arXiv:2303.02965v1 [math.ST] 6 Mar 2023

# Loca无标度随机图中的几何结构检测

Gianmarco Bet、Riccardo Michielan、Clara Stegehuis

2023年3月7日

**摘要**

我们研究如何检测幂律非齐次随机图中是否包含几何社区，并将其表述为一个假设检验问题。更具体地说，假设我们获得了一个样本，该样本来自图空间上定义于 $n$ 个顶点的某个未知分布。在原假设下，样本源自具有重尾度序列的非齐次随机图。在备择假设下，$k = o(n)$ 个顶点被赋予空间位置，并按照几何的非齐次随机图连接规则彼此连接。其余 $n - k$ 个顶点则遵循非齐次随机图的连接规则。我们提出了一种简单高效的检验方法，该方法基于统计归一化三角形的数量，以区分这两个假设。我们证明，当 $n \to \infty$ 时，该检验能够以高概率正确检测出社区的存在，并以高概率识别出社区中的大度数顶点。

## 1 引言

随机图为生物学、计算机科学、社会学以及许多其他科学领域中的复杂系统建模提供了统一框架。随机图范式通常包括指定一种生成图的概率机制，然后研究由此得到的网络的性质，无论是拓扑性质（连通性、聚类等），还是统计性质（对数据的拟合）。随机图尤其适合作为零模型，用于判断观测到的某个现实世界网络是否以统计显著的方式偏离其预期结构。在这一背景下，人们广泛观察到，现实世界网络具有两个显著特征：重尾度序列和高聚类系数 [13, 27]。经典的Erdős-Rényi随机图模型无法再现这两个特征，因此对于大多数应用而言，它并不是令人满意的零假设模型。因此，人们开发了备择模型，以匹配现实世界网络中观察到的度序列和聚类。所谓的非齐次随机图（非齐次随机图） [11] 是 Erdős-Rényi 随机图的一种流行推广：它为节点分配权重，并以与两个节点权重乘积成正比的概率连接它们。这样，非齐次随机图便能再现任意度序列，但聚类系数仍然较低。

获得高聚类模型的一种常见方法，是将顶点嵌入度量空间（如球面或环面）中，并以与它们之间距离成正比的概率连接顶点 [21]。事实上，由于三角不等式的作用，距离的存在使得给定顶点的两个邻居很可能彼此接近，因而也会相互连接。将非齐次随机图的顶点嵌入环面中，便可得到所谓的几何非齐次随机图 （GIRG） [9]。这种模型生成的网络具有现实世界网络中常见的两种现象：重尾度序列和高聚类。不过，聚类节点往往并非均匀分布在网络中，而是形成社群。

因此，从实际应用角度看，首先确定这些社群是否确实存在，其次识别它们，都具有重要意义。或许令人意外的是，早期关于后一个问题的文献并未探讨前一个问题 [16, 25, 26]。事实上，社区检测文献通常聚焦于从给定网络中提取社区结构的算法，而不论该结构是否真实存在。这些算法通常在社区结构已知的随机图模型上进行测试。随机块模型 [18], 便是一例；由于其数学上的易处理性，它受到了广泛关注。然而，这也付出了代价：模型采用了不切实际的假设，例如社群规模非常大（与图的规模同阶），且度分布均匀。为克服这一问题，Arias-Castro 和 Verzelen [3, 30] 研究了在 Erdős-Rényi 随机图中检测小型社区是否存在。他们找出了参数空间中（几乎必然）无法检测的区域，并给出了能够在该区域之外检测出社区的检验方法。随后， [6]将这些结果推广到了非齐次随机图。不过， [6] 中的结果不适用于重尾度序列，而且社区是通过调整一个临时设定的密度参数得到的。本文进一步推进了现实网络检测结果的研究。具体而言，我们的零假设模型是具有重尾度序列的非齐次随机图；若社区存在，则将全体节点中的少数节点嵌入环面，并按照通常的 GIRG 连接概率将它们连接起来。由于社区具有几何结构，其中包含许多三角形，这与基于非齐次随机图的社区所具有的树状结构相反。社区的这一现实特征使我们能够开发高效的检验方法，用于社区检测和识别。

更准确地说，当大小为 $n$的图中确实存在该社区时，网络中的 $n-k$ 个节点彼此之间以非几何的方式建立连接。其余 $k$ 个节点在某个几何空间中占据一定位置，彼此邻近的节点更有可能连接。这种几何设定会形成一个包含许多三角形的子图，因此可以将其视为网络中的一个社区。与植入一个团相比，这种几何结构的限制较少；与将稠密非齐次随机图作为社区相比，它也更贴近现实。此外，植入结构具有几何性质这一事实，使得我们能够通过高效的基于三角形的检验来检测并识别该结构。据我们所知，此前尚未有人研究过这类植入结构。

我们的贡献如下：

- 我们提出了一种统计检验，用于检测是否存在植入的几何社区。与其他用于稠密子图检测的检测方法不同， [6, 30], 该检验适用于异质度。该方法基于三角形，因此易于高效实现。我们没有采用标准的三角形计数，因为它可能无法区分几何网络与非几何网络；相反，我们的统计量会根据三角形对几何结构的支持证据进行加权。

- 我们提出了一种统计方法，用于识别植入的几何社区中度最大的顶点。该方法也基于三角形，因此易于高效实现。我们证明，该方法能够在所有高阶顶点中实现精确恢复。

- 我们提出了一种推断植入的几何社区规模的方法。该方法利用已识别的植入的几何社区中度最大的顶点，根据顺序统计量的收敛性估计社区规模。

- 我们通过数值结果表明，结合这些检验能够准确识别植入的几何社区。此外，这些检验仅需 $O(n^{3/2})$ 时间即可完成计算 [22]。

### 1.1 相关文献

我们的研究处于两个丰富研究方向的交汇处：社区检测，以及可称为结构检测的研究。在前一种情形中，给定一个来自已知随机图模型的样本，任务是判断其中是否存在统计上不太可能出现的稠密子图，并且可能还要将其识别出来。在这一背景下，植入团问题 作为社区检测算法的试验平台，受到了广泛关注。在该模型中，一个规模为 $n$ 的大型网络按照某种机制生成，并可能在其中植入一个规模为 $k$ 的小团 [1, 31]。开创性研究 [3, 30] 为更贴近现实的稠密社区研究奠定了基础。它们的零假设模型是（分别为稠密、稀疏的）Erdős-Rényi 随机图；若存在社区，则该社区是由顶点构成的小子集，其连接概率高于零假设模型中的连接概率。另请参见 [17]。在此基础上进一步推广， [6] 着重研究在非齐次随机图中检测稠密子图。具体而言，它们的零假设模型是非齐次随机图；而在备择假设下，小部分节点$C$ 的连接概率会乘以一个因子 $\rho_C > 1$。至关重要的是，它们的方法要求对图的非齐次性进行精确控制，因此无法适用于重尾度分布等情形。因此，我们的工作与 [6] 有两方面的区别。首先，我们考虑幂律顶点权重，这从建模角度来看更具吸引力。其次，植入结构之所以构成一个社区，是因为底层的几何结构，而非通过调整树状图的额外模型参数来实现。研究 [5]解决的问题与我们的恰好相反，即在几何随机图模型中检测平均场效应。具体而言，它们的零假设模型是几何随机图；而在备择假设下，一个小的顶点子集按照独立同分布的伯努利随机变量与其他每个顶点相连。它们给出了检测阈值以及渐近有力检验。

另一方面，在结构检测的情形下，给定一个来自未知随机图模型的样本，任务是判断该样本源自平均场模型还是结构化模型（例如几何模型）。在 [10] （另见 [12]）中，零假设模型是 Erdős-Rényi 图，备择模型是高维几何随机图。关于这一问题的最新进展，参见 [7]。 [14] 提出了一种基于小子图计数的检验，用于区分 Erdős-Rényi 图与一类一般的结构化模型，其中包括随机块模型和配置模型。最近， [8] 提出了一种区分平均场模型与吉布斯模型的检验，而 [24]提出了一种区分有几何结构和无几何结构的幂律随机图的检验。关于非齐次随机图的双样本假设检验，另见 [15] 。 [19]提出了所谓的 SCORE 算法，用于度修正块模型中的社区检测。他们的一个主要思路是，通过构造经过适当归一化的检验统计量来抵消顶点权重的影响，从而克服度分布异质性带来的统计问题。类似地， [20]研究了带有社区结构的非齐次随机图，并提出了一种基于短路径和短圈的归一化检验，用以检测是否存在多个社区。我们的研究同样基于图元（此处为三角形），但并非将所有三角形一视同仁，而是根据网络度的异质性对三角形赋予不同权重。由此得到一种稳健的统计量，可用于推断重尾网络中的社群。

### 1.2 论文结构

本文其余部分安排如下。第2节介绍模型以及检验所针对的假设。第3节给出我们提出的检验，并陈述其准确性的主要结果，随后在第4节进行讨论。最后，我们在第5节证明关于检测几何结构是否存在的主要结果，并在第6节证明关于识别几何结构的结果。

符号约定。我们采用统计检验问题的标准记号。原假设记为 $H_0$，备择假设记为 $H_1$。在 $H_0$ 下，即假设原假设成立时，事件 $E$ 的概率记为 $P_0(E) := \mathbb{P}(E \mid H_0)$。相对于该概率测度的期望值和方差分别记为 $E_0$ 和 $\mathrm{Var}_0$。另一方面，假设 $H_1$ 成立时，我们同样使用记号 $P_1$、$E_1$、$\mathrm{Var}_1$。本文始终采用标准的巴赫曼-朗道记号。若 $\lim_{n\to\infty} f(n)/g(n) = 0$，则记作 $f(n) = o(g(n))$；若 $\limsup_{n\to\infty} f(n)/g(n) < \infty$，则记作 $f(n) = O(g(n))$；若 $g(n) = O(f(n))$，则记作 $f(n) = \Omega(g(n))$。最后，若事件序列 $\{E_n\}_{n\ge 1}$ 以高概率（w.h.p.）发生，则称 $\lim_{n\to\infty} \mathbb{P}(E_n) = 1$。

## 2 模型

现在，我们将图中的社区检测问题表述为一个假设检验问题。给定一个简单图 $G = (V, E)$ 的单个样本，其中 $V = [n] := \{1, \dots, n\}$ 是节点集合，而 $E \subseteq \{(i, j) \in V \times V : i < j\}$ 是边集合。注意，根据假设，$G$ 不包含自环和重边。

**零假设模型**。在原假设 $H_0$下，图 $G$ 是非齐次随机图（IRG）模型的一个样本，其定义如下 [11]。我们为每个顶点 $i \in V$ 赋予一个权重 $w_i$，其中

$$F_n(x) = \frac{1}{n} \sum_{i \in V} \mathbb{1}_{\{w_i \le x\}}$$

表示经验权重分布。$F_n$ 也可以看作是从图中均匀随机选取一个顶点时的权重分布。我们要求权重序列满足以下假设。

**假设1**。存在 $\tau \in (2, 3)$ 和 $C$、$w_0 > 0$，使得对于所有满足 $x \le n^{1/(\tau-1)}/\log(n)$ 的 $x \ge w_0$

$$1 - F_n(x) = Cx^{1-\tau}(1 + o(1)).$$

给定权重序列 $\{w_i\}_{i\in V}$，任意边 $(i, j)$ 存在的概率为 y

$$p_{ij} = p(w_i, w_j) := \min\left(\frac{w_i w_j}{\mu n}, 1\right), \tag{2.1}$$

与所有其他边独立，其中 $\mu = w_0 \frac{\tau-1}{\tau-2}$。在引理 5.1 中，我们证明，当 $n$ 很大时，$\mu$ 渐近等于平均权重。

**备择模型**。 在备择假设下， $H_1$、 $k$ 这些顶点构成一个社区。不失一般性，我们假设它们是 $V_C$:
$= \{1, \ldots, k\} \subset V$。为方便起见，我们记$V_I := V \setminus V_C$，并将 $V_I$ 中的元素称为*A* 型 顶点，而将 $V_C$ 中的元素称为*B*类 顶点。现在让我们更精确地定义几何社区。令 $\mathcal{X} = [0,1]^d$ 为$d$维环面。我们赋予 $\mathcal{X}$ 范数

$$\|x - y\| = \sup_{i=1,\ldots,d} \min\{|x(i) - y(i)|, |1 - (x(i) - y(i))|\}, \tag{2.2}$$

其中 $x = (x(1) \ldots, x(d))$ 和 $y = (y(1) \ldots, y(d))$ 是 $\mathcal{X}$ 的元素。注意，这是与环面结构相容的通常无穷范数。我们为每个顶点 $i \in V_C$ 分配环面 $\mathcal{X}$ 中的一个（随机）位置 $\mathbf{x}_i$。形式化地说，$(\mathbf{x}_i)_{i \in V_C}$ 是一列在 $\mathcal{X}$ 上均匀分布的随机变量，我们将 $(x_i)_{i \in V_C}$ 记为该随机序列的一个实现。同样地，我们为每个顶点 $i \in V$ 分配一个权重 $w_i$，其中 $(w_i)_{i \in V}$ 是满足假设1的序列。此外，将几何社区中顶点权重的经验分布定义为

$$F_k(x) = \frac{1}{k} \sum_{i \in V_C} \mathbb{1}_{\{w_i \le x\}},$$

我们还要求 $F_k(x)$ 具有幂律尾部。

**假设2**。 令 $\tau$、 $C$、 $w_0$ 与假设*1*中的相同。那么，对于所有满足 $x \ge w_0$ 且 $x = O\left(k^{1/(\tau-1)} / \log(k)\right)$ 的 $x \ge w_0$，

$$1 - F_k(x) = C x^{1-\tau}(1 + o(1)).$$

在 $H_1$ 下，任意一条边 $(i, j) \in V_I \times V$ 是否存在，都独立于所有其他边，其概率由下式给出:

$$p_{ij} = p(w_i, w_j) := \frac{1}{1 + C_1} \min\left(\frac{w_i w_j}{\mu n}, 1\right), \tag{2.3}$$

其中 $C_1 := (1 + (\gamma - 1)^{-1}) 2^d$。也就是说，至少包含一个 A 类顶点的顶点对，其连接概率由两个端点的权重决定，与 $H_0$ 下的情况类似。此外，任意一条边 $(i, j) \in V_C \times V_C$ 是否存在，都独立于所有其他边，其概率为

$$p_{ij} = p(w_i, w_j, x_i, x_j) := \frac{1}{1 + C_1} \min\left(\frac{w_i w_j}{\mu k ||x_i - x_j||^d}, 1\right)^{\gamma}. \tag{2.4}$$

对于某个 $\gamma \in (1, \infty]$。这是一种作用于 $k$ 个顶点的几何连接概率，与 GIRG模型类似， [9], 乘以因子 $1/(1+C_1) \in (0, 1)$。按照约定，选择 $\gamma = \infty$ 对应于阈值 连接规则，即当 $||x_i - x_j||^d \le w_i w_j / (\mu k)$ 时为 $p_{ij} = 1/(1 + C_1)$，否则为 $p_{ij} = 0$。因此，这些 $k$ B 类顶点根据其权重及位置建立连接。特别是， $x_i$ 与 $x_j$越接近，它们建立连接的可能性就越大。三角不等式还确保了: B 类节点 $i$ 与 $j$ 、以及 $i$ 与 $k$ 之间存在连接时， $j$ 与 $k$ 之间存在一条边的可能性也会增加。因此，B 类顶点相比 A 类顶点更可能形成聚类。请注意，连接规则（2.4）还可以有另一种解释: 它是在 $n$ 个顶点上的 GIRG连接概率， [9], 其中顶点的位置 $V_C$ 均匀采样于（不断收缩的）环面 $[0, \lceil k/n \rceil]^d$上。

**随机性的来源**。 注意，在 $H_1$ 下存在两个随机性来源：位置序列（$x_i$）$_{i\in V_C}$，以及顶点之间的随机独立连接。给定一个网络样本后，仅观察网络连接时，生成该网络社区的位置通常是未知的。因此，我们假设对社区的位置向量一无所知。然而，当给定网络是一个非齐次随机图的实现时，网络中某个顶点的度以高概率接近其权重 [29,附录 C]。对于几何非齐次随机图，所有权重渐近大于 $\log^2(n)$ [9]的顶点同样满足这一点。因此，在我们的设定中，假设权重序列已知是合理的，因为可以从给定网络的度分布中推断出该序列。

**修正因子**。 在非齐次随机图中，任意顶点 $i$ 的期望度为 $w_i(1+o(1))$。另一方面，在 $H_1$下生成的随机图若不包含式（2.3）和（2.4）中的修正因子 $1/(1+C_1)$，就会使幂律度分布产生偏差。因此，只需简单检查度分布，就足以判断一个随机图是从$H_0$还是 $H_1$中抽样得到的。加入修正因子 $1/(1+C_1)$后，正如附录A中所证明的，在 $H_1$ 下任意顶点的期望度同样为 $w_i(1+o(1))$，从而排除了进行简单检测检验的可能性。

# 3 主要结果

本节介绍关于几何社区检测与识别的主要结果。首先，我们引入几个重要概念。检验 $\psi$ 是一个从$G$ 到 $\{0,1\}$的映射。这里，$\psi(G)=1$ 表示原假设 $H_0$ 被拒绝，且图中包含一个植入的几何社区；而 $\psi(G)=0$ 则表示相反情况。此类检验的风险定义为

$$R(\psi) := P_0(\psi(G)=0) + P_1(\psi(G)=1). \tag{3.1}$$

我们的目标是在图规模 $n$ 较大时，区分 $H_0$ 和 $H_1$ 。形式上，若一列检验 $(\psi_n)_{n\geq 1}$ 的风险趋于零，即 $\lim_{n\to\infty} R(\psi_n)=0$，则称其为渐近有力。这样一列检验能够在 $n\to\infty$的极限下正确识别底层模型。

## 3.1 检测

本节首先介绍一种用于植入的几何社区检测的渐近有力检验，即加权三角形检验。我们将使用简写 $\{i,j,k\}=\triangle$ 表示$\{(i,\ j),\ (j,\ k),\ (k,\ i)\}\subseteq E$。该检验使用加权三角形统计量

$$W(G) := \sum_{a,b,c\in V} \frac{1}{w_a w_b w_c} \mathbb{1}_{\{\{a,b,c\}=\triangle\}}. \tag{3.2}$$

因此，每个三角形的权重与其顶点权重的乘积成反比。这样一来，$W$ 就降低了由高权重顶点构成的三角形的权重。事实上，在几何和非几何随机图中，高权重顶点之间都很可能形成三角形。因此，标准的三角形计数甚至无法区分幂律几何图和非齐次随机图 [24], ，我们需要更高级的基于三角形的统计量。主要区别在于低度顶点之间形成的三角形，

而这类三角形在非几何模型中不太可能出现。当 $W(G)$ 大于某个阈值 $f(n)$ 时，加权三角形检验拒绝 $H_0$ 。形式上，加权三角形检验 $\psi_W$ 定义如下：

$$\psi_W(G) = \mathbb{1}_{\{W(G)\geq f(n)\}}. \tag{3.3}$$

下一个结果表明，在选择 $f(n)$ 时有很大的自由度，同时仍能得到渐近有力的检验：

**定理 3.1**。设 $f(n)$ 是一个函数，使得 $f(n) \to \infty$ 当 $n \to \infty$ 时，并且 $f(n) = o(k)$。那么，加权三角形检验是渐近有力的。

定理 3.1 表明，只要任意几何子集在 $n$中增长（即使增长得任意缓慢），就有可能检测到它的存在。不过，在选择阈值 $f(n)$ 时，检验统计量 (3.3) 依赖于几何规模 $k$ 的一个下界，因为定理 3.1 要求 $f(n) = o(k)$。

## 3.2 识别

现在我们关注在 $H_1$下识别几何顶点的问题。当检验拒绝$H_0$时，接下来的目标是识别属于植入几何部分的顶点。为此，令 $V_C \subseteq V$ 为几何顶点集合的一个估计量。我们假设已知植入几何社区的规模 $k$ 。为了衡量几何顶点估计量的表现，我们使用风险函数

$$R_{id}(\hat{V}_C) := \mathbb{E}_{V_C}\Big[\frac{|\hat{V}_C \triangle V_C|}{2|V_C|}\Big], \tag{3.4}$$

其中， $V_C \triangle V_C := (V \setminus V_C) \cap V_C) \cup (V_C \cap (V \setminus V_C))$ 表示 $\hat{V}_C$ 与 $V_C$ 之间的对称差，而 $\mathbb{E}_{V_C}$ 表示在已知集合 $V_C$ 的条件下的期望值。注意，当我们假设社区规模已知时，$|\hat{V}_C \triangle V_C| \leq 2|V_C|$，且 $\hat{V}_C$ 恰好输出 $k$ 个顶点，因此在这种情况下 $R_{id} \in [0,1]$。若一种方法满足 $R_{id}(V_C) \to 0$，则称其实现了**精确恢复**；若满足 $R_{id}(V_C) \to c$，其中 $c \in (0,1)$，则称其实现了**部分恢复**。换言之，当检验能够识别社区中正比例的顶点时，就实现了部分恢复。为了估计几何顶点集合，我们构造一个检验统计量 $T: V \to \{0,1\}$ ，使得当节点 $i \in V$ 被估计为属于社区时，$T(i) = 1$；否则， $T(i) = 0$。

GIRG 中的低权重顶点以正概率具有零度，而零度的 A 类和 B 型顶点无法被识别。这强烈表明，在我们的设定中，权重为 $O(1)$量级的 $O(n)$顶点无法实现精确恢复。事实上，由于低权重顶点占所有顶点的比例不趋于零，即使部分恢复也很困难。因此，我们着重在由所有高权重顶点诱导的图中实现部分恢复。

为进行识别，我们提出如下检验统计量 $T:\ V \to \{0,1\}$:

$$T(a) := \mathbb{1}_{\{W(a) > n/(w_a\sqrt{\log n})\}}, \tag{3.5}$$

其中

$$W(a) := \frac{n}{w_a^2} \sum_{b,c \in V} \frac{1}{w_b w_c} \mathbb{1}_{\{\{a,b,c\} = \triangle\}}. \tag{3.6}$$

接下来，我们证明当 $w_a = w_a(n) \to \infty$ 时，这会使第一类错误和第二类错误趋于零。

**定理 3.2**。 检验 $T(a)$ 能够在所有权重为$w_a \gg \log(n)$的顶点集合 $a$ 中实现精确恢复。形式化地，令

$$\hat{V}_C := \{a \in V : T(a) = 1\}, \tag{3.7}$$

我们得到

$$\lim_{n\to\infty} \mathbb{E}_{V_C}\left[\frac{|\hat{V}_C^{t_n} \triangle V_C^{t_n}|}{2|V_C^{t_n}|} \mid |V_C^{t_n}| > 0\right] = 0, \tag{3.8}$$

只要

$$(n\log(n)/k)^{1/\tau} \ll t_n \ll k^{1/(\tau-1)}. \tag{3.9}$$

其中，对于任意 $U \subseteq V$,

$$U^h := U \cap \{a \in V : w_a \geq h\}. \tag{3.10}$$

注意，虽然定理 3.2 在风险趋于 0 的权重阈值中用到了 $k$ 的知识，但通过选择 $t_n \gg (n\log(n))^{1/\tau}$，即可避开这个阈值，正如 $k \geq 1$ 所示。在这种情况下，只需了解上限是否也成立，也就是说，是否有 $k \gg (n\log(n))^{(\tau-1)/\tau}$；并且只需要对 $k$ 有一个足够大的下界。

### 3.3 僵尸网络规模的估计

接下来我们探讨另一个问题，即在$H_1$下推断植入的几何社区的规模。我们将充分利用这样一个事实：根据定理 3.2，上文引入的社区估计量在 $n \to \infty$ 极限下能够准确识别高阶顶点。

令 $X_{(1)}$，$X_{(2)}$，.....表示由定理 3.2 识别出的几何部分顶点权重的顺序统计量。也就是说，$X_{(1)}$ 是几何部分中度最高的顶点。因此，我们将基于节点的检验识别为 GIRG 一部分的 $m$ 个最高权重顶点作为输入。用 $X_{(1)}$，$X_{(2)}$ 等表示这些顶点的权重。我们提出以下估计量来估计社区规模 $k$：

$$\hat{k}_m := mX_{(m)}^{\tau-1}. \tag{3.11}$$

**定理3.3**。 假设 $k \gg (n\log(n))^{(\tau-1)/(2\tau-1)}$ 且 $m \in \mathbb{N}$。那么，当 $n \to \infty$ 时，

$$\hat{k}_m/k \xrightarrow{\mathbb{P}} 1. \tag{3.12}$$

定理*3.3*的证明。 令 $X_{(1)}$，$X_{(2)}$，.....表示 GIRG 部分顶点的度（权重）的顺序统计量。根据 [28, 式 (4.17)], ，当 $k \to \infty$时，

$$\frac{X_{(s)}}{\left(\frac{k}{s}\right)^{1/(\tau-1)}} \xrightarrow{\mathbb{P}} 1. \tag{3.13}$$

现在，只要存在满足 (3.9) 的某个 $t_n$ s，那么权重最高的 $m$ B 型顶点就能以趋近于 1 的概率被正确识别。只要.....，这样的序列就存在。

$$k \gg (n\log(n))^{(\tau-1)/(2\tau-1)}. \tag{3.14}$$

接下来,

$$\hat{k} = X_{(m)}^{\tau-1} m/k \xrightarrow{\mathbb{P}} 1. \tag{3.15}$$

□

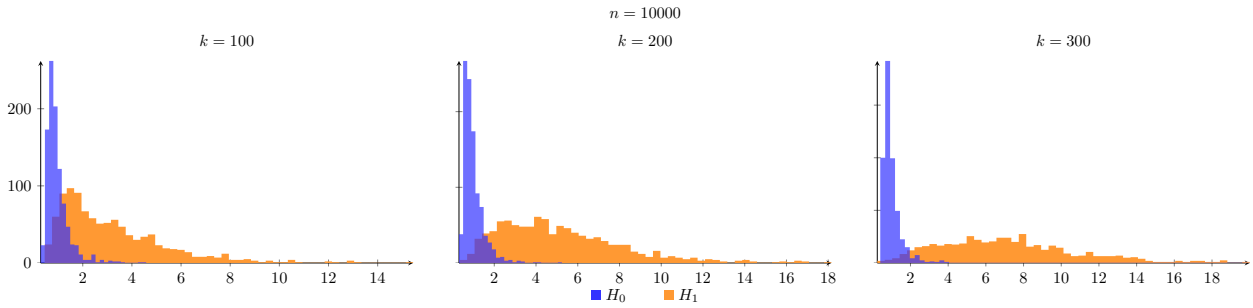



图 1：在 $H_0$（蓝色）和 $H_1$（橙色）条件下生成的 $10^4$ 个样本图中， $W$ 的值的直方图。本次模拟所选参数为 $\tau = 2.5$， $C = 1$， $w_0 = 1$， $d = 2$， $\gamma = 5$。

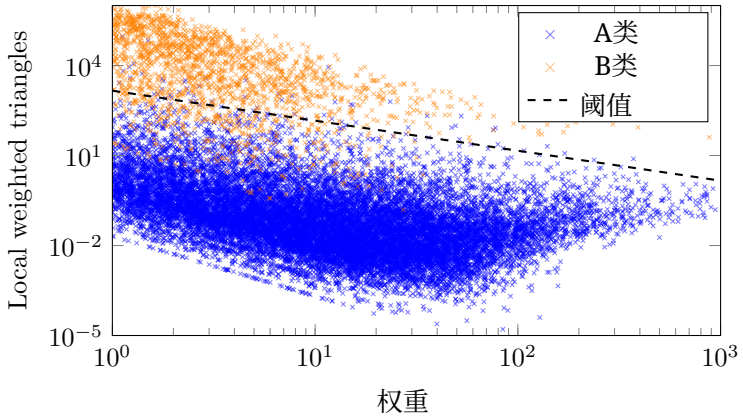



图 2：在 $H_1$ 条件下识别几何顶点。图的规模和几何社区的规模分别为 $n = 10^6$ 和 $k = 10^4$。点表示所有 A 型和 B 型顶点的局部加权三角形统计量，并以其权重为横向对照。

### 3.4 数值结果

下面我们通过一些数值实验来展示我们所提出检验的有限样本性能。

在图 1 中，我们比较了在多个零假设模型和备择模型的样本上计算得到的 $W$ 直方图（见 (3.2)）。在两种情况下，模型均基于 $n = 10^4$ 个顶点生成，几何社区的大小在 $k = 100$,200, 300 之间变化。可以看到，在 $H_0$下， $W$高度集中在其期望值附近。另一方面，在 $H_1$ 下， $W$ 的典型值更大，并且随着几何社区规模的增大而增大。这与命题 5.3 和命题 5.4 一致，也表明加权三角形检验 $\psi_W$ 即使在有限样本下也能实现较高的准确率。此外， $k$ 越大，按照 $W$ 衡量， $H_0$与 $H_1$ 之间的区分度就越高。

接下来，图 2 展示了我们的识别检验的表现。这里，我们对$H_1$ 进行了一次采样，并计算每个顶点$a$ 的局部加权三角形统计量 $W(a)$。将 $W(a)$ 对权重 $w_a$ 作图，可以看到，B 型顶点的坐标点云（$w_a$， $W(a)$) 与 A 类顶点形成的点云彼此分离。在双对数尺度上，这两个不同的区域很容易区分。图 2 中的虚线是曲线

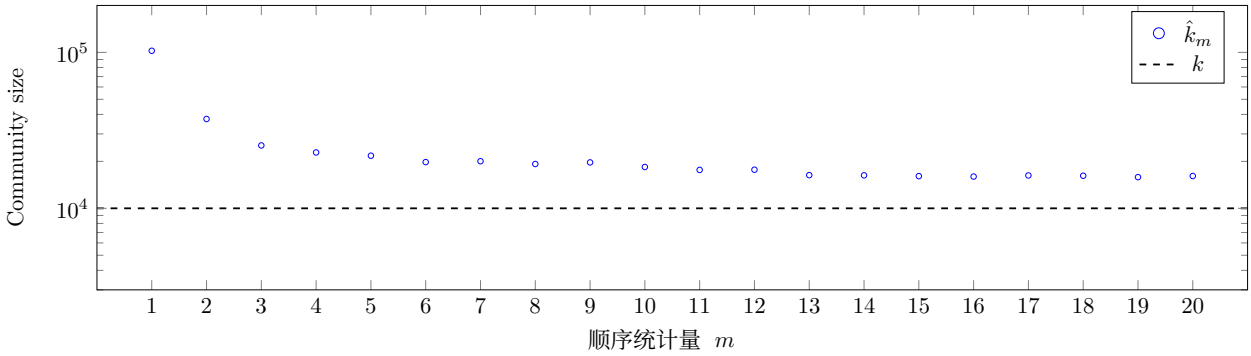



图 3：几何社区规模的估计。蓝点表示对 $k_m$ 进行 15 次模拟所得的平均值。红线表示数值 $k$，即几何社区的真实规模。此处模型选取的参数为： $n = 10^6$, $k = 10^4$, $\tau = 2.5$, $d = 1$, $\gamma = 5$, $C = 1$, $w_0 = 1$

$y = \mathcal{C}n/(x\sqrt{\log n})$，其中 $\mathcal{C}$ 是根据模型参数调整的常数值。根据定理 3.2，除一小部分权重较大的 B 型顶点外，其余顶点都位于虚线上方，而 A 类顶点则位于虚线下方。我们的模拟证实了这一点。我们还观察到，权重较低的顶点中有很大一部分被正确识别。这表明，即使定理 3.2 仅适用于高阶顶点，通过忽略局部加权三角形统计量为零的顶点，仍有可能实现对整个社区的部分恢复。

最后，图 3 展示了定理3.3 的一个实际应用。假设我们获得了模型 $H_1$ 的一个样本。使用式 (3.5) 中定义的检验 $T(a)$，并按照上述步骤，可以将上述顶点分类为 A 型或 B类。然后，对于分类为 B类的顶点的权重，我们取 $M$ 高阶顺序统计量（$X_m$）$_{m=1,\ .....,\ M}$；在本文中， $M = 20$ 。最后，根据式 (3.11) 的定义，计算社区规模估计量（$k_m$）$_{m=1,\ .....,\ M}$。对模型 $H_1$ 的多个独立样本重复这一过程（保持所有参数不变），便可对规模估计量（$k_m$）$_{m=1,\ .....,\ M}$ 取平均，以检验其表现。图 3 中的蓝点是对模型 $H_1$ 进行 15 次模拟所得到的社区规模估计量平均值，其中 $n = 10^6$, $k = 10^4$。根据定理3.3，我们知道，估计量 $k_m$ 依概率收敛到几何社区的真实规模 $k$。图 3 中无法观察到这种收敛，因为图的规模固定为 $n = 10^6$。尽管如此，该图表明，即使 $m$ 的值适中，估计量（$k_m$）$_{m} \geq 1$ 也能很好地捕捉几何社区的真实规模 $V_C$。

# 4 讨论

**计算复杂度**。 由于我们的检验是基于三角形的检验，因此只需枚举所有顶点中的三角形。此操作可在 $O(n^{1+\frac{1}{\tau}})$ [23] 或 $O(n^{3/2})$ [22] 时间内完成；若采用良好近似，则可在$O(n\log(n))$时间内完成 [4], ，从而提供一种极其高效的几何结构检测与识别方法。

**迭代过程**。 我们的识别过程能够以高概率正确识别高权重顶点。我们认为，这种识别也可以作为起点，以非平凡概率识别低权重顶点。已知某个低权重顶点的高权重邻居后，可以计算该顶点属于几何结构的可能性，并根据可能性大小识别该顶点。之后，可以利用这些更新后的识别结果再次评估这种可能性，并迭代改进对几何子集的估计。已有证明表明，这种过程适用于随机块模型 [32]，因此在这种情境下采用此类过程似乎也很有前景。

**ˆ改进** $k$。 虽然定理3.3表明，我们对 $k$ 的估计量在大网络ˆ极限下无偏，但当 $n$有限时， $k$ 会高估 $k$，如图 3 所示。这是因为该检验基于定理 3.2 所识别出的几何顶点。在大网络极限下，这些顶点都能被正确识别。然而，当 $n$有限时，部分顶点可能被误分类为几何顶点或非几何顶点。由于 $k$ 相较于 $n$较小，大多数误分类都是将非几何顶点误判为几何顶点。因此，这些误识别的顶点会使推断得到的几何顶点顺序统计量偏高，从而导致对 $k$的高估。因此，改进 $n$ 有限时的这一估计量，是一个值得进一步研究的方向。例如，可以利用误分类顶点预期数量的信息来改进该检验。

**重新缩放盒子大小**。 在我们的模型中，我们根据社区 $k$的大小重新缩放了GIRG连接概率（2.4）。这相当于在不断缩小的环面中对社区内顶点的位置进行采样，而不对连接概率进行重新缩放。另一种自然的选择是对社区内的连接概率进行重新缩放，如下所示

$$p_{ij} = \min\left(\frac{w_i w_j}{\mu n ||x_i - x_j||^d}, 1\right)^{\gamma}. \tag{4.1}$$

然而，这一连接概率会导致一个稀疏社区，其中大多数社区成员都不会与其他社区成员相连。正如其名称所示，在稀疏社区情形下，社区中给定顶点与其他社区顶点之间的平均连接数大致会按 $k/n$ 递减。正因如此，我们认为，(2.4) 的局部化社区假设是更现实的情形。不过，只要 $k \gg \sqrt{n}$，我们的检测方法同样适用于稀疏社区情形。然而，用于识别此类稀疏社区的阈值尚不明确，这将是一个值得进一步研究的方向，以探究我们方法的理论极限。−

# 5 检测：证明

本节将在假设 $H_0$ 和 $H_1$ 的条件下，为图 $\mathbb{E}[W]$ 中加权三角形的期望求出一个上界和一个下界。此外，我们还将推导加权三角形方差的一个界，并通过应用切比雪夫定理证明定理 3.1。在计算 $E[W]$ 和 Var($W$) 之前，我们先陈述并证明一些有用的引理。

**引理 5.1.** 设 $X$ 为帕累托幂律分布，其分布函数为 $F(x) = 1 - Cx^{1-\tau}$对所有 $x \geq w_0$均成立。设 $\{w_i\}_{i \in V}$ 为满足假设*1*的权重序列。那么，

*(i)* 权重之和为

$$\sum_{i\in V} w_i = \mu n(1+o(1))$$

其中 $\mu = \mathbb{E}[X]$。

*(ii)* 权重倒数之和为

$$\sum_{i\in V} w_i = \nu n(1+o(1))$$

其中 $\nu = \mathbb{E}[X^{-1}]$。

证明。令 $U$ 为顶点集 $V$ 上的均匀分布。我们注意到，$F_n$ 是从顶点 $w_U$ 中均匀选取一个顶点所得的分布函数。事实上，

$$\mathbb{P}(w_U \le x) = \sum_{i\in V} \mathbb{P}(U=i)\mathbb{1}_{\{w_i\le x\}} = \sum_{i\in V} \frac{1}{n}\mathbb{1}_{\{w_i\le x\}} = F_n.$$

此外，我们有

$$\mathbb{E}[w_U] = \frac{1}{n}\sum_{i\in V} w_i,$$
$$\mathbb{E}[w_U^{-1}] = \frac{1}{n}\sum_{i\in V} \frac{1}{w_i}.$$

$w_U$ 的期望值可通过累积分布函数计算

$$\mathbb{E}[w_U] = \int_{\mathbb{R}} x \,\mathrm{d}F_n(x) = \int_{\mathbb{R}} x(1+o(1)) \,\mathrm{d}F(x) = \mathbb{E}[X](1+o(1)), \tag{5.1}$$

with

$$\mu = \mathbb{E}[X] = \int_{w_0}^{\infty} x\cdot C(\tau-1)x^{-\tau}\,\mathrm{d}x = C\frac{\tau-1}{\tau-2}w_0^{2-\tau}. \tag{5.2}$$

同样地，我们有 $\mathbb{E}[w_U^{-1}] = \mathbb{E}[X](1+o(1))$，其中

$$\nu = \mathbb{E}[X^{-1}] = \int_{w_0}^{\infty} x^{-1}\cdot C(\tau-1)x^{-\tau}\,\mathrm{d}x = C\frac{\tau-1}{\tau}w_0^{-\tau}. \tag{5.3}$$

□

注：如果将最小权重设为 $w_0$，则帕累托随机变量 $X$ 定义中的常数 $C$ 是唯一确定的。事实上，仅当 $\int_{w_0}^{\infty} \mathrm{d}F(x) = 1$ 时，$F$ 才是一个正确的累积分布函数。这在 $C = w_0^{\tau-1}$ 时成立。

**引理 5.2**。

$$\sum_{i\in V} w_i\mathbb{1}_{\{w_i\le\sqrt{n}\}} = \mu n(1+o(1)).$$

证明。沿用引理 5.1 中的相同证明方法，我们得到

$$\begin{aligned}\frac{1}{n}\sum_{i\in V} w_i\mathbb{1}_{\{w_i\le\sqrt{n}\}} &= \int_{w_0}^{\sqrt{n}} xC(\tau-1)x^{-\tau}(1+o(1))\mathrm{d}x = [\mu - O(n^{1-\tau/2})](1+o(1))\\ &= \mu(1+o(1)).\end{aligned} \tag{5.4}$$

□

## 5.1 原假设下的加权三角形

我们现在已经准备好 dy 来计算 $H_0$ 下加权三角形的期望和方差 .

**命题 5.3**。 在 $H_0$下， $W$ 的期望值和方差为

$$\begin{aligned}\mathbb{E}_0[W] &= 1+o(1)\\ \mathrm{Var}_0(W) &\le \frac{1}{\mu^3}(1+o(1)).\end{aligned} \tag{5.5}$$

证明。 **A 部分：期望值**

$$\begin{aligned}\mathbb{E}_0[W] &= \sum_{a,b,c\in V}\mathbb{E}_0\left[\frac{\mathbb{1}_{\{\{a,b,c\}=\triangle\}}}{w_a w_b w_c}\right]\\ &= \sum_{a,b,c\in V}\frac{p_{ab}p_{bc}p_{ac}}{w_a w_b w_c}\end{aligned} \tag{5.6}$$

由于对任意 $a, b \in V$都有 $p_{ab} \le \frac{w_a w_b}{\mu n}$,

$$\begin{aligned}\mathbb{E}_0[W] &\le \left(\frac{1}{\mu n}\right)^3\sum_{a,b,c\in V} w_a w_b w_c\\ &= \left(\frac{1}{\mu n}\right)^3\left(\sum_{i\in V} w_i\right)^3\\ &= \left(\frac{1}{\mu n}\right)^3(\mu n)^3(1+o(1))\\ &= 1+o(1),\end{aligned} \tag{5.7}$$

其中，根据引理 5.1(i)， $\sum_i w_i = \mu n(1+o(1))$。为了得到下界，我们将式（5.6）中的求和限制在满足 $w_a$， $w_b$， $w_c \le \sqrt{\mu n}$的那些 $a$， $b$， $c$上。于是， $\frac{w_a w_b}{\mu n}, \frac{w_b w_c}{\mu n}, \frac{w_a w_c}{\mu n} \le 1$，并应用引理 5.2,

$$\begin{aligned}\mathbb{E}_0[W] &\ge \sum_{a:w_a\le\sqrt{\mu n}}\ \sum_{b:w_b\le\sqrt{\mu n}}\ \sum_{c:w_c\le\sqrt{\mu n}}\frac{p_{ab}p_{bc}p_{ac}}{w_a w_b w_c}\\ &= \left(\frac{1}{\mu n}\right)^3\sum_{a:w_a\le\sqrt{\mu n}}\ \sum_{b:w_b\le\sqrt{\mu n}}\ \sum_{c:w_c\le\sqrt{\mu n}} w_a w_b w_c\\ &= \left(\frac{1}{\mu n}\right)^3(n\mu)^3(1+o(1))\\ &= 1+o(1),\end{aligned} \tag{5.8}$$

这里我们应用了引理 5.2。

**B 部分：方差**

我们可以将 $W$ 的方差改写为

$$\begin{aligned}\mathrm{Var}_0(W) &= \mathrm{Var}_0\left(\sum_{a,b,c\in V}\frac{\mathbb{1}_{\{\{a,b,c\}=\triangle\}}}{w_a w_b w_c}\right)\\ &= \sum_{a_1,b_1,c_1\in V}\sum_{a_2,b_2,c_2\in V}\frac{\mathrm{Cov}_0\left(\mathbb{1}_{\{\{a_1,b_1,c_1\}=\triangle\}},\mathbb{1}_{\{\{a_2,b_2,c_2\}=\triangle\}}\right)}{w_{a_1}w_{b_1}w_{c_1}w_{a_2}w_{b_2}w_{c_2}}\end{aligned} \tag{5.9}$$

然后，利用协方差的界 $\mathsf{Cov}(X_i,\ X_j)\ \le \mathbb{E}\left[X_i X_j\right]$，

$$\begin{aligned}\mathrm{Var}_0(W) &\le \sum_{a_1,b_1,c_1\in V}\sum_{a_2,b_2,c_2\in V}\frac{\mathbb{E}_0\left[\mathbb{1}_{\{\{a_1,b_1,c_1\}=\triangle\}}\mathbb{1}_{\{\{a_2,b_2,c_2\}=\triangle\}}\right]}{w_{a_1}w_{b_1}w_{c_1}w_{a_2}w_{b_2}w_{c_2}}\\ &= \sum_{a_1,b_1,c_1\in V}\sum_{a_2,b_2,c_2\in V}\frac{P_0(a_1\leftrightarrow b_1, b_1\leftrightarrow c_1, c_1\leftrightarrow a_1, a_2\leftrightarrow b_2, b_2\leftrightarrow c_2, c_2\leftrightarrow a_2)}{w_{a_1}w_{b_1}w_{c_1}w_{a_2}w_{b_2}w_{c_2}}\end{aligned} \tag{5.10}$$

现在 六个顶点 $a_1,\ b_1,\ c_1,\ a_2,\ b_2,\ c$ 的交集现在有不同的情形 $_2$:

- 如果 $\{a_1,b_1,c_1\}$ 和 $\{a_2,b_2,c_2\}$ 不相交，或仅在一个顶点处相交，那么它们对（5.9）中方差的贡献为零。事实上，在这种情况下，两个伯努利随机变量$\mathbb{1}_{\{\{a_1,b_1,c_1\}=\triangle\}},\ \mathbb{1}_{\{\{a_2,b_2,c_2\}=\triangle\}}$ 不相关。

- 如果 $\{a_1,b_1,c_1\}$ 和 $\{a_2,b_2,c_2\}$ 在 2 个顶点处相交，在不失一般性的情况下，引入一个组合因子后，我们可以假设 $a_1=a_2\equiv a,\ b_1=b_2\equiv b$。那么，（5.10）中求和的分子有界于

$$P_0(a\leftrightarrow b, a\leftrightarrow c_1, b\leftrightarrow c_1, a\leftrightarrow c_2, b\leftrightarrow c_2)\le\frac{w_a^3w_b^3w_{c_1}^2w_{c_2}^2}{(\mu n)^5}, \tag{5.11}$$

利用以下事实：对于所有 $i,\ j,\ p_{ij}\le w_iw_j/\mu n$。因此，此类顶点对方差的贡献可界定为

$$\frac{1}{(\mu n)^5}\sum_{a,b,c_1,c_2\in V}\frac{w_a^3w_b^3w_{c_1}^2w_{c_2}^2}{w_a^2w_b^2w_{c_1}w_{c_2}}=O(n^{-1}) \tag{5.12}$$

其中等式由引理 5.1 得出。

- 如果 $\{a_1,b_1,c_1\}$ 和 $\{a_2,b_2,c_2\}$ 在全部 3 个顶点上相交，那么不失一般性地，并考虑一个组合因子，我们可以假设 $a_1=a_2\equiv a,\ b_1=b_2\equiv b,\ c_1=c_2\equiv c$。在这种情况下，(5.10) 中求和的分子可界定为

$$P_0(a\leftrightarrow b, a\leftrightarrow c, b\leftrightarrow c)\le\frac{w_a^2w_b^2w_c^2}{(\mu n)^3}. \tag{5.13}$$

因此，这类顶点对方差的贡献有如下界

$$\frac{1}{(\mu n)^3}\sum_{a,b,c\in V}\frac{w_a^2w_b^2w_c^2}{w_a^2w_b^2w_c^2}=O(1). \tag{5.14}$$

综上：

$$\mathbb{E}_0[W] = \frac{1}{6}(1 + o(1)), \tag{5.15}$$

and

$$\mathrm{Var}_0(W) = O(1). \tag{5.16}$$

□

## 5.2 备择假设下的加权三角形

接下来，我们界定备择假设$H_1$下加权三角形的期望及其方差。请记住，在 $H_1$下，顶点分为两种不同类型：A 型和 B 类，它们分别属于图的非几何部分或几何社区。备择模型中的三角形因此分为 4 种类型：

- 由 A 型顶点构成的三角形，记为 $\triangle_1$
- 由两个 A 型顶点和一个 B 类顶点构成的三角形，记为 $\triangle_2$
- 由一个 A 型顶点和两个 B 类顶点构成的三角形，记为 $\triangle_3$
- 由 B 类顶点构成的三角形，记为 $\triangle_4$

**命题 5.4**。假设 $k \equiv k(n) \to \infty$，其中 $n \to \infty$。在假设 $H_1$ 下，$W$ 的期望值和方差为

$$\begin{aligned} \mathbb{E}_1[W] &= \Omega(k) \\ \mathrm{Var}_1(W) &= O(k) \end{aligned} \tag{5.17}$$

证明。**A 部分：期望值**

我们可以将 $W$ 的期望值拆分为对所有可能三角形类型求和：

$$\mathbb{E}_1[W] = \mathbb{E}_1\left[\sum_{\alpha=1}^{4}\sum_{\{a,b,c\}\subset V}\frac{\mathbb{1}_{\{\{a,b,c\}\in\triangle_\alpha\}}}{w_a w_b w_c}\right] = \sum_{\alpha=1}^{4}\sum_{\{a,b,c\}\subset V}\frac{P_1(\{a,b,c\}\in\triangle_\alpha)}{w_a w_b w_c} \tag{5.18}$$

接下来，我们给出 $W$ 期望值的下界，其中包含来自 $\triangle_4$ 的贡献：

$$\mathbb{E}_1[W] \geq \sum_{\{a,b,c\}\subset V}\frac{P_1(\{a,b,c\}\in\triangle_4)}{w_a w_b w_c} = \sum_{\{a,b,c\}\subset V_C}\frac{P_1(a\leftrightarrow b, b\leftrightarrow c, c\leftrightarrow a)}{w_a w_b w_c} \tag{5.19}$$

现在固定 $\{a,b,c\}\subset V_C$。由于 $a$、$b$、$c$ 是 B 型顶点，

$$\begin{aligned} &P_1(a\leftrightarrow b, b\leftrightarrow c, c\leftrightarrow a) = \mathbb{E}_{\overline{x}}[p(w_a,w_b,x_a,x_b)p(w_b,w_c,x_b,x_c)p(w_a,w_c,x_a,x_c)] \\ &= \int_{(\mathbb{T}^d)^3}\left(\frac{w_a w_b}{\mu k||x_a-x_b||^d}\wedge 1\right)^\gamma\left(\frac{w_b w_c}{\mu k||x_b-x_c||^d}\wedge 1\right)^\gamma\left(\frac{w_a w_c}{\mu k||x_a-x_c||^d}\wedge 1\right)^\gamma \mathrm{d}\overline{x} \end{aligned} \tag{5.20}$$

其中，$\bar{}x = \{x_a, x_b, x_c\}$ 和$\mathbb{E}_{\bar{}x}$是对位置 $x$求得的期望。注意到

$$\begin{aligned} ||x_b - x_c|| &\leq ||x_b - x_a|| + ||x_a - x_c|| \\ &\leq 2\max\{||x_a - x_b||, ||x_a - x_c||\}. \end{aligned} \tag{5.21}$$

因此，

$$\begin{aligned}\frac{1}{||x_b - x_c||^d} &\geq \frac{1}{2^d \max\{||x_a - x_b||^d, ||x_a - x_c||^d\}} \\ &= \min\left\{\frac{1}{2^d||x_a - x_b||^d}, \frac{1}{2^d||x_a - x_c||^d}\right\}.\end{aligned} \tag{5.22}$$

然后，利用　代换 $\{x'_a, x'_b, x'_c\} = \{x_a, x_b - x_a, x_c - x_a\}$，(5.20) 的下界为　y

$$\int_{(\mathbb{T}^d)^2} \left(\frac{w_a w_b}{\mu k||x'_b||^d} \wedge 1\right)^\gamma \left(\frac{w_a w_c}{\mu k||x'_c||^d} \wedge 1\right)^\gamma \left(\frac{w_b w_c}{\mu k 2^d||x'_b||^d} \wedge \frac{w_b w_c}{\mu k 2^d||x'_c||^d} \wedge 1\right)^\gamma \mathrm{d}x'_b \, \mathrm{d}x'_c \tag{5.23}$$

接下来，我们将环面 $\mathbb{T}^d$ 表示为区间 $[-\frac{1}{2}, \frac{1}{2}]^d$，并考虑以 0 为中心的立方体

$$\mathcal{C} = \left[-\left(\frac{w_0^2}{\mu k 2^d}\right)^{1/d}, \left(\frac{w_0^2}{\mu k 2^d}\right)^{1/d}\right]^d, \tag{5.24}$$

其中 $w_0$ 是模型中的最小权重。若 $x'_b$，$x'_c \in \mathcal{C}$，则被积函数 (5.23) 中的所有最小值均等于 1。因此，若将积分域限制在子集 $\mathcal{C}^2$ 上，则 (5.23) 的下界为 $\mathcal{C}^2$ 的体积。综上，

$$P_1(a \leftrightarrow b, b \leftrightarrow c, c \leftrightarrow a) \geq \int_{\mathcal{C}^2} \mathrm{d}x'_b \, \mathrm{d}x'_c = \left(\frac{w_0^2}{\mu k}\right)^2 \tag{5.25}$$

因此，

$$\begin{aligned}\mathbb{E}_1[W] &\geq \sum_{\{a,b,c\} \subset V_C} \frac{P_1(a \leftrightarrow b, b \leftrightarrow c, c \leftrightarrow a)}{w_a w_b w_c} \\ &\geq \left(\frac{w_0^2}{\mu k}\right)^2 \sum_{\{a,b,c\} \subset V_C} \frac{1}{w_a w_b w_c} \\ &= \left(\frac{w_0^2}{\mu k}\right)^2 \Theta(k^3) \\ &= \Theta(k),\end{aligned} \tag{5.26}$$

其中，在等式中我们应用了引理 5.1(ii)。综上，$\mathbb{E}_1[W] = \Omega(k)$。

**B 部分：方差**

我们可以如下给出 $W$ 在 $H_1$ 下方差的上界：

$$\begin{aligned}\mathrm{Var}_1(W) &= \mathrm{Var}_1\left(\sum_{a,b,c\in V}\frac{\mathbb{1}_{\{\{a,b,c\}=\triangle\}}}{w_a w_b w_c}\right)\\ &= \sum_{a_1,b_1,c_1\in V}\ \sum_{a_2,b_2,c_2\in V}\frac{\mathrm{Cov}_1\left(\mathbb{1}_{\{\{a_1,b_1,c_1\}=\triangle\}},\mathbb{1}_{\{\{a_2,b_2,c_2\}=\triangle\}}\right)}{w_{a_1}w_{b_1}w_{c_1}w_{a_2}w_{b_2}w_{c_2}}\\ &\le \sum_{a_1,b_1,c_1\in V}\ \sum_{a_2,b_2,c_2\in V}\frac{\mathbb{E}_1\left[\mathbb{1}_{\{\{a_1,b_1,c_1\}=\triangle\}}\mathbb{1}_{\{\{a_2,b_2,c_2\}=\triangle\}}\right]}{w_{a_1}w_{b_1}w_{c_1}w_{a_2}w_{b_2}w_{c_2}}\\ &= \sum_{a_1,b_1,c_1\in V}\ \sum_{a_2,b_2,c_2\in V}\frac{P_1(\{a_1,b_1,c_1\}=\triangle,\{a_2,b_2,c_2\}=\triangle)}{w_{a_1}w_{b_1}w_{c_1}w_{a_2}w_{b_2}w_{c_2}}.\end{aligned}\tag{5.27}$$

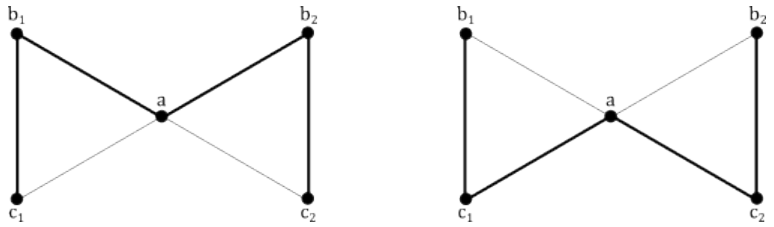
式 (5.27) 最后一项求和中的各项取决于 $\{a_1,b_1,c_1\}$ 和 $\{a_2,b_2,c_2\}$ 的顶点类型，因为顶点类型决定了连接概率。

首先，考虑 $\{a_1,b_1,c_1\}$ 和 $\{a_2,b_2,c_2\}$ 都至多包含一个 B 型顶点的情况。在这种情况下，所有连接都是非几何的。因此，根据命题 6.1 的证明（该证明给出了非几何三角形方差的界），这组顶点对方差的贡献为 $O(1)$。

此外，对于 $\{a_1,b_1,c_1\}$ 和 $\{a_2,b_2,c_2\}$ 不相交的顶点集合，随机变量 $\mathbb{1}_{\{\{a_1,b_1,c_1\}=\triangle\}}$ 和 $\mathbb{1}_{\{\{a_2,b_2,c_2\}=\triangle\}}$ 是独立的伯努利随机变量，它们对方差的贡献为 0。

因此，我们只关注以下情况：集合 $\{a_1,b_1,c_1\}$ 和 $\{a_2,b_2,c_2\}$ 中至少有一个包含两个或三个 B类顶点，且两个集合至少相交于一个顶点。

- 若 $|\{a_1,b_1,c_1\}\cap\{a_2,b_2,c_2\}|=1$，不失一般性，忽略一个组合因子，$a_1=a_2\equiv a$。在这种情况下，路径 ($c_1$、$b_1$、$a$、$b_2$、$c_2$) 或 ($b_1$、$c_1$、$a$、$c_2$、$b_2$) 中至少有一条包含至少一条几何边，也就是说，其中至少有一对相邻的B型顶点。设




路径 ($c_1$、$b_1$、$a$、$b_2$、$c_2$) 包含 $m\ge 1$ 条几何边（另一种情况的证明类似）。那么，(5.27) 的求和中的概率有界于

$$\begin{aligned}P_1(\{a,b_1,c_1\}=\triangle,\{a,b_2,c_2\}=\triangle) &\le P_1(c_1\leftrightarrow b_1\leftrightarrow a\leftrightarrow b_2\leftrightarrow c_2)\\ &\le \left(\frac{1}{1+C_1}\right)^4\frac{w_a^2w_{b_1}^2w_{b_2}^2w_{c_1}w_{c_2}}{\mu^4k^m n^{4-m}},\end{aligned}\tag{5.28}$$

其中，最后一个不等式源于这样一个事实：四条边中有 $m$ 条遵循几何连接规则，再结合引理 A.1 和引理 B.1。由于其中有 $m$ 条几何边

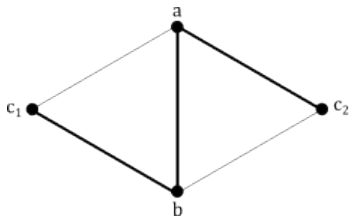
在该路径中，至少有 $m+1$ 个顶点 $a$，$b_1$，$b_2$，$c_1$，$c_2$需要属于 B 类。于是，$m+1$ 个顶点至多有 $k$ 种不同的选取方式，其余顶点至多有 $n$ 种不同的选取方式。因此，在这种情况下，式 (5.27) 中所有可能的顶点组合所作出的贡献为

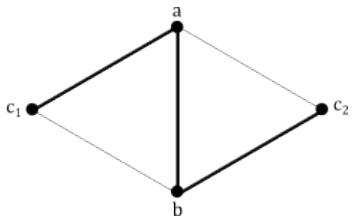
$$\begin{aligned}
&\sum_{\substack{a,b_1,b_2,c_1,c_2\in V\\ m+1 \text{ are type-B}}} \frac{P(\{a,b_1,c_1\}=\triangle,\{a,b_2,c_2\}=\triangle)}{w_a^2 w_{b_1} w_{c_1} w_{b_2} w_{c_2}} \\
&\le \sum_{\substack{a,b_1,b_2,c_1,c_2\in V\\ m+1 \text{ are type-B}}} O\left(\frac{w_a^2 w_{b_1}^2 w_{b_2}^2 w_{c_1} w_{c_2}}{k^m n^{4-m} w_a^2 w_{b_1} w_{c_1} w_{b_2} w_{c_2}}\right) \\
&= \sum_{\substack{a,b_1,b_2,c_1,c_2\in V\\ m+1 \text{ are type-B}}} O\left(\frac{w_{b_1} w_{b_2}}{k^m n^{4-m}}\right) \\
&= O(k), \qquad (5.29)
\end{aligned}$$

其中，最后一个等式由引理 5.1 得出。

对 $m=1$,2,3,4 的所有可能取值求和后，我们得到：交于一个顶点的三元组 $\{a_1,b_1,c_1\}$，$\{a_2,b_2,c_2\}$ 对方差的贡献为 $O(k)$。

- 若 $|\{a_1,b_1,c_1\}\cap\{a_2,b_2,c_2\}|=2$，不失一般性，设 $a_1=a_2\equiv a$ 且 $b_1=b_2\equiv b$。在这种情况下，路径 $(c_1,b,a,c_2)$ 或 $(c_1,a,b,c_2)$ 至少有一条包含一条几何边。假设路径 $(c_1,b,a,c_2)$ 包含 $m\ge 1$ 条几何边。那么，与前面类似，我们






可以将 (5.27) 中的概率上界估计为

$$\begin{aligned}
P_1(\{a,b,c_1\}=\triangle,\{a,b,c_2\}=\triangle) &\le P_1(c_1\leftrightarrow b\leftrightarrow a\leftrightarrow c_2) \\
&\le \left(\frac{1}{1+C_1}\right)^3 \frac{w_a^2 w_b^2 w_{c_1} w_{c_2}}{\mu^3 k^m n^{3-m}}. \qquad (5.30)
\end{aligned}$$

和之前一样，$m+1$ 个顶点需要属于 B 类。因此，在这种情况下，(5.27) 中所有可能的顶点组合所产生的贡献仍为

$$\begin{aligned}
\sum_{\substack{a,b,c_1,c_2\in V\\ m+1 \text{ are type-B}}} \frac{P_1(\{a,b,c_1\}=\triangle,\{a,b,c_2\}=\triangle)}{w_a^2 w_b^2 w_{c_1} w_{c_2}} &\le \sum_{\substack{a,b,c_1,c_2\in V\\ m+1 \text{ are type-B}}} O\left(\frac{w_a^2 w_b^2 w_{c_1} w_{c_2}}{k^m n^{3-m} w_a^2 w_b^2 w_{c_1} w_{c_2}}\right) \\
&= \sum_{\substack{a,b,c_1,c_2\in V\\ m+1 \text{ are type-B}}} O\left(\frac{1}{k^m n^{4-m}}\right) \\
&= O(k). \qquad (5.31)
\end{aligned}$$

对 $m = 1,2,3$ 的所有可能取值求和后，我们得到：两个顶点相交的三元组 $\{a_1, b_1, c_1\}$, $\{a_2, b_2, c_2\}$ 对方差的贡献为 $O(k)$。

- 若 $|\{a_1, b_1, c_1\} \cap \{a_2, b_2, c_2\}| = 3$，不失一般性地，$a_1 = a_2 \equiv a$，$b_1 = b_2 \equiv b$，$c_1 = c_2 \equiv c$。在这种情况下，路径（$a$，$b$，$c$）或（$a$，$c$，$b$）中至少有一条包含一对相邻的B型顶点。假设这条路径是（$a$，$b$，$c$），且具有 $m$ 条几何边。按照之前相同的步骤




，我们将（5.27）中的概率界定为

$$P_1(\{a,b,c\} = \triangle) \le P_1(a \leftrightarrow b \leftrightarrow c) \tag{5.32}$$

$$\le \left(\frac{1}{1+C_1}\right)^2 \frac{w_a w_b^2 w_c}{\mu^2 k^m n^{2-m}}. \tag{5.33}$$

由于这三个顶点中至少有 $m+1$ 个是B型顶点，因此这种情况下所有可能顶点组合对（5.27）的贡献再次为

$$\begin{aligned}\sum_{\substack{a,b,c\in V\\ m+1 \text{ are type-B}}} \frac{P_1(\{a,b,c\} = \triangle)}{w_a^2 w_b^2 w_c^2} &\le \sum_{\substack{a,b,c\in V\\ m+1 \text{ are type-B}}} O\left(\frac{w_a w_b^2 w_c}{k^m n^{2-m} w_a^2 w_b^2 w_c^2}\right)\\ &= \sum_{\substack{a,b,c\in V\\ m+1 \text{ are type-B}}} O\left(\frac{1}{k^m n^{2-m} w_a w_c}\right)\\ &= O(k). \end{aligned} \tag{5.34}$$

然后，对 $m = 1$，2 的所有可能取值求和，我们发现，交于三个顶点的三元组 $\{a_1, b_1, c_1\}$、$\{a_2, b_2, c_2\}$ 对方差的贡献为 $O(k)$。

综上所有可能的情形，我们得出方差为 $O(k)$。

□

定理 *3.1* 的证明。由命题 5.3 和切比雪夫不等式可知，

$$P_0(W(G) > f(n)) \le \frac{\mathrm{Var}_0(W(G))}{f(n)^2} \longrightarrow 0. \tag{5.35}$$

此外，根据命题 5.4，当 $n$ 充分大时，存在一个常数 $M$ 使得

$$P_1(W(G) < f(n)) \le P_1(W(G) - \mathbb{E}[W(G)] < f(n) - Mk). \tag{5.36}$$

然后，再由切比雪夫不等式可知，

$$P_1(W(G) < f(n)) \le \frac{\mathrm{Var}_1(W(G))}{(f(n) - Mk)^2} \to 0, \tag{5.37}$$

根据命题 5.4。 □

# 6 局部化三角形统计量：定理 3.2 的证明

现在我们将证明，在备择假设 $H_1$ 下，网络的几何社区中有正比例的高阶顶点可以通过局部化三角形统计量加以区分。

## 6.1 非齐次随机图中顶点的局部化加权三角形

**命题 6.1**。设 $a$ 是一个 $A$ 类顶点。当 $k = o(n)$ 时，

$$\mathbb{E}_1[W(a)] \le \frac{1}{2\mu}(1 + o(1)). \tag{6.1}$$

证明。我们可以写成

$$\begin{aligned}\mathbb{E}_1[W(a)] &= \frac{n}{2w_a^2}\sum_{b,c\in V}\mathbb{E}_1\left[\frac{1}{w_b w_b}\mathbb{1}_{\{\{a,b,c\}=\triangle\}}\right] \\ &= \frac{n}{2w_a^2}\sum_{b,c\in V}\frac{p_{ab}p_{bc}p_{ac}}{w_b w_c},\end{aligned} \tag{6.2}$$

因为边 $\{a,b\}$，$\{b,c\}$，$\{a,c\}$ 在图中是独立出现的。由于 $a$ 是 A 类顶点，我们可以对连接概率 $p_{ab} \le w_a w_b/\mu n$ 和 $p_{ac} \le w_a w_c/\mu n$ 给出上界。当 $b$ 或 $c$ 中至少有一个是 A 类顶点时，连接概率 $p_{bc}$ 有上界 $w_b w_c/\mu n$。在这种情况下，

$$\begin{aligned}\frac{n}{2w_a^2}\sum_{\substack{b,c\in V\\ b \text{ or } c \text{ type-A}}}\frac{p_{ab}p_{bc}p_{ac}}{w_b w_c} &\le \frac{n}{2w_a^2}\sum_{b,c\in V}\frac{w_a^2 w_b w_c}{(\mu n)^3} \\ &= \frac{1}{2\mu^3 n^2}\sum_{b,c\in V} w_b w_c \\ &\le \frac{1}{2\mu^3 n^2}(\mu n)^2(1+o(1)) = \frac{1}{2\mu}(1+o(1)).\end{aligned} \tag{6.3}$$

另一方面，当 $b$ 和 $c$ 都是 B 类顶点时，连接概率 $p_{bc}$ 为$O(w_b w_c/k)$。于是，

$$\begin{aligned}\frac{n}{2w_a^2}\sum_{\substack{b,c\in V\\ b \text{ and } c \text{ type-B}}}\frac{p_{ab}p_{bc}p_{ac}}{w_b w_c} &\le \frac{n}{2w_a^2}\sum_{b,c\in V\setminus[n-k]} O\left(\frac{w_a^2 w_b w_c}{\mu^3 n^2 k}\right) \\ &= O\left(\frac{k}{2\mu n}\right) = o(1).\end{aligned} \tag{6.4}$$

□

## 6.2 几何社区中节点的基于节点的统计量

我们现在共同考虑几何社区中一个节点的基于节点的三角形统计量。

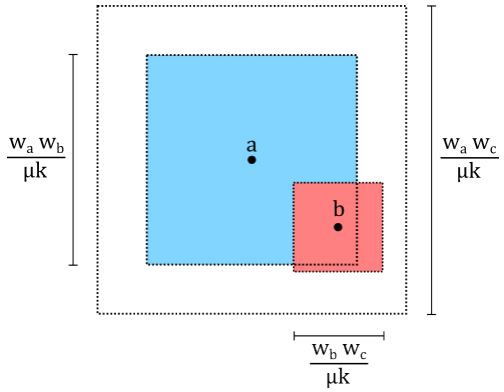



图4：当 $d=2$ 时，方程6.8中概率的直观描述。

**命题 6.2**。 假设 $a$ 是一个 $B$型顶点。当 $k=o(n)$ 且 $k\to\infty$ 作为 $n\to\infty$ 时，

$$\mathbb{E}_1\left[W(a)\right]=\Omega(n/w_a) \tag{6.5}$$

证明。 假设 $w_a\gg\log(n)$。我们仅考虑由以下三角形构成的基于节点的统计量：其中一个顶点为 $a$，且该三角形的所有其他顶点 $b$，$c$ 也都是 B 型顶点。

$$\begin{aligned}\mathbb{E}_1\left[W(a)\right]&\geq\frac{n}{2w_a^2}\sum_{b,c\in V_C}\mathbb{E}_1\left[\frac{1}{w_bw_c}\mathbb{1}_{\{\{a,b,c\}=\triangle\}}\right]\\&=\frac{n}{2w_a^2}\sum_{b,c\in V_C}\frac{P_1(\{a,b,c\}=\triangle)}{w_bw_c}\\&=\frac{n}{2w_a^2}\sum_{b,c\in V_C}\frac{\mathbb{E}_x\left[p_{ab}p_{bc}p_{ac}\right]}{w_bw_c}\end{aligned} \tag{6.6}$$

其中，对于任意 $i$，$j$ B 型顶点，$p_{ij}$ 是式(2.4)中定义的连接概率。注意，对于任意 $i$，$j$ B 型顶点，我们可以将 $p_{ij}$ 下界估计为

$$p_{ij}\geq\tilde{p}_{ij}:=\begin{cases}\frac{1}{1+C_1},&\text{if }||x_i-x_j||^d\leq w_iw_j/\mu k\\0,&\text{otherwise.}\end{cases} \tag{6.7}$$

因此，由式(6.6)可得

$$\mathbb{E}_1\left[W(a)\right]\geq\frac{n}{2w_a^2(1+C_1)^3}\sum_{b,c\in V_C}\frac{\mathbb{P}_1\left(||x_a-x_b||^d\leq\frac{w_aw_b}{\mu k},||x_b-x_c||^d\leq\frac{w_bw_c}{\mu k},||x_a-x_c||^d\leq\frac{w_aw_c}{\mu k}\right)}{w_bw_c}. \tag{6.8}$$

假设 $w_b<w_c\leq\log(n)$，后一个求和中的概率与 $\frac{w_aw_b}{\mu k}\frac{w_bw_c}{\mu k}=w_aw_b^2w_c\ (\mu k)^2$ 成正比。为了直观地理解后一个陈述，请看图4中的示意图：选定 $a$ 的位置；当 $b$ 的位置位于蓝色正方形内时，$a$ 会连接到 $b$；接下来，当 $c$ 的位置处于红色正方形与白色正方形的交集内时，它会同时连接到 $a$ 和 $b$，该交集的体积与红色正方形成正比（后者渐近地更小）。

于是，对于某个 $K>0$,

$$\begin{aligned}\mathbb{E}_1[W(a)] &\geq \frac{nK}{2w_a^2(1+C_1)^3} \sum_{\substack{b,c\in V_C\\ w_b<w_c\leq\log(n)}} \frac{1}{w_b w_c}\frac{w_a w_b^2 w_c}{(\mu k)^2} \\ &\geq \frac{nK}{2w_a(1+C_1)^3} \sum_{\substack{b,c\in V_C\\ w_b<w_c\leq\log(n)}} \frac{w_0}{(\mu k)^2} \\ &= \frac{w_0}{(1+C_1)^3\mu^2}\frac{nK}{w_a}(1+o(1)).\end{aligned} \tag{6.9}$$

□

## 6.3 局部统计量的方差

**命题 6.3.** 假设 $k=o(n)$，且当 $n\to\infty$ 时 $k\to\infty$。对于任意 $a$,

$$\mathrm{Var}_1(W(a)) = O(n^2/w_a^3). \tag{6.10}$$

证明. 在任意顶点 $a$ 上计算得到的局部统计量的方差为

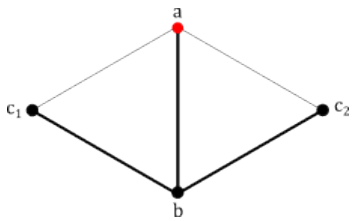
$$\begin{aligned}\mathrm{Var}_1(W(a)) &= \frac{n^2}{w_a^4}\sum_{b_1,c_1,b_2,c_2}\mathrm{Cov}_1\left(\frac{\mathbb{1}_{\triangle_{a,b_1,c_1}}}{w_{b_1}w_{c_1}},\frac{\mathbb{1}_{\triangle_{a,b_2,c_2}}}{w_{b_2}w_{c_2}}\right) \\ &= \frac{n^2}{w_a^4}\sum_{b_1,c_1,b_2,c_2}\frac{P_1(\triangle_{a,b_1,c_1},\triangle_{a,b_2,c_2})-P(\triangle_{a,b_1,c_1})P(\triangle_{a,b_2,c_2})}{w_{b_1}w_{b_2}w_{c_1}w_{c_2}} \\ &\leq \frac{n^2}{w_a^4}\sum_{b_1,c_1,b_2,c_2}\frac{P_1(\triangle_{a,b_1,c_1},\triangle_{a,b_2,c_2})}{w_{b_1}w_{b_2}w_{c_1}w_{c_2}}.\end{aligned} \tag{6.11}$$

由 (6.11) 的第二行可知，对于顶点 $\{b_1,c_1\}$、$\{b_2,c_2\}$ 的不相交组合，其对方差的贡献始终为零，因为随机变量 $\triangle_{a、b_1、c_1}$、$\triangle_{a、b_2、c_2}$ 相互独立。

因此，我们只关注 (6.11) 上界中 $\{b_1,c_1\}$、$\{b_2,c_2\}$ 相交时的求和项。

- 如果 $|\{b_1,c_1\}\cap\{b_2,c_2\}|=1$，不失一般性（忽略一个组合因子），我们可以假设 $b_1=b_2\equiv b$。注意，(6.11) 求和项中的分子还可以进一步上界为

$$P_1(\triangle_{a,b,c_1},\triangle_{a,b,c_2})\leq P_1(b\leftrightarrow a, b\leftrightarrow c_1, b\leftrightarrow c_2). \tag{6.12}$$



根据顶点 $a$、$b$、$c_1$、$c_2$ 的类型，顶点对 $(a,b)$、$(b,c_1)$、$(b,c_2)$ 可能是几何的，也可能不是。

假设这些顶点对中有 $m$ 个是几何对（当 $m=0$ 时，其中没有几何对）。那么至少有 $m$ 个顶点 $b,c_1,c_2$ 是 B 类。这些 $m$ 个顶点最多有 $k$ 种不同的选法，其余顶点最多有 $n$ 种不同的选法。由引理 A.1 可知，每个几何对 $i,j$ 的连接概率为 $p_{ij}=\Theta(1\wedge w_iw_j/k=O(w_iw_j/k)$。因此，根据 (6.12) 并应用引理 B.1，可得由满足条件 $m$ 的顶点组合对 (6.11) 的贡献，其中 $b,c_1,c_2$ 中的顶点为 B类:

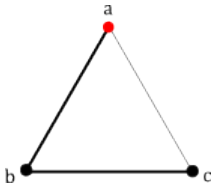
$$\begin{aligned}\frac{n^2}{w_a^4}\sum_{\substack{b,c_1,c_2\\ m\text{ are type-B}}}\frac{P_1(\triangle_{a,b,c_1},\triangle_{a,b,c_2})}{w_b^2w_{c_1}w_{c_2}}&\le\frac{n^2}{w_a^4}\sum_{\substack{b,c_1,c_2\\ m\text{ are type-B}}}\frac{O(w_aw_b^3w_{c_1}w_{c_2})}{k^mn^{3-m}w_b^2w_{c_1}w_{c_2}}\\&\le\frac{n^2}{w_a^4}\binom{3}{m}O(w_a)=O(n^2/w_a^3).\end{aligned}\tag{6.13}$$

因此，对 $m=0$,1,2,3 的所有可能取值求和，可得

$$\frac{n^2}{w_a^4}\sum_{b,c_1,c_2}\frac{P_1(\triangle_{a,b,c_1},\triangle_{a,b,c_2})}{w_b^2w_{c_1}w_{c_2}}=O(n^2/w_a^3).\tag{6.14}$$

- 若 $|\{b_1,c_1\}\cap\{b_2,c_2\}|=2$，不失一般性，我们可以假设 $b_1=b_2\equiv b$ 和 $c_1=c_2\equiv c$。公式 (6.11) 求和项中的分子还可以进一步上界为

$$P_1(\triangle_{a,b,c})\le P_1(a\leftrightarrow b,b\leftrightarrow c)\tag{6.15}$$




与之前类似，连接概率可能是几何的，也可能是非几何的，具体取决于顶点的类型。令 $m$ 表示 $(a,b)$,$(b,c)$ 中几何的顶点对数量。如前一步所示，由 (6.15) 并应用引理 5.1-B.1 可得

$$\begin{aligned}\frac{n^2}{w_a^4}\sum_{b,c}\frac{P_1(\triangle_{a,b,c})}{w_b^2w_c^2}&\le\frac{n^2}{w_a^4}\sum_{\substack{b,c\\ m\text{ are type-B}}}\frac{O(w_aw_b^2w_c)}{k^mn^{2-m}w_b^2w_c^2}\\&\le\frac{n^2}{w_a^4}\binom{2}{m}O(w_a)=O(n^2/w_a^3).\end{aligned}\tag{6.16}$$

因此，对 $m=0$,1,2 的所有可能取值求和，可得

$$\frac{n^2}{w_a^4}\sum_{b,c}\frac{P_1(\triangle_{a,b,c})}{w_b^2w_c^2}=O(n^2/w_a^3).\tag{6.17}$$

因此，对集合 $\{b_1,c_1\}$、$\{b_2,c_2\}$ 的所有可能交集求和后，我们得出公式 (6.11) 的右侧为 $O(n^2/w_a^3)$。□

## 6.4 基于节点的检测检验

定理 *3.2* 的证明。假设 $w_a \gg \log(n)$。

我们首先计算第一类错误。当 $a$ 为 A 型时，根据命题 6.1 和命题 6.3，可得 $\mathbb{E}_1[W(a)] = O(1)$，$\mathrm{Var}(W(a)) = O(n^2/w_a^3)$。根据切比雪夫不等式，

$$P\left(W(a) > \frac{n}{w_a\sqrt{\log(n)}}\right) \leq \frac{\mathrm{Var}(W(a))}{\left(\frac{n}{w_a\sqrt{\log(n)}}(1+o(1))\right)^2} = O\left(\frac{\log(n)}{w_a}\right). \tag{6.18}$$

令 $\tilde{n}$ 表示权重大于或等于 $t_n$的顶点数， $k$ 表示权重大于或等于 $t_n$的 B 类顶点数。对于第二类错误，根据命题 6.2 和命题 6.3，可得$\mathbb{E}_1[W(a)] = \Omega(n/w_a)$，$\mathrm{Var}(W(a)) = O(n^2/w_a^3)$。对于任意 B类顶点 $a$。然后，根据切比雪夫不等式

$$\begin{aligned} P\left(W(w_i) < \frac{n}{w_a\sqrt{\log(n)}}\right) &\leq \frac{\mathrm{Var}(W(a))}{\left(\frac{n}{w_a\sqrt{\log(n)}} - \mathbb{E}_1[W(a)]\right)^2} \\ &= O\left(\frac{n^2/w_a^3}{(n/w_a)^2}\right) = O\left(\frac{1}{w_a}\right). \end{aligned} \tag{6.19}$$

因此，根据 (5.3)，误分类的 B 类顶点的期望数量等于

$$\tilde{k}K_2(h\log(n))^{-\tau} = o(\tilde{k}) \tag{6.20}$$

对于某个 $K_2 > 0$。

取用于检验的权重阈值 $t_n$。由于 $\tilde{k}$ 服从均值为 $nt_n^{1-\tau}$的二项分布，定理 A.1.4 在 [2] 中给出，对所有 $\varepsilon > 0$

$$\mathbb{P}(\mathcal{E}_1) := \mathbb{P}\left(\tilde{k} < (1-\varepsilon)kt_n^{1-\tau}\right) \leq \exp(-kt_n^{1-\tau}((1-\varepsilon)\log(1-\varepsilon)+\varepsilon)) = \exp(-\tilde{\varepsilon}kt_n^{1-\tau}), \tag{6.21}$$

对于某个 $\tilde{\varepsilon} > 0$，并且

$$\mathbb{P}(\mathcal{E}_2) := \mathbb{P}\left(\tilde{n} < (1+\varepsilon)nt_n^{1-\tau}\right) \leq \exp(-nt_n^{1-\tau}((1+\varepsilon)\log(1-\varepsilon)-\varepsilon)) = \exp(-\hat{\varepsilon}nt_n^{1-\tau}), \tag{6.22}$$

对于某个 $\hat{\varepsilon} > 0$。因此，当 $k < n$ 时，

$$\mathbb{P}\left(\bar{\mathcal{E}}_1 \cap \bar{\mathcal{E}}_2\right) \geq 1 - \exp(-\zeta kt_n^{1-\tau}) \tag{6.23}$$

对于某个 $\zeta > 0$。由 (6.18) 和 (5.2)，取 $w_0 = t_n$，在 $\mathcal{E}_2$ 上，误分类的 A 类顶点的期望数量等于

$$O\left(\sum_{a\in[n-k]:w_a>t_n} \frac{\log(n)}{w_a}\right) = O(\log(n)t_n^{-\tau}nt_n^{1-\tau}), \tag{6.24}$$

当 $nt_n^{1-\tau}$如 $\bar{\mathcal{E}}_2$所示时，权重至少为$t_n$的 A类顶点至多有 $(1+\varepsilon)nt_n^{1-\tau}$ 个。此外，根据 (6.19)，被错误分类的类型 $B$ 顶点为 $o(k)$。

Now,

$$\begin{aligned}\mathbb{E}\left[\frac{|\hat{V}_C^{t_n}\triangle V_C^{t_n}|}{2|V_C^{t_n}|}\mid V_C^{t_n}\geq 1\right] &= \mathbb{P}\left(\mathcal{E}_1\cup\mathcal{E}_2\right)\mathbb{E}\left[\frac{|\hat{V}_C^{t_n}\triangle V_C^{t_n}|}{2|V_C^{t_n}|}\mid \mathcal{E}_1\cup\mathcal{E}_2, V_C^{t_n}\geq 1\right]\\ &\quad+\mathbb{P}\left(\bar{\mathcal{E}}_1\cap\bar{\mathcal{E}}_2\right)\mathbb{E}\left[\frac{|\hat{V}_C^{t_n}\triangle V_C^{t_n}|}{2|V_C^{t_n}|}\mid \bar{\mathcal{E}}_1\cap\bar{\mathcal{E}}_2\right]\\ &\leq n\exp(-\tilde{\varepsilon}kt_n^{1-\tau})+\frac{O(\log(n)t_n^{1-2\tau}n)}{(1-\varepsilon)kt_n^{1-\tau}}\\ &= o(1),\end{aligned} \tag{6.25}$$

只要

$$(n\log(n)/k)^{1/\tau}\ll t_n\ll k^{1/(\tau-1)}. \tag{6.26}$$

□

# A $H_1$ 中的期望度

现说明为何在备择假设 $H_1$ 下，必须对连接概率作出一项小的修正。我们首先计算权重为$w_i$的顶点的连接概率，然后利用它求出权重为 $w_i$ 的顶点的期望度。

在本附录中，我们称 $\mathbb{E}_X$ 为随机变量 $X$ 上的期望值。

**引理 A.1**。 在 $H_1$ 下，任意两个 $B$ 型顶点 $i$ 和 $j$ 以如下概率连接

$$P_1(i\leftrightarrow j)=\mathbb{E}_{x_i,x_j}[p_{ij}(w_i,w_j,x_i,x_j)]=\begin{cases}\frac{1}{1+C_1} & \text{if } \frac{w_iw_j}{k}\geq\frac{\mu}{2^d}\\ \frac{C_1}{1+C_1}\frac{w_iw_j}{k}+o(\frac{w_iw_j}{k}) & \text{otherwise.}\end{cases} \tag{A.1}$$

证明。 首先，注意到

$$\begin{aligned}\mathbb{E}_{x_j}[p_{ij}(w_i,w_j,x_i,x_j)] &= \frac{1}{1+C_1}\int_{\mathbb{T}^d}\left(\frac{w_iw_j}{n\mu||x_i-x_j||^d}\wedge 1\right)^\gamma \mathrm{d}x_j\\ &= \frac{1}{1+C_1}\int_0^{1/2}\left(\frac{w_iw_j}{n\mu r^d}\wedge 1\right)^\gamma \mathbb{P}\left(||x_j||=r\right)\mathrm{d}r\\ &= \frac{1}{1+C_1}\int_0^{1/2}\left(\frac{w_iw_j}{n\mu 2^{-d}V(r)}\wedge 1\right)^\gamma V'(r)\mathrm{d}r\end{aligned} \tag{A.2}$$

因为 $\mathbf{x}_i$ 是固定的，而 $\mathbf{x}_j$ 服从均匀分布，因此 $V(r)=\mathbb{P}(||x_j||\leq r)=(2r)^d$。随后，对 $x_i$ 取平均，并进行积分换元

$$\begin{aligned}\mathbb{E}_{\mathbf{x}_i}\left[\mathbb{E}_{\mathbf{x}_j}[p_{ij}(w_i,w_j,x_i,x_j)]\right] &= \frac{1}{1+C_1}\int_{[0,1]^d}\left(\int_0^1\left(\frac{w_iw_j}{n\mu 2^{-d}V}\wedge 1\right)^\gamma dV\right)\mathrm{d}x_i\\ &= \frac{1}{1+C_1}\int_0^1\left(\frac{w_iw_j}{n\mu 2^{-d}V}\wedge 1\right)^\gamma \mathrm{d}V.\end{aligned} \tag{A.3}$$

我们记作 $r_0=\frac{w_iw_j}{n\mu 2^{-d}\circ}$

- 当 $r_0 \geq 1$ 时，后者被积函数始终为 1，且 $\mathbb{E}_{x_i,\ x_j}[p_{ij}(w_i,\ w_j,\ x_i,\ x_j)] = 1$。
- 当 $r_0 < 1$ 时，

$$\begin{aligned}\mathbb{E}_{x_i,x_j}[p_{ij}(w_i,w_j,x_i,x_j)] &= \frac{1}{1+C_1}\left(\int_0^{r_0} 1dV + \int_{r_0}^{1}\left(\frac{r_0}{V}\right)^{\gamma} dV\right)\\ &= \frac{1}{1+C_1}\left(r_0 + \left[\frac{r_0^{\gamma}}{\gamma-1}\left(r_0^{-\gamma+1}-1\right)\right]\right)\\ &= \frac{1}{1+C_1} r_0(1+(\gamma-1)^{-1}) + O(r_0^{\gamma}). \end{aligned} \tag{A.4}$$

□

**引理 A.2.** 在 $H_1$ 下，任意顶点 $i$ 的期望度为 $\mathbb{E}[d_i] = w_i(1+o(1))$。

证明。 任意顶点 $i$ 的期望度为:

$$E_1[d_i] = E_1\left[\sum_{j\in V} \mathbb{1}_{\{i\sim j\}}\right]$$

- 若顶点 $i$ 是 A 类顶点，则它根据 (2.1) 与任意其他顶点相连。在这种情况下， $i$ 的期望度有如下上界和下界:

$$E_1\left[\sum_{j\in V} \mathbb{1}_{\{i\sim j\}}\right] = \sum_{j\in V} p_{ij}(w_i,w_j) \leq \frac{w_i}{\mu n}\sum_{j\in V} w_j = w_i(1+o(1))$$

and

$$E_1\left[\sum_{j\in V} \mathbb{1}_{\{i\sim j\}}\right] = \sum_{j\in V} p_{ij}(w_i,w_j) \geq \frac{w_i}{k}\sum_{j\in V\setminus[n-k]} w_j \mathbb{1}_{\{w_j\leq \mu n/w_i\}} = w_i(1+o(1))$$

于是， $E_1[d_i] = w_i(1+o(1))$。

- 若顶点 $i$ 是 B 型顶点，则它根据 (2.3) 与一个 A 类顶点 $j$ 相连，并根据 (2.4) 与一个 B 型顶点$k$ 相连。在这种情况下，

$$E_1\left[\sum_{j\in V} \mathbb{1}_{\{i\sim j\}}\right] = \sum_{j\in[n-k]} p_{ij}(w_i,w_j) + \sum_{j\in V\setminus[n-k]} \mathbb{E}_{x_i,x_j}[p_{ij}(w_i,w_j,x_i,x_k)] =: A+B.$$

根据与前一种情况相同的计算，我们有 $A = \frac{n-k}{n}\frac{w_i}{1+C_1}(1+o(1)) = \frac{w_i}{1+C_1}(1+o(1))$。另一方面，当 $i$， $j$ 为 B类时，

$$P(i\leftrightarrow j) = \begin{cases} \frac{1}{1+C_1} & \text{if } \frac{w_iw_j}{k} \geq \frac{\mu}{2^d} \\ \frac{C_1}{1+C_1}\frac{w_iw_j}{k} + o(\frac{w_iw_j}{k}) & \text{otherwise} \end{cases}$$

根据引理 A.1。因此，我们得到上界和下界

$$B \le \frac{C_1}{1+C_1}\frac{w_i}{k}\sum_{j\in V\setminus[n-k]} w_j(1+o(1)) = \frac{C_1}{1+C_1}w_i(1+o(1)),$$
$$B \ge \frac{C_1}{1+C_1}\frac{w_i}{k}\sum_{j\in V\setminus[n-k]} w_j(1+o(1))\mathbb{1}_{\{w_j\le \mu k/w_i 2^d\}} = \frac{C_1}{1+C_1}w_i(1+o(1)).$$

那么， $E_1[d_i] = A + B = w_i(1+o(1))$。

□

# B 辅助结果

**引理 B.1.** 给定权重序列 $(w_i)_{i\in V}$，假设 $G$ 是根据备择假设 $H_1$抽样得到的图。设 $p \ge 2$ 和 $\mathcal{P} = (a_1, ..., a_p)$ 是图 $G$中的顶点有序集。记$m = |\{i : a_i, a_{i+1}$ 是 $A$ 类顶点$\}|$ 的数量，其中 $\mathcal{P}$ 中的顶点对按照 *GIRG* 连接概率相连。那么，

$$P_1(a_1 \leftrightarrow a_2 \leftrightarrow ... \leftrightarrow a_p) = \prod_{i=1}^{p-1} P_1(a_i \leftrightarrow a_{i+1}) \tag{B.1}$$
$$= O\left(\frac{w_{a_1}^{-1}w_{a_p}^{-1}\prod_{i=1}^{p} w_{a_i}^2}{k^m n^{p-m-1}}\right) \tag{B.2}$$

证明。 我们首先使用关于路径 $\mathcal{P}$长度的归纳论证证明 (B.1)。当 $p = 2$时，该等式显然成立。接下来，假设该等式对 $p = s$成立，并证明它对 $p = s + 1$也成立。考虑顶点对 $(a_s, a_{s+1})$。

如果 $a_s$ 和 $a_{s+1}$ 中至少有一个是 A 类顶点，那么随机变量 $\mathbb{1}_{a_s\leftrightarrow a_{s+1}}$ 是参数为 $\min(1, w_{a_s}w_{a_{s+1}}/\mu n)$ 的伯努利随机变量，并且与路径的其余部分独立。因此， $P_1(a_1 \leftrightarrow a_2 \leftrightarrow ... \leftrightarrow a_{s+1}) = \prod_{i=1}^{s-1} P_1(a_i \leftrightarrow a_{i+1}) \cdot P_1(a_s \leftrightarrow a_{s+1})$。

如果 $a_s$ 和 $a_{s+1}$ 都是 B类，那么边 $(a_s, a_{s+1})$ 存在的概率取决于位置 $\mathrm{x}^{a_s}$和 $\mathrm{x}^{a_{s+1}}$。 然而，整条路径 $(a_1, ..., a_s)$ 存在于图中的概率可能取决于位置 $\mathrm{x}^{a_1}\cdots\mathrm{x}^{a_s}$。 于是，我们可以对位置 $\mathrm{x}^{a_s}$ 进行条件化，并得到

$$P_1(a_1 \leftrightarrow a_2 \leftrightarrow ... \leftrightarrow a_{s+1}) = \int_{\mathbb{T}^d} P_1(a_1 \leftrightarrow a_2 \leftrightarrow ... \leftrightarrow a_s \mid \mathrm{x}_{a_s}) P_1(a_s \leftrightarrow a_{s+1} \mid \mathrm{x}_{a_s}) \mathrm{d}P_1(\mathrm{x}_{a_s} = x). \tag{B.3}$$

现在，由于环面上无穷范数的对称性，可得 $P_1(a_s \leftrightarrow a_{s+1} \mid \mathrm{x}^{a_s}) = P_1(a_s \leftrightarrow a_{s+1})$。因此，

$$P_1(a_1 \leftrightarrow a_2 \leftrightarrow ... \leftrightarrow a_{s+1}) = P_1(a_s \leftrightarrow a_{s+1}) \int_{\mathbb{T}^d} P_1(a_1 \leftrightarrow a_2 \leftrightarrow ... \leftrightarrow a_s \mid \mathrm{x}_{a_s}) \mathrm{d}\mathbb{P}(\mathrm{x}_{a_s} = x) \tag{B.4}$$
$$= P_1(a_s \leftrightarrow a_{s+1}) P_1(a_1 \leftrightarrow a_2 \leftrightarrow ... \leftrightarrow a_s), \tag{B.5}$$

应用归纳假设即可完成证明。

接下来，我们证明 (B.2)。在路径 $m$ 中，顶点对按照 GIRG 的边概率相连。假设 $(a_r,\ a_{r+1})$ 是这样一对顶点，其中 $1 \le r < p$。那么，应用引理 A.1 可得 $P_1(a_r \leftrightarrow a_{r+1}) = O(w_{a_r} w_{a_{r+1}}/k)$。其余 $p-m-1$ 顶点对则按照非齐次随机图的边规则相连。在这种情况下，若 $(a_r,\ a_{r+1})$ 是这样一对顶点，其中 $r$，则有上界 $P_1(a_r \leftrightarrow a_{r+1}) = \min(1,\ w_{a_r} w_{a_{r+1}}/\mu n) = O(w_{a_r} w_{a_{r+1}}/n)$。综合这两个事实，我们得出

$$\prod_{i=1}^{p-1} P_1(a_i \leftrightarrow a_{i+1}) = O\left(\frac{w_{a_1}^{-1} w_{a_p}^{-1} \prod_{i=1}^{p} w_{a_i}^2}{k^m n^{p-m-1}}\right) \tag{B.6}$$

无论哪些顶点对形成于 B 型顶点之间。 □

# 参考文献


[1] N. Alon、M. Krivelevich 和 B. Sudakov。在随机图中寻找大型隐藏团。随机结构与算法，13(3-4):457–466，1998年。[2] N. Alon 和 J. H. Spencer。概率方法。Wiley 离散数学与优化丛书。John Wiley & Sons, Inc.，美国新泽西州霍博肯，第四版，2016年。[3] E. Arias-Castro 和 N. Verzelen。稠密随机网络中的社区检测。《统计学年鉴》，42(3):940–969，2014年。[4] L. Becchetti、P. Boldi、C. Castillo 和 A. Gionis。大规模局部三角形计数的高效算法。*ACM 数据知识发现汇刊*，4(3):1–28，2010年10月。[5] G. Bet、K. Bogerd、R. M. Castro 和 R. van der Hofstad。检测网络中的僵尸网络。数学统计与学习，3(3):315–343，2021年。[6] K. Bogerd、R. M. Castro、R. van der Hofstad 和 N. Verzelen。在非齐次随机图中检测植入社区。*Bernoulli*，27(2):1159–1188，2021年。[7] M. Brennan、G. Bresler 和 D. Nagaraj。随机图中潜在几何检测的相变。概率论及相关领域，178(3):1215–1289，2020年。[8] G. Bresler 和 D. Nagaraj。结构化与非结构化网络数据的最优单样本检验。载于学习理论会议，第1657–1690页。PMLR，2018年。[9] K. Bringmann、R. Keusch 和 J. Lengler。几何非齐次随机图。《理论计算机科学》，760:35–54，2019年。[10] S. Bubeck、J. Ding、R. Eldan 和 M. Z. Rácz。随机图中的高维几何检验。随机结构与算法，49(3):503–532，2016年。[11] F. Chung 和 L. Lu。给定期望度的随机图中的平均距离。美国国家科学院院刊，99(25):15879–15882，2002年。[12] R. Eldan 和 D. Mikulincer。各向异性随机几何图的信息与维数。载于泛函分析的几何方面，第273–324页。Springer，2020年。

[13] M. Faloutsos、P. Faloutsos 和 C. Faloutsos。互联网拓扑中的幂律关系。*ACM SIGCOMM*《计算机通信评论》，29(4):251–262，1999年。[14] C. Gao 和 J. Lafferty。利用小子图概率之间的关系检验网络结构。*arXiv*预印本 *arXiv:1704.06742*，2017年。[15] D. Ghoshdastidar、M. Gutzeit、A. Carpentier 和 U. Von Luxburg。非齐次随机图的双样本假设检验。《统计学年鉴》，48(4):2208–2229，2020年。[16] M. Girvan 和 M. E. Newman。社会网络和生物网络中的社区结构。《美国国家科学院院刊》，99(12):7821–7826，2002年。[17] B. Hajek、Y. Wu 和 J. Xu。随机图社区检测的计算下界。载于学习理论会议，第899–928页。PMLR，2015年。[18] P. W. Holland、K. B. Laskey 和 S. Leinhardt。随机区块模型：初步探索。《社会网络》，5(2):109–137，1983年6月。[19] J. Jin。基于得分的快速社区检测。《统计学年鉴》，43(1):57–89，2015年。[20] J. Jin、Z. Ke 和 S. Luo。通过图元计数进行网络全局检验。载于国际机器学习会议，第2333–2341页。PMLR，2018年。[21] D. Krioukov、F. Papadopoulos、M. Kitsak、A. Vahdat 和 M. Boguná。复杂网络的双曲几何。《物理评论*E*》，82(3):036106，2010年。[22] M. Latapy。超大（稀疏（幂律））图中三角形计算的实用算法。《理论计算机科学杂志》，407，2007年。[23] M. Latapy。超大（稀疏（幂律））图的主存三角形计算。《理论计算机科学》，407(1-3):458–473，2008年11月。[24] R. Michielan、N. Litvak 和 C. Stegehuis。网络中的双曲几何检测：为什么三角形还不够。《物理评论*E*》，106:054303，2022年11月。[25] M. E. Newman。网络中的模块度和社区结构。《美国国家科学院院刊》，103(23):8577–8582，2006年。[26] M. E. Newman 和 M. Girvan。发现并评估网络中的社区结构。《物理评论*E*》，69(2):026113，2004年。[27] M. E. J. Newman。高度聚类网络的性质。《物理评论*E*》，68(2)，2003年。[28] S. I. Resnick。重尾现象：概率与统计建模。Springer Science & Business Media，2007年。[29] C. Stegehuis、R. van der Hofstad、J. S. H. van Leeuwaarden 和 A. J. E. M. Janssen。无标度网络的聚类谱。《物理评论*E*》，96(4):042309，2017年。[30] N. Verzelen 和 E. Arias-Castro。稀疏随机网络中的社区检测。《应用概率年鉴》，25(6):3465–3510，2015年。

[31] Y. Wu 和 J. Xu。《植入结构的统计问题：信息论与计算极限》。数据科学中的信息论方法，第 383 页，2021 年。

[32] S.-Y. Yun 和 A. Proutiere。通过谱算法实现随机块模型中的精准社区检测。*arXiv* 预印本 *arXiv:1412.7335*，2014 年。