使用 k-核估计量匹配相关非齐次随机图

Miklós Z. Rácz 统计与数据科学系 计算机科学系 西北大学 埃文斯顿，伊利诺伊州
miklos.racz@northwestern.edu

Anirudh Sridhar 电气与计算机工程系 普林斯顿大学 普林斯顿，新泽西州
anirudhs@princeton.edu

arXiv:2302.05407v1 [math.ST] 10 Feb 2023

摘要——我们研究如何估计两个具有一般非齐次结构的边相关随机图之间的潜在顶点对应关系。我们研究所谓的 k-核估计量，它输出一种顶点对应关系，使两个图都包含一个较大的公共子图，且该子图的最小度至少为 k。我们推导出充分条件，在这些条件下， k-核估计量能够精确或部分恢复潜在顶点对应关系。最后，我们将这一通用框架具体应用于相关随机块模型、相关 Chung-Lu 图和相关随机几何图，推导出关于精确恢复和部分恢复的新结果。

I. 引言

过去十年来，人们对理解图匹配的基本极限产生了浓厚且日益增长的兴趣，既包括信息论层面 [6]–[9], [14], [16],[26]–[29], [31], [33] ，也包括算法层面 [2], [10], [11], [13],[22]–[25], [30], [34], ，并由此取得了若干近期突破。迄今为止，大多数研究聚焦于最简单的相关 Erd˝os–Rényi 随机图情形，只有少数工作研究了更广泛的模型，例如相关随机块模型 [15], [26], [28], 、相关随机增长图 [19], [29], 、相关 Chung–Lu 图 [3], [4], [35], 和相关几何模型 [32]。

A. 贡献

在本研究中，我们首次系统地发展了用于匹配具有一般非齐次结构的相关网络的技术。我们研究的匹配过程称为 k-核估计器，它对两个图 G_1 和 G_2 生成一个（可能是部分的）匹配，并由此诱导出一个最小度至少为 k 的大型公共子图。在 [8], [15], 的分析基础上，我们证明，对于根据图的密度适当选择的 k 值，k-核估计器能够保证生成两个图的完全正确的（可能是部分的）匹配。

接着，我们推导出一般条件，在这些条件下， k-核估计量能够完全或部分恢复潜在匹配。对于这两种目标，我们证明， k-核估计量能否成功取决于 G_1 和 G_2的交图的最小期望度。具体而言，当最小期望度大于 $\log n$时，即可保证精确恢复匹配，其中 n 是图中的顶点数。这一条件被认为是 Erd˝os–Rényi 图 [6], [7] 和具有两个平衡社区的随机块模型 [28]中精确图匹配的信息论阈值；我们猜想，对于范围广得多的模型，这一条件也可能在信息论意义上达到最优。对于部分恢复这一目标，我们证明，当交图的最小期望度大于图的非齐次性的某个函数时， k-核估计量能够成功。

最后，我们将一般性结果应用于若干受到广泛研究且值得关注的生成式图模型。具体而言，我们的工作首次对一般随机块模型、Chung-Lu 模型和带噪随机几何图中的无种子图匹配进行了理论分析。这些结果凸显了 k-核估计器的强大能力：作为一种与模型无关的实用过程，它有助于理解更贴近现实的网络模型中的图匹配。

B. 相关工作

近年来，图匹配信息论刻画的研究取得了显著进展。对于相关 Erd˝os–Rényi 图，文献 [6],[7], [33] 推导了精确图匹配可行与不可行的信息论条件。另有多项研究探讨了部分图匹配（即目标是正确恢复真实匹配中的任意正比例部分）的信息论刻画：文献 [8], [9], [13], [16], [33], 给出了不可行性结果，而 Ding 和 Du 最近确定了匹配的可达性结果 [9]。此外，我们指出，本文用于图匹配的 k-核估计器与文献 [9], 中使用的稠密子图估计器密切相关；后者在信息论意义上对部分恢复是最优的。有关这一点的更多详情，请参见注释 III.1。我们还指出，尽管本文以及上述文献主要关注图匹配的信息论条件，但一个引人入胜的相关研究方向是设计能够在可达区域内运行的高效算法（例如，参见 [2], [10], [11], [13], [22]–[24]）。

据我们所知，关于 Erd˝os–Rényi 图之外的网络，其图匹配的信息论极限鲜有研究。Onaran、Garg 和 Erkip 首先研究了相关随机块模型 [26]，，随后 Rácz 和 Sridhar [28] 确定了这一设定下精确恢复的信息论极限。近期，Wang、Wu、Xu 和 Yolou [32] 研究了两张带有对应于潜在几何结构的边权重的完全图之间的精确图匹配和近乎精确图匹配。这些研究的共同特点是，用于恢复匹配的统计量（例如最大后验（MAP）估计器）以及分析方法，都高度依赖于所考虑的网络模型类型。与之相反，我们的工作提供了一套与模型无关的工具，用于推导精确和部分图匹配的可达性结果。

我们简要讨论一类相关但有所不同的研究：利用辅助信息进行图匹配，其中辅助信息是许多事先已知且正确匹配的“种子节点”。在这种设定下，已有许多高效算法能够提供可证明的保证，从而在相关 Erd˝os-Rényi 图 [18], [25], [27], [30], [34], 、相关 Chung-Lu 图 [3], [4], [35] 以及相关优先连接图 [19]中精确恢复潜在匹配。另一方面，我们的工作研究的是在没有此类辅助信息的情况下能够获得哪些信息。

C. 结构安排

本文其余部分安排如下：首先在第 II 节介绍所研究的模型，随后在第 III 节详细说明我们的结果。证明见第 IV 节，部分证明则放在附录中。最后，我们将在第 V 节进行简要讨论。

D. 记号

我们将图表示为 $G = (V, E)$，其中 V 是顶点集，E 是边集。对于顶点 $v \in V$，令 $\deg_G(v)$ 表示其在 G 中的度。令 $\mathrm{core}_k(G)$ 表示 G 的 k-核，即最小度为 k 的极大子图。对于集合 $S \subseteq V$，$G\{S\}$ 表示 G 中对应于 S 的诱导子图。对于随机变量 X 和 Y，我们用 $X \preceq Y$ 表示 X 被 Y 随机支配。全文中，我们使用标准的渐近记号（例如，$O(\cdot)$ $o(\cdot)$）。

II. 相关非齐次随机图

我们首先描述一种非齐次随机图的一般模型。

定义 II.1（非齐次随机图）。设 n 为正整数，并设 $\mathbf{p} := \{p_{ij}\}_{i,\ j\in[n]}$ 是 $[0,1]$ 中表示边概率的一组元素。我们通过以概率 p_{ij} 将边 (i, j) 添加到 G 中来构造 $G \sim \mathcal{G}(\mathbf{p})$，且对 $[n]$ 中所有元素对而言，这些操作相互独立。

这是一个研究得非常充分的一般模型（例如，参见书籍 [12]）。许多著名的网络概率生成模型都可以视为上述非齐次随机图模型的特殊情形。例如，如果对于所有不同的 i、$j \in [n]$，均有 $p_{ij} = p$，我们便得到 Erd˝os-Rényi 图——这或许是最基本的生成网络模型。

在更复杂的模型中，p_{ij} 可能由潜在结构诱导，而这种结构可能会显著影响最终随机图的拓扑结构。也就是说，每个顶点 $i \in [n]$ 都关联一个潜变量（例如社区归属、权重、潜在位置），该变量通常从某个分布中随机抽取；在给定这些潜变量的条件下，该图是定义 II.1 中的非齐次随机图，其中 p_{ij} 是潜变量的函数。

接下来，我们介绍这类模型中三种常见的模型。我们的通用结果（见定理 III.5 和 III.9）适用于定义 II.1 中的非齐次随机图；因此，也适用于给定潜变量条件下的特定模型。正因如此，我们介绍这些模型时，均假设已给定潜变量，而不指定潜变量的分布。不过，正如我们将看到的，这些定理的条件很简单，因此不难理解它们何时以高概率（相对于潜变量的分布）成立。关于这些细节，我们留给读者自行探究。

例 II.2。在随机块模型（SBM）中，*[1], [17],*由潜在社区结构诱导。具体而言，假设网络中有 m 个社区，它们由 $[n]$ 的一个划分 $V_1, \ldots, V_m$ 表示。我们还指定一组 $\mathbf{q} = \{q_{ab}\}_{a,\ b\in[m]}$ 个取值于 $[0,1]$ 的元素，表示社区内部及社区之间的边形成概率。若对于任意不同的 i、$j \in [n]$，只要 $i \in V_a$ 且 $j \in V_b$，就有 $p_{ij} = q_{ab}$，则称其为 $G \sim \mathrm{SBM}(V_1, \ldots, V_m, \mathbf{q})$。

例 II.3. 在 Chung-Lu模型中，*[5],* 的 p_{ij} 由潜在顶点权重诱导。令 $\mathbf{w} \in \mathbb{R}_+^n$ 为一组满足 $\max_{i\in[n]} w_i \le \sqrt{\sum_{i\in[n]} w_i}$ 的顶点权重。若对于不同的 i、$j \in [n]$，有 $p_{ij} = w_i w_j / (\sum_{k\in[n]} w_k)$，则称 $G \sim \mathrm{CL}(\mathbf{w})$。

例 II.4. 考虑一个噪声随机几何图（参见 *[20]* 及其中的参考文献），其构造如下。令 d 为一个正整数（可以是常数，也可以随 n 增大），并令 $\mathbf{x} := \{x_1, \ldots, x_n\}$ 为 d 维球面 $\mathbb{S}^{d-1}$ 上的元素。给定参数 $r \in [0,2]$ 和 $p \in [0,1]$，若 $G \sim \mathrm{RGG}(\mathbf{x}, r, p)$，则

$$p_{ij} = p\mathbf{1}(\|x_i - x_j\|_2 \le r) \tag{II.1}$$

对于不同的 i、$j \in [n]$。

接下来，我们定义相关非齐次随机图。

定义 II.5（相关非齐次随机图）。令 n 为正整数，令 $\mathbf{p}$ 为边概率的集合，并令 $s \in [0,1]$ 为相关参数。根据以下过程生成一对相关非齐次随机图 (G_1, G_2)。首先，生成一个父图 $G \sim \mathcal{G}(\mathbf{p})$。给定 G 后，通过以概率 s 对 G 中的每条边进行子采样，独立构造一对与 G 具有相同顶点集的图 (G_1, G_2')。最后，将均匀随机置换 π_* 应用于

G_2' 的顶点标签，以生成 G_2。为简便起见，如果图对 (G_1, G_2) 是以这种方式构造的，我们称 $(G_1, G_2) \sim \mathcal{CG}(\mathrm{p}, s)$。

相应地，我们还可以构造相关随机块模型、Chung-Lu 图和带噪声随机几何图。我们分别将这些分布称为 CSBM、CCL 和 CRGG。

III. 结果：通过 k-核估计器进行图匹配

本节将介绍我们的主要结果，给出相关非齐次随机图中 k-核估计器正确性的充分条件。

A. k-核估计器的正确性

我们先给出一些定义。

定义 III.1（匹配）。一对 (M, μ) 是 G_1 与 G_2 之间的匹配，当且仅当 $M \subseteq [n]$，$\mu : M \to [n]$，且 μ 是单射。

我们将 $G_1 \wedge_\mu G_2$ 记为关于匹配 (M, μ) 的交图。具体而言，$G_1 \wedge_\mu G_2$ 的顶点集为 M，当且仅当 (i, j) 是 G_1 中的一条边，且 $(\mu(i)\ \mu(j))$ 是 G_2中的一条边时，(i, j) 才是 $G_1 \wedge_\mu G_2$ 中的一条边。

定义 III.2（k-核匹配）。若 $G_1 \wedge_\mu G_2$ 中的最小度至少为 k，则匹配 (M, μ) 称为 k-核匹配。

这一术语源自 k-核的概念：图的 k-核是最小度为 k 的极大子图。

定义 III.3（k-核估计量）。给定一对定义在 n 个顶点和 $k \in [n]$上的图 (G_1, G_2)，k-核估计量 $\widehat{(M_k, \widehat{\mu}_k)}$ 是一个 k-核匹配，且在所有 k-核匹配中涉及的顶点数最多。（若其定义不唯一，则任取一个这样的匹配。）

下面的引理给出了一个一般且简单的充分条件，在此条件下，k-核估计器能够以高概率成功。更确切地说，该引理保证 k-核估计器能够准确恢复交图的 k-核上的潜在匹配，不产生任何错误（并且不会在此k-核之外进行匹配）。

在陈述结果之前，对于匹配 (M, μ) 和 $S \subseteq M$，我们定义集合 $\mu\{S\} := \{(i, \mu(i)) : i \in S\}$。

引理 III.4. 设 $(G_1, G_2) \sim \mathcal{CG}(\mathrm{p}, s)$。假设 $p_{\max} \le n^{-\alpha+o(1)}$ 对某个 $\alpha \in (1/2, 1]$ 成立，且 $k > 12/(2\alpha - 1)$。那么，以 $1 - o(1)$ 的概率，有

$$\widehat{M_k} = \mathsf{core}_k(G_1 \wedge_{\pi_*} G_2) \text{ and } \widehat{\mu}_k\{\widehat{M_k}\} = \pi_*\{\widehat{M_k}\}.$$

该引理的证明见附录 A。

注释 III.1. k-核估计器与 [9]，的稠密子图估计器有些相似，后者输出一个估计量 $\widehat{\mu}$，使得 $G_1 \wedge_{\widehat{\mu}} G_2$ 包含一个平均度较高的子图。值得注意的是，稠密子图估计器能够成功实现部分图匹配，直至信息论阈值。

从某种意义上说，k-核估计器可以视为稠密子图估计器的稳健版本：通过寻找最小度有下界（而非平均度有下界）的子图，我们可以保证匹配中的所有顶点都得到正确匹配。另一方面，稠密子图估计器输出的是完整匹配而非部分匹配，因此在稀疏区域会产生若干错误。然而，k-核估计器的保证能否以某种方式推广到稠密子图估计器，仍是一个值得探讨的开放问题。

B. 精确图匹配

引理 III.4 给出了使 k-核估计量 $\widehat{\mu}_k$ 恢复 π_*（无论是精确恢复还是部分恢复）的充分条件。下一个结果是我们关于相关非齐次随机图中精确图匹配的主要一般性定理。

定理 III.5. 设 $(G_1, G_2) \sim \mathcal{CG}(\mathrm{p}, s)$。假设 $p_{\max} \le n^{-\alpha+o(1)}$ 对某个 $\alpha \in (1/2, 1]$ 成立，并且 $k > 12/(2\alpha - 1)$ 是一个常数。此外，假设存在 $\epsilon > 0$ 使得

$$\min_{i \in [n]} \sum_{j=1}^{n} p_{ij} s^2 \ge (1 + \epsilon) \log n. \tag{III.1}$$

Then $\mathbb{P}(\widehat{\mu}_k = \pi_*) = 1 - o(1)$ *as* $n \to \infty$.

也就是说，在定理的条件下，k-核估计量 $\widehat{\mu}_k$ 以高概率精确恢复 π_*。条件 (III.1) 简单且自然：换言之，它表示交图$G_1 \wedge_{\pi_*} G_2$中每个节点的期望度至少为 $(1 + \epsilon) \log n$。在 Erd˝os–Rényi 情形下，这一条件简化为 $nps^2 \ge (1 + \epsilon) \log n$，已知这是信息论意义上的最优条件 [6], [7]。更一般地，视 p 而定，可能可以改进条件 (III.1)；事实上，我们在下文的定理 IV.3 中证明了一个略强的结果（所需的充分条件更弱）。不过，我们猜想，对于很大范围的 p，定理 III.5 中这个简单的充分条件已接近最优。

接下来，我们将定理 III.5 专门应用于三个重要情形：相关随机块模型、相关 Chung-Lu 图以及相关随机几何图。证明较为简短，可见附录 B。值得注意的是，在这三种情形中，精确恢复的充分条件都是关于最小期望度的条件；对于大多数自然的潜变量分布，这一条件都很容易计算。

推论 III.6. 设 $(G_1, G_2) \sim \mathrm{CSBM}(V_1, \ldots, V_m, \mathrm{q}, s)$。假设 $\max_{a,b \in [m]} q_{ab} \le n^{-\alpha+o(1)}$，其中某个 $\alpha \in (1/2, 1]$，并且 $k > 12/(2\alpha - 1)$ 是常数。此外，假设存在 $\epsilon > 0$ 使得

$$\min_{a \in [m]} \sum_{b=1}^{m} |V_b| q_{ab} s^2 \ge (1 + \epsilon) \log n.$$

则当 $n \to \infty$ 时， $\mathbb{P}(\widehat{\mu}_k = \pi_*) = 1 - o(1)$。

推论 III.7. 令 $(G_1, G_2) \sim \mathsf{CCL}(\mathsf{w}\ s)$。假设

$$\max_{i\in[n]} w_i \le n^{-\alpha/2+o(1)} \sqrt{\sum_{i=1}^{n} w_i}, \tag{III.2}$$

对某个 $\alpha \in (1/2, 1]$，并令 $k > 12/(2\alpha - 1)$ 为常数。此外，假设存在 $\epsilon > 0$ 使得

$$\min_{i\in[n]} w_i s^2 \ge (1+\epsilon)\log n.$$

则当 $n \to \infty$ 时， $\mathbb{P}(\widehat{\mu}_k = \pi_*) = 1 - o(1)$。

推论 III.8. 令 $(G_1, G_2) \sim \mathsf{CRGG}(\mathsf{x}\ r, p)$。假设 $p \le n^{-\alpha+o(1)}$ 对某个 $\alpha \in (1/2, 1]$ 成立，并令 $k > 12/(2\alpha - 1)$ 为常数。此外，假设存在 $\epsilon > 0$ 使得

$$ps^2 \min_{i\in[n]} |\{j \in [n] : \|x_i - x_j\| \le r\}| \ge (1+\epsilon)\log n.$$

则当 $n \to \infty$ 时， $\mathbb{P}(\widehat{\mu}_k = \pi_*) = 1 - o(1)$。

C. 部分图匹配

超越精确匹配，引理 III.4 还使我们能够在无法实现精确恢复时，利用 k-核估计量 $\widehat{\mu}_k$ 对 π_* 进行部分恢复并获得定量结果。在陈述主要结果之前，我们先定义

$$R := \frac{(n-1)p_{\max}}{\min_{i\in[n]} \sum_{j=1}^{n} p_{ij}}.$$

量 R 在某种意义上衡量了非齐次图模型中的异质性。事实上，如果 p_{ij} 是常数，那么 $R = 1$。然而，如果 p_{ij} 的最小值和最大值之间存在显著差异，那么 R 的取值就会更大。

接下来的结果是关于相关非齐次随机图中部分图匹配的主要一般性定理。它突显了 R 在 π_* 部分恢复中的作用。为简便起见，我们在此聚焦于 $p_{\max} = n^{-1+o(1)}$ 的情形，不过这些技术也可推广到更广的 p 范围。

定理 III.9。设 $(G_1, G_2) \sim \mathcal{CG}(\mathsf{p}\ s)$。假设 $p_{\max} = n^{-1+o(1)}$ 且 $k = 13$。若

$$\min_{i\in[n]} \sum_{j=1}^{n} p_{ij} s^2 \ge \max\{R, 154\},$$

那么，以概率 $1 - o(1)$ 有

$$|\widehat{M}_k| \ge \left(1 - 3e^{-\min_{i\in[n]} \sum_{j=1}^{n} p_{ij} s^2/7}\right) n - o(n) \tag{III.3}$$

并且

$$\widehat{\mu}_k\{\widehat{M}_k\} = \pi_*\{\widehat{M}_k\}. \tag{III.4}$$

也就是说，在该定理的条件下， k-核估计器以高概率匹配了占节点总数 $(1 - 3\exp(-\min_{i\in[n]} \sum_{j=1}^{n} p_{ij} s^2/7) - o(1))$ 的节点（见 (III.3)），且所有这些匹配均正确（见 (III.4)）。我们强调，不产生任何错误匹配这一性质非常有用，尤其是在匹配结果进一步用于下游任务（例如社群恢复 [15]）时。

我们再次将一般定理具体应用于这三种情形。证明很简短，可见附录 B。这里，在这三种情形中， π_* 部分恢复的充分条件也取决于最小期望度以及 R （异质性的度量）；对于大多数常见的潜变量分布，这些量都很容易计算。

推论 III.10. 设 $(G_1, G_2) \sim \mathsf{CSBM}(V_1, \ldots, V_m, \mathsf{q}\ s)$，且令 $q_{\max} := \max_{a,\ b\in[m]} q_{ab}$ 以及 $q_{\min} := \min_{a,\ b\in[m]} q_{ab}$。假设 $q_{\max} = n^{-1+o(1)}$ 且 $k = 13$。若存在 $\epsilon > 0$ 使得

$$\min_{a\in[m]} \sum_{b=1}^{m} |V_b| q_{ab} s^2 \ge \max\left\{\frac{q_{\max}}{q_{\min}}, 154\right\} + \epsilon.$$

推论 III.11. 设 $(G_1, G_2) \sim \mathsf{CCL}(\mathsf{w}\ s)$，并记 $w_{\max} := \max_{i\in[n]} w_i$ 和 $w_{\min} := \min_{i\in[n]} w_i$。若 $w_{\max}$ 有界（$n \to \infty$），且存在 $\epsilon > 0$ 使得，则定理 III.9 的结论成立。

$$w_{\min} s^2 \ge \max\left\{\left(\frac{w_{\max}}{w_{\min}}\right)^2, 154\right\} + \epsilon.$$

推论 III.12. 设 $(G_1, G_2) \sim \mathsf{CRGG}(\mathsf{x}\ r, p)$。若 $p = n^{-1+o(1)}$，则定理 III.9 的结论成立；并且对某个 $\epsilon > 0$，有

$$ps^2 \min_{i\in[n]} |\{j \in [n] : \|x_i - x_j\| \le r\}| \ge \max\{s\sqrt{np}, 154\} + \epsilon.$$

IV. 证明：恢复潜在匹配

A. 度分布的性质

对于图 G 和正整数 k，定义集合

$$Z_k := \{i \in [n] : \deg_G(i) \le k\}. \tag{IV.1}$$

我们陈述一些关于非齐次随机图中 Z_k 大小的有用结果。其证明推迟至附录 C。

引理 IV.1. 设 $G \sim \mathcal{G}(\mathsf{p})$。那么，对于任意正整数 k 和任意 $c \in (0,1)$，有

$$\mathbb{E}[|Z_k|] \le \left(\frac{1}{1-c}\right)^k \sum_{i=1}^{n} \exp\left(-c\sum_{j=1}^{n} p_{ij}\right).$$

引理 IV.2. 设 $G \sim \mathcal{G}(\mathsf{p})$，并假设 $p_{\max} := \max_{i,\ j\in[n]} p_{ij} = o(1/\sqrt{n})$。那么对每个 $k \in [n]$，

$$\mathbb{P}\left(||Z_k| - \mathbb{E}[|Z_k|]| \le \frac{1}{3}n^{3/4}\right) = 1 - o(1), \quad as\ n \to \infty.$$

B. 精确恢复

我们在此陈述并证明一个略强于定理 III.5 的结果（定理 III.5 随即便可推出，见下文）。

定理 IV.3. 设 $(G_1, G_2) \sim \mathcal{CG}(\mathsf{p}\ s)$。假设 $p_{\max} \le n^{-\alpha+o(1)}$ 对某个 $\alpha \in (1/2, 1]$ 成立，且 $k > 12/(2\alpha - 1)$。那么对于任意 $c \in (0, 1)$，

$$\mathbb{P}(\widehat{\mu}_k \ne \pi_*) \le \left(\frac{1}{1-c}\right)^{k-1} \sum_{i=1}^{n} \exp\left(-c\sum_{j=1}^{n} p_{ij} s^2\right) + o(1),$$

其中 $o(1) \to 0$ 随 $n \to \infty$而趋于。

证明。我们可以将所关注的概率界定如下:

$$\begin{aligned}
&\mathbb{P}(\widehat{\mu}_k \neq \pi_*) \le \mathbb{P}(\mathsf{core}_k(G_1 \wedge_{\pi_*} G_2) \neq [n]) \\
&\quad + \mathbb{P}\Big(\widehat{M}_k \neq \mathsf{core}_k(G_1 \wedge_{\pi_*} G_2) \text{ or } \widehat{\mu}_k\{\widehat{M}_k\} \neq \pi_*\{\widehat{M}_k\}\Big) \\
&\overset{(a)}{\le} \mathbb{P}(\mathrm{d}_{\min}(G_1 \wedge_{\pi_*} G_2) < k) + o(1) \\
&\overset{(b)}{\le} \sum_{i=1}^{n} \mathbb{P}(\deg_{G_1 \wedge_{\pi_*} G_2}(i) < k) + o(1) \\
&= \mathbb{E}\Big[\big|\{i \in [n] : \deg_{G_1 \wedge_{\pi_*} G_2}(i) \le k-1\}\big|\Big] + o(1). \quad \text{(IV.2)}
\end{aligned}$$

上式中的 (a) 可由引理 III.4 得出，并结合如下观察：如果图的 k-核未包含整个顶点集，则该图的最小度必然小于 k。不等式 (b) 由并集界得出。最后，我们利用引理 IV.1 对 (IV.2) 中的期望进行界定。

□

定理 III.5 的证明。令定理 IV.3 中的 $c = (1+\epsilon)^{-1/2}$。利用假设——对每个 $i \in [n]$，我们都有下界 $\sum_{j=1}^{n} p_{ij} s^2 \ge (1+\epsilon)\log n$——可得

$$\mathbb{P}(\widehat{\mu}_k \neq \pi_*) \le \left(\frac{1}{1-(1+\epsilon)^{-1/2}}\right)^{k-1} n^{1-\sqrt{1+\epsilon}} + o(1).$$

当 k 固定时，此界在 $n \to \infty$ 时趋于零。 □

C. 部分恢复

我们首先陈述 Łuczak 关于 Erd˝os-Rényi 图中小型诱导子图密度的一个有用结果。

引理 IV.4 ([21])。假设 $\gamma = o(\sqrt{n})$，并令 $G \sim \mathrm{ER}(n, \gamma/n)$。那么，以概率 $1-o(1)$，对于每个满足 $|S| \le \frac{3}{4\gamma^2} n$ 的 $S \subset [n]$，$G\{S\}$ 至多有 $2|S|$ 条边。

通过一个简单的耦合论证，我们可以将此引理推广到非齐次随机图的情形。

引理 IV.5。假设 $p_{\max} \le \gamma/n$，其中 $\gamma = o(\sqrt{n})$。那么，引理 IV.4 的结论对 $G \sim \mathcal{G}(\mathbf{p})$ 仍成立。

证明。假设我们将 G 与 $G' \sim \mathcal{G}(n, \gamma/n)$ 耦合，使得 G 总是 G' 的一个子图。因此，对于任意满足 $|S| \le \frac{3}{4\gamma^2} n$ 的 $S \subset [n]$，如果 $G'\{S\}$ 至多有 $2|S|$ 条边，那么 $G\{S\}$ 也至多有 $2|S|$ 条边。所需结论由此得出。 □

令 $F_k := [n] \setminus \mathsf{core}_k(G)$ 表示不属于 G 的 k-core 的顶点集合。还记得 (IV.1) 中 Z_k 的定义；在下文中每当使用 Z_k 时，所指的图都可以从上下文中明确得知。

引理 IV.6. 设 $G \sim \mathcal{G}(\mathbf{p})$。假设对所有 i，都有 $p_{ij} \le \gamma/n$，$j \in [n]$，其中 $\gamma = o(\sqrt{n})$。假设引理 IV.5 中的事件成立，且 $|Z_{k+1}| \le n/(4\gamma^2)$。那么 $|F_k| \le 3|Z_{k+1}|$。

证明采用与 Łuczak [21] 类似的方法；证明过程详见附录D。

引理 IV.7. 假设 $p_{\max} = o(n^{-7/8})$；也就是说，假设 $\gamma = o(n^{1/8})$。还假设 $\mathbb{E}[|Z_{k+1}|] \le n/(5\gamma^2)$。那么，以概率 $1-o(1)$，有

$$|F_k| \le 3\mathbb{E}[|Z_{k+1}|] + n^{3/4}.$$

证明。引理 IV.2 蕴含，以概率 $1-o(1)$，有 $3|Z_{k+1}| \le 3\mathbb{E}[|Z_{k+1}|] + n^{3/4}$，并且

$$|Z_{k+1}| \le \mathbb{E}[|Z_{k+1}|] + \frac{n^{3/4}}{3} \le \frac{n}{5\gamma^2} + \frac{n^{3/4}}{3} \le \frac{n}{4\gamma^2},$$

其中在最后一个不等式中，我们用到了 $\gamma = o(n^{1/8})$。应用引理 IV.6 即可得到所需结果。 □

现在我们可以证明部分图匹配的主要结果了。

定理 III.9 的证明。简写定义 $d_{\min} := \min_{i \in [n]} \sum_{j=1}^{n} p_{ij}$。根据引理 III.4，只需证明

$$|\mathsf{core}_k(G_1 \wedge_{\pi_*} G_2)| \ge \Big(1 - 3e^{-d_{\min} s^2/7}\Big) n - o(n) \quad \text{(IV.3)}$$

以高概率成立。因此，根据引理 IV.7，只需对 $G_1 \wedge_{\pi_*} G_2$ 中度至多为 $k+1 = 14$ 的顶点的期望数量进行界定。通过在引理 IV.1 中令 $c = 1/2$，该数量至多为

$$\begin{aligned}
&2^{k+1} \sum_{i=1}^{n} \exp\left(-\frac{1}{2}\sum_{j=1}^{n} p_{ij} s^2\right) \le n 2^{k+1} \exp\left(-\frac{1}{2} d_{\min} s^2\right) \\
&\le n \exp\left\{-\left(\frac{1-\log 2}{2}\right) d_{\min} s^2\right\} \le n \exp\left(-\frac{d_{\min} s^2}{7}\right).
\end{aligned}$$

上式中，第二个不等式使用了我们的假设 $d_{\min} s^2 \ge 2(k+1)$，最后一个不等式则利用 $(1-\log 2)/2$ 的下界为 $1/7$。

接下来，根据上面的公式，当

$$\exp\left(-\frac{d_{\min} s^2}{7}\right) \le \frac{1}{5(n p_{\max} s^2)^2} = \frac{((n-1)/n)^2}{5R^2 (d_{\min} s^2)^2}, \quad \text{(IV.4)}$$

时，引理 IV.7 的结论成立，其中引理 IV.7 中的参数 γ 取为 $np_{\max}$；上式中的最后一个等式由 R 的定义得到。为简化条件 (IV.4)，我们作几点观察。首先，当 n 足够大且 $d_{\min} s^2 \ge R$ 时，若

$$\exp\left(-\frac{d_{\min} s^2}{7}\right) \le \frac{1}{6(d_{\min} s^2)^4}. \quad \text{(IV.5)}$$

接下来可以通过数值方式看出，当$x \ge 154$时，$e^{-x/7} \le 1/(6x^4)$。因此，当 n 足够大且 $d_{\min} s^2 \ge \max\{R, 154\}$时，条件 (IV.5) 得以满足；结合引理 IV.7，可推出 $|F_k| \le 3e^{-d_{\min} s^2/7} + o(n)$。(IV.3) 中的界可直接得到。 □

V. 结论与讨论

本文开启了对相关非齐次随机图图匹配的系统研究。我们的主要结果表明， k-核估计器具有理想的信息论保证，无论是在一般设置下，还是针对若干经过深入研究的特定模型。

我们的工作为未来研究提出了许多问题。一个具体的开放问题是：我们能否放宽关于 $p_{\max}$ 的假设，允许 $p_{\max} \le n^{-\alpha+o(1)}$ 具有 $\alpha \le 1/2$。一个非常有趣的未来方向是，针对特定模型开发能够适应底层模型异质性的反向界。最后，近期针对 Erd˝os–Rényi 情形提出的计算高效估计量方面的突破，能否推广到非齐次情形?

参考文献

[1] E. Abbe. 社区检测与随机块模型：近期进展. 机器学习研究期刊, 18 6531, 2017 1 6446( ): – .[2] B. Barak, C.-N. Chou, Z. Lei, T. Schramm, and Y. Sheng. 相关随机图上的图匹配问题（近乎）高效算法. 收录于神经信息处理系统进展（ NeurIPS ） , 第 9190–9198 页 , 2019. [3] K. Bringmann, T. Friedrich, and A. Krohmer. 拟线性时间内对异质随机图进行去匿名化. 收录于欧洲算法研讨会论文集, 第197–208页, 2014. [4] C.-F. Chiasserini, M. Garetto, and E. Leonardi. 无标度用户关系下的社交网络去匿名化. IEEE/ACM 网络汇刊, 24(6):3756–3769, 2016. [5] F. Chung and L. Lu. 具有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. [6] D. Cullina and N. Kiyavash. Erd˝os-Rényi 图匹配的改进可达界与反向界. 收录于 ACM SIGMETRICS, 第44卷, 第63–72页, 2016. [7] D. Cullina and N. Kiyavash. 相关 Erd˝os-Rényi 图的精确对齐恢复. 预印本见 https://arxiv.org/abs/1711.06783, 2018. [8] D. Cullina, N. Kiyavash, P. Mittal, and H. V. Poor. 通过 k-核对齐实现 Erd˝os-Rényi 图对齐的部分恢复. SIGMETRICS Perform. Eval. Rev., 48(1):99–100, 2020年7月. [9] J. Ding and H. Du. 相关随机图的匹配恢复阈值. 预印本见 https://arxiv.org/abs/2205.14650, 2022. [10] J. Ding, Z. Ma, Y. Wu, and J. Xu. 通过度分布实现高效随机图匹配. 概率论及相关领域, 179 115, 2021 1 29( ): – .[11] Z. Fan, C. Mao, Y. Wu, and J. Xu. 谱图匹配与正则化二次松弛：算法与理论. 收录于第37届国际机器学习会议（IC ML）论文集, 机器学习研究论文集第119卷, 第2985–2995页. PMLR, 2020年7月13–18日. [12] A. Frieze and M. Karo´nski. 随机图导论. 剑桥大学出版社, 2016. [13] L. Ganassali and L. Massoulié. 从树匹配到稀疏图对齐. 收录于第三十三届学习理论会议（COLT）论文集, 机器学习研究论文集第125卷, 第1633–1665页. PMLR, 2020年7月9–12日. [14] L. Ganassali, L. Massoulié, and M. Lelarge. 图对齐问题中部分恢复的不可能性. 收录于学习理论会议, 第2080–2102页. PMLR, 2021. [15] J. Gaudio, M. Z. Rácz, and A. Sridhar. 相关随机块模型中的精确社区恢复. 收录于第三十五届学习理论会议（ COLT ）论文集 , 机器学习研究论文集第178卷, 第2183–2241页. PMLR, 2022年7月2–5日. [16] G. Hall and L. Massoulié. 图对齐问题中的部分恢复. 预印本见 https://arxiv.org/abs/2007.00533, 2020. [17] P. W. Holland, K. B. Laskey, and S. Leinhardt. 随机块模型：初步研究. 社会网络, 5(2):109–137, 1983.

[18] E. Kazemi、S. H. Hassani 和 M. Grossglauser。从少量种子节点开始扩展图匹配。VLDB Endowment 论文集，8(10):1010–1021，2015年6月。[19] N. Korula 和 S. Lattanzi。面向社交网络的高效协调算法。VLDB Endowment 论文集，7 388，2014 5 377( ): – 。[20] S. Liu 和 M. Z. Rácz。含噪高维随机几何图中的相变。预印本见 https://arxiv.org/abs/2103.15249，2021。[21] T. Łuczak。随机图 k-核的大小与连通性。离散数学，91(1):61–68，1991。[22] C. Mao、M. Rudelson 和 K. Tikhomirov。常数相关性下随机图的精确匹配。预印本见 https://arxiv.org/abs/2110.05000，2021。[23] C. Mao、M. Rudelson 和 K. Tikhomirov。具有更强噪声鲁棒性的随机图匹配。发表于第34届学习理论会议（COLT）论文集，页3296–3329。PMLR，2021。[24] C. Mao、Y. Wu、J. Xu 和 S. H. Yu。通过计数枝形图，在 Otter阈值处实现随机图匹配。预印本见 https://arxiv.org/abs/2209.12313，2022。[25] E. Mossel 和 J. Xu。通过大邻域统计量实现带种子图匹配。发表于第三十届 ACM-SIAM离散算法研讨会（SODA）年会论文集，页1005–1014，2019。[26] E. Onaran、S. Garg 和 E. Erkip。具有社区结构的随机图中的最优去匿名化。发表于2016年第50届阿西洛马信号、系统与计算机会议，页709–713。IEEE，2016。[27] P. Pedarsani 和 M. Grossglauser。匿名网络的隐私问题。发表于第17届 ACM SIGKDD知识发现与数据挖掘国际会议（KDD）论文集，页1235–1243，2011。[28] M. Z. Rácz 和 A. Sridhar。相关随机块模型：结合社区恢复应用的精确图匹配。发表于神经信息处理系统进展（NeurIPS），2021。[29] M. Z. Rácz 和 A. Sridhar。相关随机增长图。应用概率年刊，32(2):1058–1111，2022。[30] F. Shirani、S. Garg 和 E. Erkip。带种子图匹配：高效算法与理论保证。发表于2017年第51届阿西洛马信号、系统与计算机会议，页253–257，2017。[31] F. Shirani、S. Garg 和 E. Erkip。基于测度集中的相关图匹配方法。IEEE信息论选定领域期刊，2(1):338–351，2021。[32] H. Wang、Y. Wu、J. Xu 和 I. Yolou。几何模型中的随机图匹配：完全图情形。发表于第三十五届学习理论会议（COLT）论文集，机器学习研究论文集第178卷，页3441–3488。PMLR，2022年7月2日至5日。[33] Y. Wu、J. Xu 和 S. H. Yu。确定随机图匹配的精确重构阈值。IEEE信息论汇刊，68(8):5391–5417，2022。[34] L. Yartseva 和 M. Grossglauser。渗流图匹配的性能。发表于首届 ACM在线社交网络会议论文集，COSN '13，页119–130，美国纽约州纽约市，2013年。计算机协会。[35] L. Yu、J. Xu 和 X. Lin。匹配幂律图中 D 跳的威力。ACM 计算系统测量与分析，5(2):1–43，2021。

附录 A 引理 III.4 的证明

对于匹配 (M,μ)，定义

$$f(M,\mu,G_1,G_2,\pi_*) := \sum_{i\in M:\mu(i)\neq\pi_*(i)} \deg_{G_1\wedge_\mu G_2}(i).$$

为简洁起见，我们用 $f(\mu)$ 代替 $f(M,\mu,G_1,G_2,\pi_*)$。换言之，$f(\mu)$ 表示在 $G_1\wedge_\mu G_2$ 中被 μ 错误匹配的顶点的度数之和。

定义 A.1（弱 k-核匹配）。若匹配 (M,μ) 满足 $f(\mu) \geq k|\{i\in M\colon \mu(i)\neq\pi_*(i)\}|$，则称其为弱 k-核匹配。

通俗地说，若 (M,μ) 是弱 k-核匹配，则 μ 的错误匹配区域中的平均度至少为 k。注意，若 (M,μ) 是 k-核匹配，则它也是弱 k-核匹配，但反过来则未必成立。

接下来，我们定义极大匹配这一有用概念。

定义 A.2（π_*-极大匹配）。若匹配 (M,μ) 对每个 $i\in[n]$ 都满足以下条件之一：$i\in M$，或 $\pi_*(i)\in\mu(M)$，其中 $\mu(M)$ 是 M 在 μ 下的像，则称其为 π_*-极大匹配。此外，我们将 $\mathcal{M}(d)$ 定义为会产生 d 个错误的 π_*-极大匹配集合（即 $|\{i\in M\colon \mu(i)\neq\pi_*(i)\}|=d$）。

以下引理的证明见 [8], [15], ，它为任意 k-核匹配的正确性提供了一个通用的充分条件。

引理 A.3. 设 (G_1,G_2) 是定义在顶点集 $[n]$ 上的一对随机图，其真实匹配为 π_*。对于任意正整数 k，定义如下量：

$$\xi := \max_{1\leq d\leq n}\max_{(M,\mu)\in\mathcal{M}(d)} \mathbb{P}(f(\mu)\geq kd)^{1/d}.$$

设 $(M_k,\widehat{\mu}_k)$ 是 G_1 和 G_2 的 k-核估计量。那么

$$\mathbb{P}\Big(\widehat{M}_k = \mathsf{core}_k(G_1\wedge_{\pi_*}G_2) \text{ and } \widehat{\mu}_k\{\widehat{M}_k\}=\pi_*\{\widehat{M}_k\}\Big) \geq 2-\exp(n^2\xi).$$

换言之，ξ 表示给定的一个极大匹配是弱 k-核匹配的概率，并按 (M,μ) 产生的错误数进行了归一化。关键的是，引理 A.3 表明，如果 $\xi=o(n^{-2})$，那么 k-核估计量将以高概率得到正确结果。Cullina、Kiyavash、Mittal 和 Poor [8] 首次针对相关 Erd˝os–Rényi 图证明了引理 A.3。随后，Gaudio、Rácz 和 Sridhar [15] 指出，[8] 中的证明可以直接推广到一般的相关随机图对 (G_1,G_2)。关于引理 A.3 的证明，请参阅 [15，引理 19 和推论 20]。

我们接下来通过界定 ξ来进行分析，具体形式见下文引理。需要指出的是，证明过程基本沿用了 [15,引理 23], ，但为完整起见，我们仍将其列出。

引理 A.4. 设 $p_{\max} := \max_{i,j\in[n]} p_{ij}$。对于任意匹配 $(M,\mu)\in\mathcal{M}(d)$ 以及任意 $\theta>0$，有

$$\mathbb{P}(f(\mu)\geq kd) \leq 3\exp\big\{-d\big(\theta k - e^{2\theta}p_{\max}s^2 - ne^{6\theta}p_{\max}^2s^2\big)\big\}.$$

证明。定义以下集合：

$$\mathcal{A}(\mu) := \{(i,j)\in M^2 : \mu(i)\neq\pi_*(i)\}$$
$$\mathcal{B}(\mu) := \{(i,j)\in\mathcal{A}(\mu) : \mu(i)=\pi_*(j) \text{ and } \mu(j)=\pi_*(i)\}$$
$$\mathcal{C}(\mu) := \mathcal{A}(\mu)\setminus\mathcal{B}(\mu).$$

关于这些集合，我们作几点说明。首先，由于 μ 会产生d个错误，$\mathcal{A}(\mu)\leq d|M|\leq dn$。此外，对于 $(i,j)\in\mathcal{B}(\mu)$，i 是被 μ 误分类的 d 个顶点之一，并且不存在另一个 $k\in[n]$ 使得 $(i,k)\in\mathcal{B}(\mu)$。因此，$|\mathcal{B}(\mu)|\leq d$。

利用 $\mathcal{A}(\mu)$、$\mathcal{B}(\mu)$ 和 $\mathcal{C}(\mu)$，我们可以将 $f(\mu)$ 分解如下：

$$\begin{aligned} f(\mu) &= \sum_{i\in M:\mu(i)\neq\pi_*(i)} \deg_{G_1\wedge_\mu G_2}(i) = \sum_{(i,j)\in\mathcal{A}(\mu)} A_{ij}B_{\mu(i)\mu(j)} \\ &= \sum_{(i,j)\in\mathcal{B}(\mu)} A_{ij}B_{\mu(i)\mu(j)} + \sum_{(i,j)\in\mathcal{C}(\mu)} A_{ij}B_{\mu(i)\mu(j)} \\ &= 2\sum_{(i,j)\in\mathcal{B}(\mu):i<j} A_{ij}B_{\mu(i)\mu(j)} + \sum_{(i,j)\in\mathcal{C}(\mu)} A_{ij}B_{\mu(i)\mu(j)} \end{aligned}$$

为简便起见，我们将第一个求和记为 $X_{\mathcal{B}}$，第二个求和记为 $X_{\mathcal{C}}$。事实证明，$X_{\mathcal{B}}$ 和 $X_{\mathcal{C}}$ 相互独立。实际上，注意到当且仅当 $\{\mu(a)\ \mu(b)\}=\{\pi_*(i)\ \pi_*(j)\}$时，$A_{ij}$ 和 $B_{\mu(a)\mu(b)}$ 才相关。根据 $\mathcal{B}(\mu)$ 的定义，对于 $(i,j)\in\mathcal{B}(\mu)$，有 $\{\mu(i)\ \mu(j)\}=\{\pi_*(i)\ \pi_*(j)\}$，由此可知，$X_{\mathcal{B}}$ 中求和项与 $X_{\mathcal{C}}$中求和项相互独立。相同的论证表明，构成 $X_{\mathcal{B}}$ 的各项也相互独立，其中 $A_{ij}B_{\mu(i)\mu(j)}=A_{ij}B_{\pi_*(i)\pi_*(j)}\sim\mathsf{Bern}(p_{ij}s^2)$。因此

$$X_{\mathcal{B}} \preceq \mathrm{Bin}(|\mathcal{B}(\mu)|,p_{\max}s^2) \preceq \mathrm{Bin}(d,p_{\max}s^2), \tag{A.1}$$

其中，在第二个随机支配中，我们用到了$|\mathcal{B}(\mu)|\leq d$。

处理 $X_{\mathcal{C}}$ 更为复杂，因为相应的求和项可能彼此相关。为解决这一问题，我们将 $\mathcal{C}(\mu)$划分为 $\mathcal{C}_1(\mu)$、$\mathcal{C}_2(\mu)$和 $\mathcal{C}_3(\mu)$，并定义

$$X_{\mathcal{C}_m} := \sum_{(i,j)\in\mathcal{C}_m(\mu):i<j} A_{ij}B_{\mu(i)\mu(j)}, \qquad m\in\{1,2,3\}.$$

特别注意，$X_{\mathcal{C}}\leq 2(X_{\mathcal{C}_1}+X_{\mathcal{C}_2}+X_{\mathcal{C}_3})$，其中因子 2 是为了考虑$(i,j)$和$(j,i)$可能同时属于 $\mathcal{C}(\mu)$的情况。

关键在于，我们将选择一种划分，使得对每个 $m\in\{1,2,3\}$, $X_{\mathcal{C}_m}$ 都是独立伯努利随机变量之和。为此，考虑 $\binom{M}{2}$中的两个无序对 $\{i,j\}$ 和 $\{a,b\}$。当且仅当以下两个条件之一成立时，随机变量 $A_{ij}B_{\mu(i)\mu(j)}$ 和$A_{ab}B_{\mu(a)\mu(b)}$ 相依：

$$\{\mu(i),\mu(j)\}=\{\pi_*(a),\pi_*(b)\}, \tag{A.2}$$
$$\{\mu(a),\mu(b)\}=\{\pi_*(i),\pi_*(j)\}. \tag{A.3}$$

现在，我们在顶点集$V(H)$上构造依赖图 $H := \{\{i,j\} : (i,j) \in \mathcal{C}(\mu)\}$，使得当且仅当 $A_{ij}B_{\mu(i)\mu(j)}$和 $A_{ab}B_{\mu(a)\mu(b)}$ 相关时，$\{i,j\},\{a,b\} \in V(H)$之间有一条边。根据条件(A.2)和(A.3)，由于H 中的每个顶点至多有两个邻居，因此 H 是3-可着色的。令 $\mathcal{C}_1,\mathcal{C}_2,\mathcal{C}_3$ 为对应于三种颜色的 $V(H)$的划分，则可知$X_{\mathcal{C}_1}, X_{\mathcal{C}_2}, X_{\mathcal{C}_3}$ 都是独立随机变量之和，正如所需。

接下来，我们研究 $X_{\mathcal{C}_1}$、 $X_{\mathcal{C}_2}$和$X_{\mathcal{C}_3}$的分布。对于每个$(i,j) \in \mathcal{C}(\mu)$，根据$\mathcal{C}(\mu)$的定义，我们有 $\{\pi_*(i)\ \pi_*(j)\} \neq \{\mu(i)\ \mu(j)\}$，因此

$$A_{ij}B_{\mu(i)\mu(j)} \sim \mathrm{Bern}(p_{ij}p_{\pi_*^{-1}(\mu(i)),\pi_*^{-1}(\mu(j))}s^2).$$

特别地，我们有 $A_{ij}B_{\mu(i)\mu(j)} \preceq \mathsf{Bern}(p_{\max}^2 s^2)$。此外，对于 $m \in \{1,2,3\}$,

$$X_{\mathcal{C}_m} \preceq \mathrm{Bin}(|\mathcal{C}_m(\mu)|, p_{\max}^2 s^2) \preceq \mathrm{Bin}(dn, p_{\max}^2 s^2), \quad \text{(A.4)}$$

其中，在第二个随机支配关系中，我们用到了$|\mathcal{C}_m(\mu)| \le |\mathcal{C}(\mu)| \le |\mathcal{A}(\mu)| \le dn$。

现在，我们可以如下界定所关注事件的概率：

$$\begin{aligned}
\mathbb{P}(f(\mu) \ge kd) &\le \mathbb{P}(2X_{\mathcal{B}} + X_{\mathcal{C}} \ge kd) \\
&\overset{(a)}{\le} \mathbb{P}(2(X_{\mathcal{B}} + X_{\mathcal{C}_1} + X_{\mathcal{C}_2} + X_{\mathcal{C}_3}) \ge kd) \\
&\overset{(b)}{\le} \sum_{m=1}^{3} \mathbb{P}(2X_{\mathcal{B}} + 6X_{\mathcal{C}_m} \ge kd) \\
&\overset{(c)}{\le} \sum_{m=1}^{3} e^{-\theta kd}\mathbb{E}\big[e^{2\theta X_{\mathcal{B}}}\big]\mathbb{E}\big[e^{6\theta X_{\mathcal{C}_m}}\big] \\
&\overset{(d)}{\le} 3e^{-\theta kd}(1 + p_{\max}s^2(e^{2\theta}-1))^d(1 + p_{\max}^2 s^2(e^{6\theta}-1))^{dn} \\
&\overset{(e)}{\le} 3\exp\big\{-d(\theta k - e^{2\theta}p_{\max}s^2 - ne^{6\theta}p_{\max}^2 s^2)\big\}.
\end{aligned}$$

在上面的推导中，(a)用到了 $X_{\mathcal{C}} \le 2(X_{\mathcal{C}_1} + X_{\mathcal{C}_2} + X_{\mathcal{C}_3})$；$(b)$是由并集界得到的；对于 $\theta > 0$，(c)由切尔诺夫界以及 $X_{\mathcal{B}}$ 和 $X_{\mathcal{C}_m}$的独立性推出；(d)通过对$X_{\mathcal{B}}$ 和 $X_{\mathcal{C}}$的矩母函数进行界定得到，这里用到了 $X_{\mathcal{B}}$ 和 $X_{\mathcal{C}_m}$ 可随机支配于二项随机变量这一事实（见(A.1)和(A.4)）；(e)用到了不等式$(1+x)^t \le e^{tx}$。 □

现在我们可以证明引理 III.4 了，这正是本节的主要目标。

引理 III.4 的证明。令 $\theta := c\log n$ for some $2/k < c < (2\alpha-1)/6$。由于假设 $k > 12/(2\alpha-1)$，这样的 c 存在。取此 θ 后，我们有

$$e^{2\theta}p_{\max}s^2 \le s^2 n^{-(\alpha+1)/3+o(1)} = o(1).$$

我们还有

$$ne^{6\theta}p_{\max}^2 s^2 \le s^2 n^{6c-(2\alpha-1)+o(1)} = o(1),$$

这是因为 $6c < 2\alpha - 1$。引理 A.4 现在表明

$$\mathbb{P}(f(\mu) \ge kd) \le 3\exp\{-d(ck\log n - o(1))\},$$

这进而意味着 $\xi \le 3e^{-ck\log n + o_{(1)}}$。由于我们选择了$c$ 使得 $ck > 2$，因此有 $\xi = o(n^{-2})$，结合引理 A.3 即可证明所需结果。 □

附录 B 推论的证明

在本节中，我们用 $d_{\min}\mathsf{min}_{i\in[n]} \sum_{j=1}^n p_{ij}$ 表示。 :=

A. 精确图匹配的证明

推论 III.6 的证明。令 $a \in [m]$，并假设 $i \in V_a$。当 n 足够大时，有

$$\begin{aligned}
d_{\min}s^2 &= \sum_{b=1}^{m} |V_b| q_{ab}s^2 - q_{aa}s^2 \\
&\ge (1+\epsilon)\log n - 1 \ge (1+\epsilon/2)\log n. \quad \text{(B.1)}
\end{aligned}$$

在上面的第一个不等式中，我们用到了 $q_{aa}s^2 \le 1$。应用定理 III.5 即可证明该推论。 □

推论 III.7 的证明。我们先证明，若满足 (III.2)，则 $p_{\max} \le n^{-\alpha+o(1)}$。对于任意 $i,j \in [n]$,

$$\frac{w_i w_j}{\sum_{i=1}^n w_i} \le \left(\frac{\max_{i\in[n]} w_i}{\sqrt{\sum_{i=1}^n w_i}}\right)^2 \le n^{-\alpha+o(1)}, \quad \text{(B.2)}$$

其中，最后一个界是 (III.2) 的推论。特别地，由于对于不同的 $i,j \in [n]$, $p_{ij} = w_i w_j/(\sum_{k=1}^n w_k)$，因此可得 $p_{\max} \le n^{-\alpha+o(1)}$。剩下的任务是给出最小期望度的下界。对于任意 $i \in [n]$，当 n 足够大时，有

$$\begin{aligned}
d_{\min}s^2 &= \min_{i\in[n]} \sum_{j\in[n]:j\neq i} \frac{w_i w_j s^2}{\sum_{k=1}^n w_k} \\
&\ge \min_{i\in[n]} \sum_{j=1}^{n} \frac{w_i w_j s^2}{\sum_{k=1}^n w_k} - p_{\max}s^2 \\
&\ge \min_{i\in[n]} w_i s^2 - n^{-\alpha+o(1)} \ge (1+\epsilon/2)\log n. \quad \text{(B.3)}
\end{aligned}$$

在上面的推导中，第一个不等式用到了 (B.2)。由此即可得到所需结果。 □

推论 III.8 的证明。由 (II.1) 可明显看出 $p_{\max} \le p$，因此 $p_{\max} \le n^{-\alpha+o(1)}$。在此情形下写出定理 III.5 中的条件 (III.1)，即可得到所需结果。 □

B. 部分图匹配的证明

推论 III.10 的证明。由于 $d_{\min} \ge (n-1)q_{\min}$，我们有 $R \le q_{\max}/q_{\min}$。此外，我们还有

$$\begin{aligned}
d_{\min}s^2 &= \min_{a\in[m]} \sum_{b=1}^{m} |V_b| q_{ab}s^2 - o(1) \\
&\ge \max\left\{\frac{q_{\max}}{q_{\min}}, 154\right\} \ge \max\{R, 154\},
\end{aligned}$$

上式中的等号由 (B.1) 得出。应用定理 III.9 即可得到所需结果。 □

推论 III.11 的证明。我们首先如下给出 R 的上界:

$$R \le \frac{p_{\max}}{p_{\min}} \le \left(\frac{w_{\max}^2}{\sum_{k=1}^n w_k}\right)\left(\frac{w_{\min}^2}{\sum_{k=1}^n w_k}\right)^{-1} = \left(\frac{w_{\max}}{w_{\min}}\right)^2.$$

此外，我们有

$$\begin{aligned} d_{\min}s^2 &= w_{\min}s^2 - o(1) \\ &\ge \max\left\{\left(\frac{w_{\max}}{w_{\min}}\right)^2, 154\right\} \\ &\ge \max\{R, 154\}, \end{aligned}$$

现在应用定理 III.9 即可得到所需结果。 □

推论 III.12 的证明。简记 $D := \min_{i\in[n]} |\{j \in [n] \setminus \{i\} : \|x_i - x_j\| \le r\}|$。注意到 $p_{\max} \le p$ 对于该模型成立，我们有上界$R \le n/D$。注意到 $d_{\min} = pD$，当且仅当 $ps^2D \ge \max\{n/D, 154\}$，或等价地，当且仅当 $ps^2D \ge \max\{s\sqrt{np}, 154\}$时，定理 III.9 的条件得到满足。最后，由于 $ps^2D = \min_{i\in[n]} |\{j \in [n] : \|x_i - x_j\| \le r\}| - o(1)$，所需结果便可直接得出。 □

附录 C 度相关引理的证明

引理 IV.1 的证明。我们先写出 $\mathbb{E}[|Z_k|] = \sum_{i=1}^n \mathbb{P}(\deg_G(i) \le k)$；接下来对求和中的各项进行界定。对于任意 $i \in [n]$，我们有分布表示 $\deg_G(i) \overset{d}{=} \sum_{j=1}^n Y_j$，其中各个 Y_j 相互独立，且服从 $Y_j \sim \mathrm{Bern}(p_{ij})$。对于 $\theta > 0$，由马尔可夫不等式可得

$$\begin{aligned} &\mathbb{P}(\deg_G(i) \le k) \le e^{\theta k}\mathbb{E}[\exp(-\theta \deg_G(i))] \\ &= e^{\theta k}\prod_{j=1}^n \mathbb{E}[\exp(-\theta Y_j)] = e^{\theta k}\prod_{j=1}^n \big(1 + p_{ij}(e^{-\theta} - 1)\big) \\ &\le \exp\left\{\theta k - (1 - e^{-\theta})\sum_{j=1}^n p_{ij}\right\}. \end{aligned}$$

所需结论可由令 $\theta = -\log(1 - c)$ 得出。 □

引理 IV.2 的证明。我们的策略是界定 $|Z_k|$ 的方差，并应用切比雪夫不等式。令 $E_i := \mathbf{1}(\deg_G(i) \le k)$，于是 $|Z_k| = \sum_{i=1}^n E_i$，且

$$\mathrm{Var}(|Z_k|) = \sum_{i=1}^n \mathrm{Var}(E_i) + \sum_{i,j\in[n]:i\ne j} \mathrm{Cov}(E_i, E_j).$$

由于 E_i 是一个指示变量，我们可以对 $\mathrm{Var}(E_i) \le \mathbb{E}[E_i]$进行界定，从而将第一个求和界定为 $\mathbb{E}[|Z_k|]$。我们现在转而考虑第二个求和中的各项。定义指示变量 $A_{ij} := \mathbf{1}((i, j) \in E(G))$，我们有

$$\begin{aligned} \mathbb{E}[E_iE_j] &= \mathbb{E}[\mathbb{E}[E_i|A_{ij}]\mathbb{E}[E_j|A_{ij}]] \\ &\overset{(a)}{\le} p_{ij} + (1 - p_{ij})\mathbb{E}[E_i|A_{ij} = 0]\mathbb{E}[E_j|A_{ij} = 0] \\ &\overset{(b)}{\le} p_{ij} + (1 - p_{ij})\frac{\mathbb{E}[E_i]\mathbb{E}[E_j]}{(1 - p_{ij})^2} \\ &\le p_{ij} + (1 - p_{ij})^{-1}\mathbb{E}[E_i]\mathbb{E}[E_j]. \end{aligned} \tag{C.1}$$

上文中，(*a*) 成立是因为 $A_{ij} = 1$ 以概率 p_{ij} 发生，且 $|E_i|, |E_j| \le 1$；(*b*) 用到了如下关系

$$\mathbb{E}[E_i|A_{ij} = 0] = \frac{\mathbb{P}(E_i = 1, A_{ij} = 0)}{\mathbb{P}(A_{ij} = 0)} \le \frac{\mathbb{E}[E_i]}{1 - p_{ij}}.$$

利用 (C.1)，我们可以将 $\mathrm{Cov}(E_i, E_j)$ 界定为

$$\begin{aligned} \mathrm{Cov}(E_i, E_j) &= \mathbb{E}[E_iE_j] - \mathbb{E}[E_i]\mathbb{E}[E_j] \\ &\le p_{ij} + \big((1 - p_{ij})^{-1} - 1\big)\mathbb{E}[E_i]\mathbb{E}[E_j] \\ &\le p_{ij} + 2p_{ij}\mathbb{E}[E_i]\mathbb{E}[E_j] \\ &\le 3p_{ij} \le 3p_{\max}. \end{aligned} \tag{C.2}$$

上文中，第三行的不等式成立是因为对于充分小的 x，$(1-x)^{-1} - 1 = x/(1 - x) \le 2x$；第三行中的第一个不等式成立是因为 $0 \le E_i, E_j \le 1$。综合以上结果可得

$$\mathrm{Var}(|Z_k|) \le \mathbb{E}[|Z_k|] + 3n^2p_{\max} = o(n^{3/2}), \tag{C.3}$$

其中，第一个不等式源于 (C.2) 中的协方差界。最终的 $o(n^{3/2})$ 界成立是因为 $|Z_k| \le n$，且我们假设 $p_{\max} = o(1/\sqrt{n})$。有了方差估计 (C.3)，切比雪夫不等式便推出所需结论。 □

附录 D Łuczak 扩张引理的证明

引理 IV.6 的证明。该证明采用了与 Łuczak [21] 类似的方法。定义如下的顶点子集序列$\{U_m\}_{m\ge 0}$。令 $U_0 := Z_{k+1}$。对于 $m \ge 0$，若存在 $v \in V \setminus U_m$，使其在 U_m 中至少有 3 个邻居，则令$U_{m+1} := U_m \cup \{v\}$。若不存在这样的顶点，我们就停止构造。

令 U_ℓ 为该构造得到的最终集合，并用反证法假设 $|U_\ell| > 3|U_0|$。那么必定存在 $0 \le b \le \ell$，使得 $|U_b| = 3|U_0|$，因为每次迭代构造只添加一个顶点。令 E_b 表示 $G\{U_b\}$ 中边的数量，则有

$$E_b \ge 3(|U_b| - |U_0|) = 2|U_b|.$$

由于$|U_b| = 3|U_0| \le \frac{3}{4\gamma^2}n$，上面的式子与引理 IV.5 中的事件矛盾。因此 $|U_\ell| \le 3|U_0| = 3|Z_{k+1}|$。

最后，我们将此分析与 G 的 k-核联系起来。注意，如果 $v \in V \setminus U_\ell$，那么 $\deg_G(v) \ge k + 2$，且 v 在 U_ℓ 中至多有 2 个邻居。因此，$G\{V \setminus U_\ell\}$ 中的最小度至少为 k，所以 $F_k \subseteq U_\ell$。 □