统计量 T 比 T^+和 TW 更能适应稀疏网络。

4.3. 稀疏网络下的第一类错误和第二类错误

在本次模拟中，我们展示检验统计量 T的表现，因为另外两个检验统计量 T^+和 TW在稀疏网络设置中表现不佳。在稀疏网络设置中，社区之间的边概率 k和 l分别为$B_{1,\ kl} = 0.5r + 0.5r \times I\{k = l\}$、 $B_{2,\ kl} = B_{1,\ kl} + \epsilon r \times I\{k = l\}$，其中 $r = \frac{n}{\text{对数}}$。在原假设 $\epsilon = 0$下，我们有 $B_1 = B_2$。在备择假设下，我们设定 $\epsilon = 0.4$。成员向量 g通过从 $\{1, 2\}$ 中以相同概率独立抽取每个元素生成。我们令网络规模 n 从1000增加到5000。在进行1,00次独立重复试验后，名义显著性水平为0.05时的拒绝比例见图4。原假设情形对应的虚线

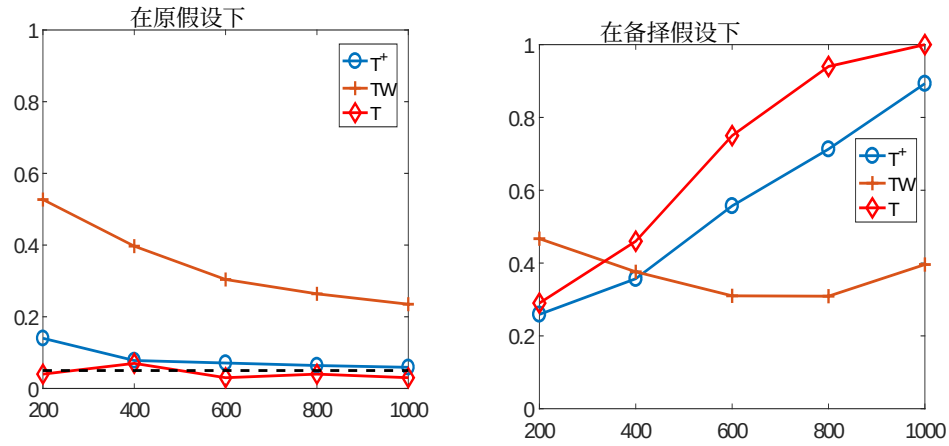


图3：随着顶点数量 n 和 $r = 1$ 的增加，对检验进行比较。

对应显著性水平5%。这表明，我们的检验统计量 T 在稀疏网络设置中表现良好。

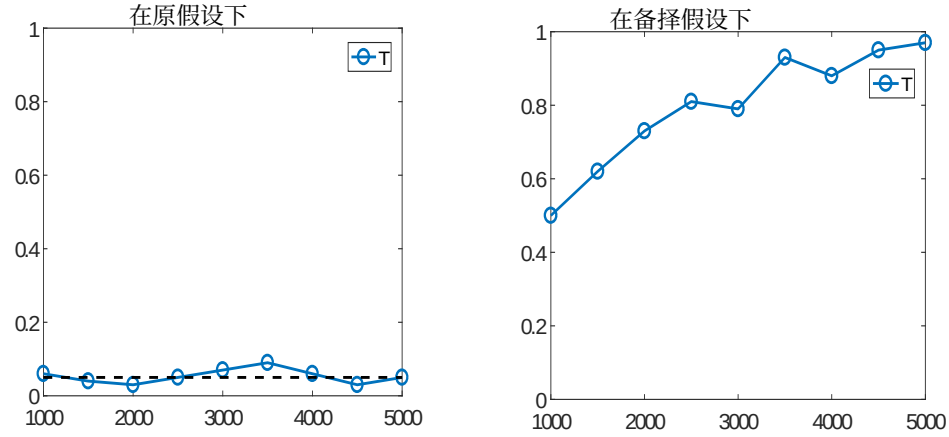


图4：随着顶点数量 n 和 $r = 1$ 的增加，对检验进行比较。

5. 真实数据示例

本节将对真实数据集进行实验。该数据集采集自奥胡斯大学计算机科学系，涉及员工互动。研究总体包括61名自愿参与调查问卷的员工，其中包括教授、博士后研究人员、博士生和行政人员，占全体142名员工的一部分。该数据集包含五个邻接矩阵，重点用于测量5个结构变量，即当前的工作关系、反复进行的休闲活动、定期共进午餐、论文的合作著述，以及在脸书上的友谊。参与者之间的关系是二元的，并基于他们之间的互动，不带任何权重。前3个变量通过在线发放的调查问卷进行评估，测量参与者对自身情况的评价。合作著述关系取自在线DBLP书目数据库，未与参与者进行任何直接互动。研究人员使用定制应用程序，获取所有报告拥有脸书账号的个体之间的友谊关系。更多详情请参见 [33]。

我们希望刻画工作、休闲、午餐、合作著述和友谊这五个社区层面的边概率之间的差异，分别记为 B_W、 B_{L_1}、 B_{L_2}、 B_C 和 B_F。在统计上，我们对这五个矩阵进行两两比较，共开展十个双样本检验问题。我们使用 T 作为检验统计量，并计算其取值，结果列于表1。由于 Gumbel 分布的 $t_{0.95} = 4.79$，我们以充分证据在 0.05 的显著性水平下拒绝上述十个原假设。对于

说明，图5展示了这五个网络，其中用户根据聚类结果着色。

表1：五个矩阵之间十个双样本检验问题对应的检验统计量值。

Work	47.05	14.99	50.19	12.95
	休闲	11.43	13.44	11.58
		午餐	20.50	42.47
			合作著述	22.76
				友谊

6. 结论

本文提出了一种新方法，用于检验两个稀疏网络是否来自具有 K 个社区的同一随机块模型。我们推导了原假设下检验统计量的渐近零分布，并给出了备择假设下的理论检验效能。现实网络通常表现出度异质性，可通过将所提方法扩展到度修正随机块模型来处理。然而，当度修正参数未知时，由于重缩放邻接矩阵各元素之间存在复杂的依赖关系，推导渐近分布会面临挑战。我们将此留待未来研究。模拟和真实数据示例验证了我们的理论结果。

致谢

Jiang Hu 得到了国家自然科学基金委员会的资助（项目编号：12171078 和 11971097）。

参考文献

[1] Abbe, E., 2018. 社区检测与随机块模型：最新进展。机器学习研究杂志 18, 1–86。

[2] Amini, A.A., Chen, A., Bickel, P.J., Levina, E., 2013. 大型稀疏网络中的社区检测伪似然方法。《统计年鉴》41, 2097–2122。

[3] Bassett, D.S., Bullmore, E., Verchinski, B.A., Mattay, V.S., Weinberger, D.R., Meyer-Lindenberg, A., 2008. 健康与精神分裂症状态下人类皮层网络的层级组织。《神经科学杂志》28, 9239–9248。

[4] Bickel, P., Choi, D., Chang, X., Zhang, H., 2013. 随机块模型中最大似然及其变分近似的渐近正态性。《统计年鉴》41, 1922–1943。

[5] Bickel, P.J., Sarkar, P., 2016. 网络自动社区检测的假设检验。英国皇家统计学会会刊：B辑（统计方法学）78, 253–273。

[6] Chen, J., Yuan, B., 2006. 在酵母蛋白质相互作用网络中检测功能模块。《生物信息学》（英国牛津）22，2283–2290。

[7] Chen, K., Lei, J., 2018. 用于确定网络数据中社区数量的网络交叉验证。《美国统计学会杂志》113，241–251。

[8] Chen, L., Zhou, J., Lin, L., 2021. 网络总体的假设检验. 统计学通讯：理论与方法 0, 1–24.

[9] Dong, Z., Wang, S., Liu, Q., 2020. 复杂网络中用于社区检测的谱方法假设检验. 信息科学 512, 1360–1371.

[10] Gangrade, A., Venkatesh, P., Nazer, B., Saligrama, V., 2019年. 平衡随机区块模型中社区变化的高效近最优检验. 神经信息处理系统进展 32.

[11] Gao, C., Ma, Z., Zhang, A.Y., Zhou, H.H., 2017. 在随机区块模型中实现最优误分类比例. 机器学习研究期刊 18, 1–45.

[12] Ghoshdastidar, D., Gutzeit, M., Carpentier, A., von Luxburg, U., 2020. 非齐次随机图的双样本假设检验. 统计学年刊 48, 2208–2229.

[13] Ghoshdastidar, D., von Luxburg, U., 2018. 图双样本检验的实用方法. Advances in Neural Information Processing Sys- tems 31.[14] Holland, P.W., Laskey, K.B., Leinhardt, S., 1983年. 随机区块模型：初步研究. Social Networks 5, 109–137.

[15] Hu, J., Qin, H., Yan, T., Zhao, Y., 2020a. 随机区块模型的修正贝叶斯信息准则. 《美国统计学会杂志》115, 1771–1783.

[16] Hu, J., Zhang, J., Qin, H., Yan, T., Zhu, J., 2020b. 使用最大逐元素偏差检验随机区块模型的拟合优度。《美国统计学会杂志》0, 1–10。

[17] 季, P., 金, J., 2016. 统计学家的合作著述与引文网络。《应用统计年鉴》10, 1779–1812。

[18] 季, P., 金, J., 柯, Z.T., 李, W., 2021. 统计学家的共引网络与合作著述网络。《商业与经济统计杂志》0, 1–17。

[19] 金, J., 2015. 通过 SCORE 快速进行社区检测。《统计年鉴》43, 57–89。

[20] Jing, B., Li, T., Ying, N., Yu, X., 2021. 使用对称化拉普拉斯逆矩阵（SLIM）进行稀疏网络社区检测。Sta- tistica Sinica .[21] Krishna Reddy, P., Kitsuregawa, M., Sreekanth, P., Srinivasa Rao, S., 2002. 一种基于图的方法：提取用于协同过滤的邻域客户社区，载于：网络信息系统中的数据库，施普林格，柏林，海德堡。第 188–200 页。[22] Le, C.M., Levina, E., 2022. 用谱方法估计社区数量。《电子统计学杂志》16, 3315–3342。

[23] Le, C.M., Levina, E., Vershynin, R., 2017. 随机图的集中性与正则化。Random Structures & Algorithms 51，538–561。[24] Lei, J., 2016. 随机区块模型的拟合优度检验。Annals of Statistics 44，401–424。[25] Lei, J., Rinaldo, A., 2015. 随机区块模型中谱聚类的一致性。Annals of Statistics 43，215–237。[26] Li, T., Levina, E., Zhu, J., 2020. 通过边采样进行网络交叉验证。Biometrika 107，257–276。[27] Ma, X., Wang, B., Yu, L., 2018. 基于聚类方法等价性的复杂网络社区检测半监督谱算法。Physica A: Statistical Mechanics and its Applications 490，786–802。[28] Newman, M.E.J., 2006. 使用矩阵特征向量寻找网络中的社区结构。Physical Review E. Statistical, Nonlinear, and Soft Matter Physics 74，036104，19。[29] Newman, M.E.J., Leicht, E.A., 2007. 网络中的混合模型与探索性分析。Proceedings of the National Academy of Sciences 104，9564。[30] Pal, S., Zhu, Y., 2021. 稀疏超图随机块模型中的社区检测。Random Structures & Algorithms 59，407–463。arXiv:1904.05981。

[31] Pontes, B., Gir´aldez, R., Aguilar-Ruiz, J.S., 2015. 表达数据的双聚类：综述。《生物医学信息学杂志》57, 163–180.

[32] Rohe, K., Chatterjee, S., Yu, B., 2011. 谱聚类与高维随机块模型。《统计年鉴》39, 1878–1915.

[33] Rossi, L., Magnani, M., 2015. 迈向多重网络与多层网络的有效可视分析。《混沌、孤子与分形》72, 68– 76.

[34] Salda˜na, D.F., Yu, Y., Feng, Y., 2017. 有多少个社区？《计算与图形统计学杂志》26, 171–181.

[35] Tang, M., Athreya, A., Sussman, D.L., Lyzinski, V., Park, Y., Priebe, C.E., 2017a. 随机图的半参数双样本假设检验问题。《计算与图形统计学杂志》26, 344–354.

[36]Tang, M., Athreya, A., Sussman, D.L., Lyzinski, V., Priebe, C.E., 2017b. 随机图的非参数双样本假设检验问题。伯努利 23，1599–1630。

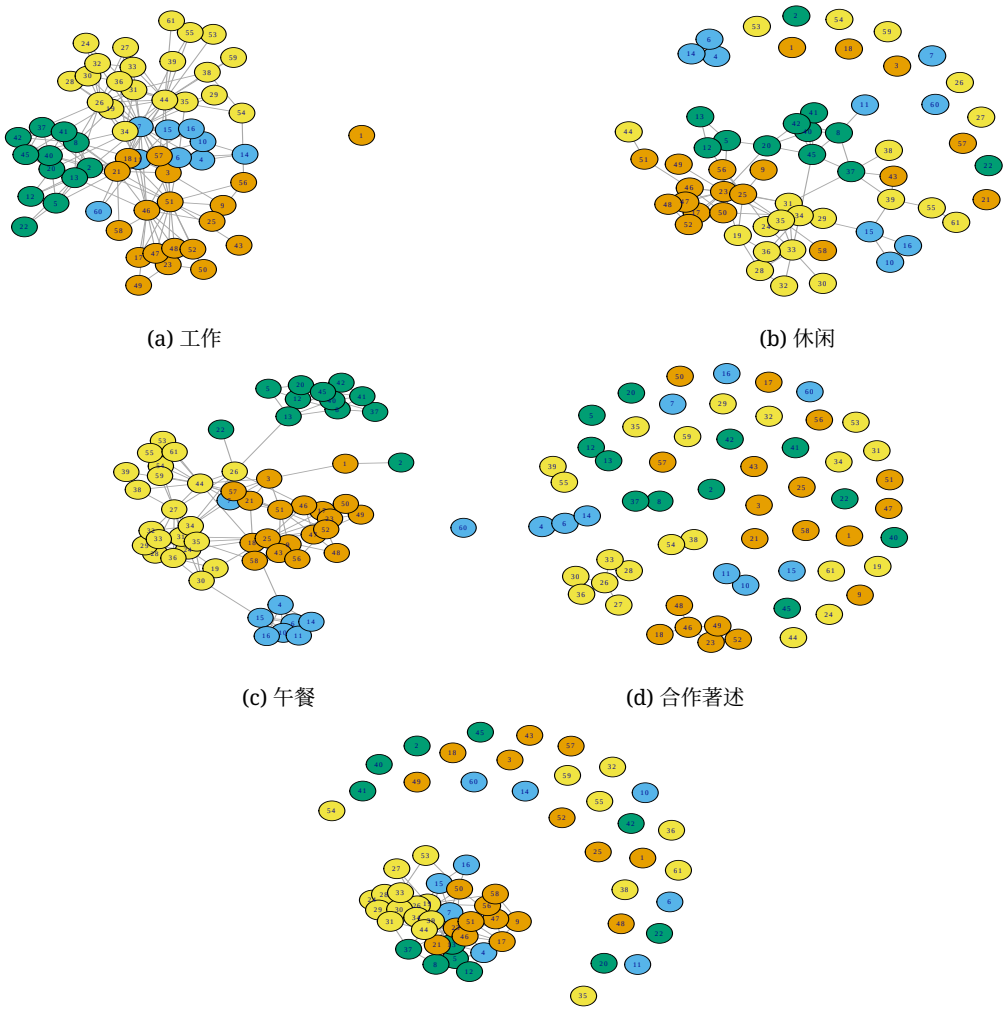
[37] Wang, Y.X.R., Bickel, P.J., 2017. 随机区块模型的基于似然的模型选择。Annals of Statistics 45，500–528。

[38] Westveld, A.H., Hoff, P.D., 2011. 用于纵向关系与网络数据的混合效应模型及其在国际贸易和冲突中的应用。《应用统计年鉴》5。

[39] Wu, Q., Hu, J., 2022年. 随机区块模型双样本检验。审稿中。

[40] Wu, Y., Lan, W., Feng, L., Tsai, C.L., 2022年. 通过最大采样逐项偏差检验随机区块模型. 手稿。

[41] Zhang, B., Li, H., Riggins, R.B., Zhan, M., Xuan, J., Zhang, Z., Hoff- man, E.P., Clarke, R., Wang, Y., 2009年. 差异依赖网络分析，用于识别生物网络中的条件特异性拓扑变化。《生物信息学》25, 526–532。



(a) 工作

(b) 休闲

(c) 午餐

(d) 合作著述

(e) 友谊

图5：五个网络的可视化。