arXiv:2602.05852v1 [cs.LG] 5 Feb 2026

数据块模型中的精确恢复

Amir R. Asadi*[1]、Akbar Davoodi[2], 、Ramin Javadi†[3]和Farzad Parvaresh‡[4]

[1]剑桥大学数学科学中心统计实验室，剑桥，英国[3]伊斯法罕理工大学数学科学系，伊斯法罕，伊朗[4]伊斯法罕大学工程学院电气工程系，伊斯法罕，伊朗

2026年2月5日

摘要

网络中的社区检测是机器学习和统计推断中的一个基础问题，在社交网络、生物系统和通信网络中均有应用。随机块模型（SBM）是研究社区结构的经典框架，而精确恢复——以高概率识别真实社区——是一个核心理论问题。经典研究刻画了仅基于图连通性的精确恢复相变，但许多现实网络还包含节点属性或标签等额外数据。本文研究数据块模型（DBM）中的精确恢复；该模型是在随机块模型中加入节点关联数据，由Asadi、Abbe和Verd´u（2017）加以形式化。我们引入Chernoff–TV散度，并用其刻画数据块模型的尖锐精确恢复阈值。我们还提出一种高效算法，可达到该阈值，并给出相匹配的逆向结果，证明在阈值以下无法实现精确恢复。最后，模拟结果验证了我们的发现，并展示了在社区检测中将顶点数据作为侧信息的优势。

1 引言

1.1 动机与背景

社区检测和聚类是机器学习中的基础问题，应用于各个领域。其目标是通过分析交互模式，揭示网络中隐藏的社区结构。许多现实世界的数据集都可以表示为具有社区结构的网络，包括社交网络、生物网络和协作网络（Girvan and Newman，2002；Backstrom et al.，2006）。在这些情形下，社区会自然地将数据划分为由具有共同属性的节点组成的群组；具有相似特征的节点更有可能属于同一群组。

社区检测支持众多任务，包括推荐系统（Gasparetti等人，2021）、欺诈检测（Sarma等人，2020）以及生物网络分析（Rahiminejad

*asadi@statslab.cam.ac.uk
†rjavadi@iut.ac.ir
‡f.parvaresh@eng.ui.ac.ir

等人，2019），并且通常被视为无监督学习问题（Abbe，2017）。自20世纪80年代以来，该领域的研究显著扩展，催生了在机器学习、计算机科学、网络科学、社会科学和统计物理学中发展出的大量模型与算法。

块模型是一类带有植入簇的随机图（Goldenberg 等，2010）。其中最经典的模型是随机块模型（SBM），这是一种用于生成具有植入社区结构的随机图的生成模型，也是社区检测领域广泛研究的框架。随机块模型为研究社区检测提供了理论基础：它将图建模为节点属于潜在社区、边则根据社区内部和社区间的连接概率生成。通常，两个节点之间存在边的概率取决于它们的成员关系，而且社区内部的边往往比社区之间更密集。随机块模型最初由 Holland 等（1983）在社交网络分析领域提出，此后已成为统计学、机器学习和网络科学中的基础工具。它为社区结构提供真实标注，因此是评估社区检测算法的重要基准，并使严谨的理论分析成为可能。尽管模型简单，随机块模型仍能捕捉社区检测中的关键瓶颈，并启发了高效算法。与信息论中的离散无记忆信道类似，随机块模型既易于处理又富有启发性，推动了可扩展至更复杂真实世界图的方法的发展。它在不同领域得到广泛应用，这一点从其丰富的名称和解读中可见一斑：理论计算机科学中的植入划分模型（McSherry，2001）、数学中的非齐次随机图（Bollob´等，2007）、物理学中的植入自旋玻璃模型（Decelle 等，2011），以及随机矩阵理论中的低秩随机矩阵模型（McSherry，2001）。这些不同视角彰显了该模型的多功能性及其在社区检测研究中的核心作用。

传统的随机块模型主要关注网络连通性，但现实世界的网络通常还包含其他信息，例如节点属性、标签或特征。例如，在社交媒体网络中，个体不仅会建立友谊，还会发布帖子，并展现出位置、职业或兴趣等特征。同样，引文网络会根据研究主题对论文进行分类，生物网络则会依据功能属性对蛋白质进行分组。在这些示例中，社区结构不仅源于连通性，也源于这些共享的外部特征。因此，社区效应既会影响网络的结构，也会影响与其节点相关的数据分布。网络拓扑与节点特定数据之间的这种相互作用表明，纳入顶点数据能够显著提升社区检测的效果。

对这一情形进行建模的一种自然方式是采用数据块模型 (DBM)。它是随机块模型（SBM）的扩展，将额外数据作为节点属性纳入模型。该模型最初由Abbe (2016)提出，随后由 Asadi 等 (2017)加以形式化。DBM 假设属于同一社区的节点，其属性共享相同的先验分布。 Asadi 等 (2017)通过基于 Chernoff 信息的论证，将Abbe 和 Sandon(2015a)的 CH 散度框架加以扩展，开创了研究此类顶点数据如何影响精确恢复阈值的工作。DBM 为分析侧信息在精确恢复中的作用提供了严谨的理论基础，将基于网络的方法与数据驱动的聚类方法联系起来。将节点特定数据用作侧信息，对于实现精确恢复可能至关重要，尤其是在仅凭图结构不足以完成恢复的情形下。通过同时纳入连通模式和节点特征，DBM 能够更细致地揭示正确识别社区所需的条件。接下来，我们将回顾这一方向上更多相关工作。

1.2 进一步相关工作

大量文献研究了随机块模型中社区检测的基本极限与算法。对于双社群对称随机块模型，精确恢复在常数度数区间内不可能实现，但在对数度数区间内会出现明显的相变；此外，该阈值可以通过高效算法达到（Abbe 等人，2016；Mossel 等人，2016）。对于一般随机块模型，Abbe 和 Sandon 提出了 CH 散度，确立了精确恢复的充要阈值，并给出了能够达到该阈值的近线性时间算法（Abbe 和 Sandon，2015a；b）；另见综述（Abbe，2017）。

附加在顶点上的侧信息已在若干互补模型中得到研究。对于平衡二元随机块模型，Saad 和 Nosratinia (2018) 分析了噪声标签（错误概率 α）和部分揭示的标签（在 $1-\epsilon$比例的节点中已知），确定了侧信息在何种情况下以及在多大程度上能够降低精确恢复阈值。这项研究在 Saad 和 Nosratinia(2019) 中扩展至连续值侧信息，并弥合了 Saad 和 Nosratinia (2018) 中充要条件之间的差距。当每个节点观测到 $O(\log n)$ 个独立同分布属性样本时，也已获得精确阈值（Sima 等人，2021），并提出了能够达到该极限的高效半定规划算法。与此一致，Esmaeili 等人 (2021) 证明，在若干辅助信息模型（噪声标签、部分揭示的标签以及每个节点的多个特征）中，半定规划算法能够达到最大似然精确恢复阈值。

一类密切相关的研究探讨了上下文随机块模型，其中除了图之外，还会观测节点协变量，并将二者结合起来用于社区检测。从建模角度看，上下文 SBM 可视为数据块模型的特例，只需指定特定的节点属性分布族及其对潜在社区标签的依赖关系即可得到。Deshpande 等人(2018) 提出了一种上下文 SBM，其中边的形成和节点特征均由潜在社区驱动，并确定了这样一些情形：结合图和协变量信息能够实现仅凭图无法达到的恢复效果。在此框架基础上，Braun 等人(2022) 提出了一种用于上下文 SBM 的迭代聚类算法，并在图稀疏性和特征分布满足适当假设时，证明了其最优性保证。这些研究共同表明，以节点特征为形式的辅助信息会显著影响恢复阈值和算法性能。

带节点属性的随机块模型也曾从信息论视角进行研究。特别是，Dreveton 等人(2023) 研究了带节点属性的 SBM 的精确恢复问题，并基于散度准则刻画了一个相变，同时采用 Bregman 硬聚类的视角。本文在对数度数设定下重新考察这一情形，并基于 Chernoff–TV 散度给出更精细的刻画。在这一受限设定下（详见小节 4.2），我们证明，按原文表述，(Dreveton 等人, 2023) 中的条件并非精确恢复的必要条件，并推导出一个修正后的阈值。

除了这些设定外，其他研究还考察了不同形式的辅助信息和推断目标。在一般随机块模型中，针对部分揭示标签的标签传播已有研究（Saha 等人，2020）。在Esmaeili 和 Nosratinia (2020中，研究者考虑了一个包含两个潜在标签变量的二元随机块模型，其中边概率取决于这两个潜在因子；研究还给出了基于 SDP 的恢复保证。在Yoon 等人(2018中，研究者考虑了具有 k 个社区的对称随机块模型，其中每个节点都配有一个二元 m维向量。这里，侧信息是相应向量的含噪且部分观测版本。任务是恢复 k 个社区以及节点的特征向量，并刻画了该问题的信息论极限和计算极限。针对这些问题的超图类比也已得到发展。对于均匀超图随机块模型，Zhang 和 Tan(2021提出了

精确恢复阈值（通过广义 CH 型判据），而Zhao 等人 (2021则在节点标签含噪或仅部分揭示时推导出了紧确界。

1.3 我们的贡献

本文的主要贡献如下：

1. 我们引入 Chernoff–TV 散度，并用它精确刻画数据块模型（DBM）中的精确恢复阈值。该准则适用于任意数量的社区；当属性不提供信息时，它恢复经典 SBM 阈值；当图不提供信息时，它退化为仅数据极限。2. 我们设计了一种多项式时间算法，联合利用边数据和节点数据以达到该 Chernoff–TV 阈值，从而在信息论极限处确立计算可行性。3. 我们证明了一个匹配的反向结论：低于 Chernoff–TV 阈值时，精确恢复在信息

理论上不可能，即使计算资源不受限制。4. 我们阐明了本文阈值与带节点属性的 SBM 中精确恢复现有准则之间的关系，并在小节 4.2 的设定下，对（Dreveton 等人，2023）中先前声称的必要性方向给出了修正表述。5. 我们在合成数据集上通过实验验证了该理论，实验呈现出尖锐的经验相变，并表明纳入节点属性能带来一致的增益。

1.4 论文结构

本文结构如下：第2 节介绍数据块模型和基本定义。第3节汇总预备结果，包括 CH 散度及相关工具。第4 节定义 Chernoff–TV 散度，陈述数据块模型中精确恢复的精确阈值，并介绍两阶段恢复算法及其可实现性和反向分析。第5节探讨实验设置、展示实验结果，并讨论其意义。最后，第6节总结本文的主要发现，并提出未来研究方向。

1.5 符号约定

本文通篇使用以下符号约定。对于任意正整数 k，记 $[k]$ 表示集合 $\{1, 2, \ldots, k\}$。给定矩阵 A，用 A_r表示其第 r 列。符号$\mathcal{P}_{\boldsymbol{\mu}}$ 表示各分量相互独立、均值向量为 $\boldsymbol{\mu}$的多元泊松分布。对于向量 $v \in \mathbb{R}^n$，记 diag(v)表示以 v 的各分量为对角元素的 $n \times n$ 对角矩阵。我们用 $\mathbb{R}_+$ 和 $\mathbb{Z}_+$ 分别表示非负实数集和非负整数集。 $P \ll Q$ 表示概率分布 P 关于 Q绝对连续。

2 模型与定义

本节将介绍模型并给出一些基本定义。

定义 2.1。令 $n \in \mathbb{N}$ 表示顶点的数量，$k \in \mathbb{N}$ 表示社区的数量。固定 $P = (p_1, \dots, p_k)$，这是定义在 $[k]$ 上的概率质量函数，称为先验分布，并令

$$\mathbf{W} = [W_{ab}]_{1 \le a,b \le k} \in [0,1]^{k \times k}$$

是一个边概率的对称矩阵。

随机块模型（G_n, X^n）定义如下。社区标签 X^n =（$X_1, \dots, X_n$）是独立同分布的随机变量，满足

$$\mathbb{P}(X_i = a) = p_a, \quad a \in [k].$$

随机图 $G_n = (Y_{ij})_{1 \le i < j \le n}$ 定义在顶点集 $\mathcal{V}_n = \{v_1, \dots, v_n\}$上，其中每个$Y_{ij} \in \{0,1\}$ 表示顶点 v_i 与 v_j之间一条边的缺失（0）或存在（1）。在给定 X^n的条件下，边变量 $\{Y_{ij}\}$相互独立并满足

$$\mathbb{P}\big(Y_{ij} = 1 \mid X_i = a, X_j = b\big) = W_{ab}, \quad 1 \le i < j \le n, \quad a, b \in [k].$$

我们记

$$(G_n, X^n) \sim \mathrm{SBM}(n, k, P, \mathbf{W})$$

以表示（G_n, X^n）服从一个具有 n 个顶点、 k 个社区、先验分布 P以及边概率矩阵$\mathbf{W}$的SBM。

本文关注对数度数区间，其中期望度数为$\Theta(\log n)$。相应地，我们假设对每个 n ，边概率矩阵满足$\mathbf{W}^{(n)} = (\log\ n/n)\mathbf{Q}$，其中$\mathbf{Q} \in \mathbb{R}_+^{k \times k}$为某个固定的对称矩阵。

定义 2.2. 数据块模型（DBM）由三元组 (G_n, X^n, U^n) 定义。其中，(G_n, X^n) 是一个随机块模型，称为底层随机块模型，而

$$U^n := (U^{(1)}, U^{(2)}, ..., U^{(n)})$$

表示附属于各顶点的数据文件集合，其中 $U^{(i)}$ 是位于顶点 i 处的数据文件，$i = 1, 2, ..., n$。 $U^{(i)}$ 的分布取决于顶点 i 所属的社区。

我们假设 (G_n, X^n, U^n) 满足马尔可夫链 $G_n \leftrightarrow X^n \leftrightarrow U^n$，这意味着给定社区标签 X^n 后，图 G_n 与数据 U^n 相互独立。此外，给定 X^n 后，数据文件 $\{U_{(i)}\}_{i=1}^n$ 相互独立，并满足 $U_{(i)} \mid \{X_i = x\} \sim P_{U|X}(\cdot \mid x)$，其中 $x \in [k]$。因此，(G_n, X^n, U^n) 的联合分布为

$$\begin{aligned}
&\mathbb{P}(G_n = g_n, X^n = x^n, U^n = u^n) \\
&= \mathbb{P}\big(X^n = x^n\big)\mathbb{P}\big(G_n = g_n \mid X^n = x^n\big)\mathbb{P}\big(U^n = u^n \mid X^n = x^n\big) \\
&= \prod_{i=1}^n P(x_i) \prod_{1 \le i < j \le n} \big((W_{x_i x_j})^{y_{ij}} (1 - W_{x_i x_j})^{1 - y_{ij}}\big) \prod_{i=1}^n P_{U|X}\big(u^{(i)} \mid x_i\big).
\end{aligned}$$

我们将 $P_{U|X}$ 称为数据—社区随机变换。我们记

$$(G_n, X^n, U^n) \sim \mathrm{DBM}(n, k, P, \mathbf{W}, P_{U|X})$$

表示（G_n， X^n， U^n）服从一个 DBM，其中包含 n 个顶点、 k 个社区、先验分布 P、边概率矩阵 $\mathbf{W}$ 以及数据—社区随机变换 $P_{U|X}$。

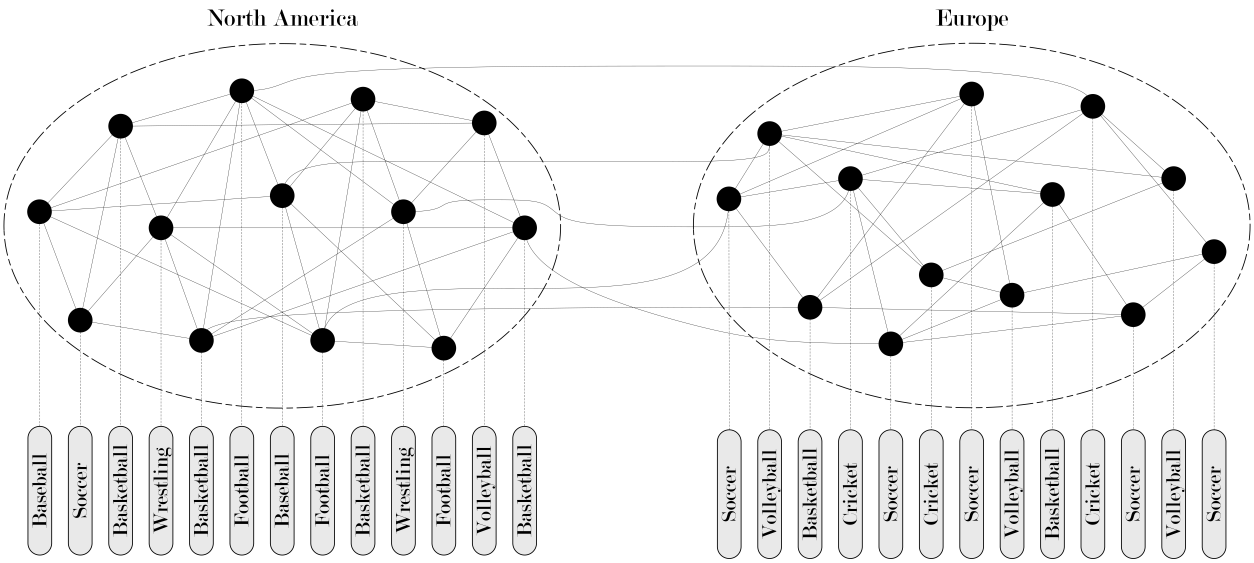


图 1：DBM 示例

请考虑以下 DBM 示例：

例 2.3。 图 1 展示了一个包含两个社区的数据块模型实例，这两个社区分别对应居住在北美洲和欧洲的人群。每个顶点代表一个人，每条边表示两个人之间的友谊，每个顶点都关联一个记录该人最喜爱的运动的数据文件。

该模型在图和数据两方面都呈现出结构。两个社区中的体育运动分布有所不同，并且图中社区内部的边密度高于社区之间的边密度。此示例突出了在 DBM 中，社区成员关系如何同时影响连通模式和节点级属性。

我们对一系列随机块模型的近乎精确恢复和精确恢复的可解性定义如下：

定义 2.4。 设 P 是 $[k]$ 上的先验分布， $\{\mathbf{W}^{(n)}\}_{n=1}^{\infty}$ 是一系列 $k \times k$ 边概率矩阵。对于每个 $n \geq 1$，令 $(G_n,\ X^n) \sim \mathrm{SBM}(n,\ k,\ P,\ \mathbf{W}^{(n)})$。如果对于每个 $n \geq 1$，都存在一种算法，以 G_n 作为输入，并输出 $X^n = X^n(G_n)$，使得以下条件成立，则序列 $\{(G_n,\ X^n)\}_{n=1}^{\infty}$ 上的下列恢复问题是可解的：

- 近乎精确恢复：

$$\mathbb{P}\left[\max_{\pi \in \mathrm{Sym}(k)} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\hat{X}_i = \pi(X_i)\}} \geq 1 - o(1)\right] = 1 - o(1).$$

- 精确恢复：

$$\mathbb{P}\left[\exists\, \pi \in \mathrm{Sym}(k) :\ \hat{X}^n = \pi(X^n)\right] = 1 - o(1).$$

这里，$\mathrm{Sym}(k)$ 表示 $[k]$ 上的对称群，且 $\pi(X^n) :=(\pi(X_1) \ldots, \pi(X_n))$.

类似地，我们如下定义一列 DBM 的近乎精确恢复和精确恢复的可解性：

定义 2.5。设 P 是 $[k]$ 上的先验分布，$\{\mathbf{W}^{(n)}\}_{n=1}^{\infty}$ 是一列 $k \times k$ 边概率矩阵，且 $\{P_{U|X}^{(n)}\}_{n=1}^{\infty}$ 是一列数据—社区随机变换。对每个 $n \geq 1$，令 $(G_n, X^n, U^n) \sim \mathrm{DBM}(n, k, P, \mathbf{W}^{(n)}, P_{U|X}^{(n)})$。若对每个 $n \geq 1$，都存在一个算法，以 G_n 和 $\hat{} \hat{} U^n$ 为输入，并输出 $X^n = X^n(G_n, U^n)$，使得

- 近乎精确恢复：

$$\mathbb{P}\left[\max_{\pi \in \mathrm{Sym}(k)} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\hat{X}_i = \pi(X_i)\}} \geq 1 - o(1)\right] = 1 - o(1).$$

- 精确恢复：

$$\mathbb{P}\left[\exists\, \pi \in \mathrm{Sym}(k) :\ \hat{X}^n = \pi(X^n)\right] = 1 - o(1).$$

这里，$\mathrm{Sym}(k)$ 表示作用于 $[k]$ 上的对称群，且 $\pi(X^n) := (\pi(X_1) \ldots, \pi(X_n))$.

3 预备知识

本节介绍支撑我们分析的主要信息论量。我们首先介绍 Chernoff–Hellinger（CH）散度，它在刻画对数度数区间内随机块模型的精确恢复阈值方面发挥着核心作用。

定义 3.1（Chernoff–Hellinger 散度）。设 R 和 S 是定义在有限集合 $\mathcal{X}$ 上的正测度（不一定是概率测度）。R 和 S 之间的 Chernoff–Hellinger（CH）散度定义为

$$D_+(R\|S) := \max_{t \in [0,1]} \sum_{x \in \mathcal{X}} f_t\left(\frac{R(x)}{S(x)}\right) S(x),$$

其中 $f_t(y) := 1 - t + ty - y^t$。

术语 *Chernoff–Hellinger* 散度 由Abbe 和 Sandon (2015a) 引入，旨在强调它与 Chernoff 信息和 Hellinger 散度的联系。它之所以重要，是因为它为对数度数区间内随机块模型的精确恢复阈值提供了明确且尖锐的刻画，以下定理对此进行了形式化表述。

定理 3.2 (Abbe 和 Sandon (2015a))。设 P 是 $[k]$上的先验分布，且 $\mathbf{Q} \in \mathbb{R}_+^{k \times k}$是一个对称矩阵。对每个 $n \geq 1$，令

$$(G_n, X^n) \sim \mathrm{SBM}\left(n, k, P, \tfrac{\log n}{n}\mathbf{Q}\right).$$

序列 $\{(G_n, X^n)\}_{n \geq 1}$ 的精确恢复是可实现的（且可高效实现），当且仅当

$$D_+((\mathrm{diag}(P)\mathbf{Q})_i \,\|\, (\mathrm{diag}(P)\mathbf{Q})_j) > 1,$$

对于所有 i、$j \in [k]$ 满足 $i \neq j$。逆命题是，如果存在 $i \neq j$ 使得 t

$$D_+((\mathrm{diag}(P)\mathbf{Q})_i \,\|\, (\mathrm{diag}(P)\mathbf{Q})_j) < 1,$$

其中，$(\mathrm{diag}(P)\mathbf{Q})_i$表示$\mathrm{diag}(P)\mathbf{Q}$的第 i列。

定理3.2 表明，在对数度数区间内，SBM 的精确恢复存在明显的相变。这一转变由以下量决定：

$$\min_{i\neq j} D_+((\mathrm{diag}(P)\mathbf{Q})_i \,\|\, (\mathrm{diag}(P)\mathbf{Q})_j)\,.$$

接下来我们回顾 CH 散度的另一种表示形式，该表示形式由 Asadi 等 （2017）推导得出，并将其与多元泊松分布之间的 Chernoff 信息联系起来。为此，我们首先回顾两种标准的散度度量。

定义 3.3 （相对熵）。设 P_X 和 Q_X 是定义在同一集合 $\mathcal{A}$ 上的离散概率分布。P_X 与 Q_X 之间的相对熵（或 Kullback–Leibler 散度）定义为

$$D(P_X\|Q_X) = \mathbb{E}\left[\log\left(\frac{P_X(X)}{Q_X(X)}\right)\right],$$

其中，当 $P_X \ll Q_X$ 时，$X \sim P_X$；否则，$D\left(P_X\|Q_X\right) = \infty$。

定义 3.4 （Chernoff 信息（Chernoff，1952））。设 P_1 和 P_2 是定义在有限或可数集合 $\mathcal{X}$上的离散概率分布。P_1 与 P_2 之间的 Chernoff 信息定义为

$$C(P_1\|P_2) := -\log\left(\min_{\lambda\in[0,1]} \sum_{x\in\mathcal{X}} P_1(x)^{\lambda} P_2(x)^{1-\lambda}\right). \tag{1}$$

下面的命题确立了对于具有独立坐标的多元泊松分布，CH 散度与 Chernoff 信息之间的一个关键恒等式。

命题 3.5 (CH 散度即 Chernoff 信息（Asadi 等，2017））。设 $\mathbf{a}$，$\mathbf{b}$ $\in \mathbb{R}^n_+$。则

$$D_+(\mathbf{a}\|\mathbf{b}) = C(\mathcal{P}_{\mathbf{a}}\|\mathcal{P}_{\mathbf{b}})\,,$$

其中 $\mathcal{P}_{\mathbf{a}}$ 表示均值向量为 $\mathbf{a}$ 的乘积泊松分布。

最后，我们回顾总变差距离的定义。

定义 3.6 （全变差）。设 P和 Q是可测空间 Ω 上的概率分布。它们的总变差距离定义为

$$\mathrm{TV}(P,Q) := \sup_{A\subseteq\Omega} |P(A) - Q(A)|.$$

4 主要理论结果

4.1 Chernoff–TV 散度

我们首先介绍一种新的散度度量，称为*Chernoff–TV* 散度。该量以统一的方式刻画图结构和节点属性对数据块模型精确恢复阈值的共同贡献。

定义 4.1 （Chernoff–TV 散度）。设 $P_X^{(1)}$和 $P_{(2)\,X}$是有限集合 $\mathcal{X}$上的离散概率测度，$Q_U^{(1)}$和 $Q_U^{(2)}$是有限集合 $\mathcal{U}$上的离散概率测度。两对测度 $\left(P_X^{(1)}, Q_U^{(1)}\right)$和 $\left(P_X^{(2)}, Q_U^{(2)}\right)$之间的 Chernoff–TV 散度定义为

$$D_{\mathrm{CT}}\left(P_X^{(1)}, Q_U^{(1)} \middle\| P_X^{(2)}, Q_U^{(2)}\right) :=$$
$$-\log\left(\sum_{u\in\mathcal{U}} \min_{\lambda_u\in[0,1]} \sum_{x\in\mathcal{X}} \left(P_X^{(1)}(x)Q_U^{(1)}(u)\right)^{\lambda_u} \left(P_X^{(2)}(x)Q_U^{(2)}(u)\right)^{1-\lambda_u}\right).$$

下一个命题表明，在特殊情况下， D_{CT} 可以还原经典信息度量。

命题 4.2. 设 $P_{(1)X}$、$P_{(2)X}$ 为 $\mathcal{X}$ 与 $Q_U^{(1)}$ 上的概率质量函数，且 $Q_U^{(2)}$ 为 $\mathcal{U}$ 上的概率质量函数。

(i) 若 $Q_U^{(1)} = Q_U^{(2)}$ ，则

$$D_{\mathrm{CT}}\Big(P_X^{(1)}, Q_U^{(1)} \,\Big\|\, P_X^{(2)}, Q_U^{(2)}\Big) = C\Big(P_X^{(1)} \,\Big\|\, P_X^{(2)}\Big).$$

(ii) 若 $P_X^{(1)} = P_{(2)X}$，则

$$D_{\mathrm{CT}}\Big(P_X^{(1)}, Q_U^{(1)} \,\Big\|\, P_X^{(2)}, Q_U^{(2)}\Big) = -\log\Big(1 - \mathrm{TV}\Big(Q_U^{(1)}, Q_U^{(2)}\Big)\Big).$$

证明. (i) 若 $Q_U^{(1)} = Q_U^{(2)} =: Q_U$，则对 $\{\lambda_u\}_{u\in\mathcal{U}}$ 的任意选取，

$$\sum_{u\in\mathcal{U}}\sum_{x\in\mathcal{X}} \big(P_X^{(1)}(x)Q_U(u)\big)^{\lambda_u}\big(P_X^{(2)}(x)Q_U(u)\big)^{1-\lambda_u} = \sum_{u\in\mathcal{U}} Q_U(u)\sum_{x\in\mathcal{X}} (P_X^{(1)}(x))^{\lambda_u}(P_X^{(2)}(x))^{1-\lambda_u}.$$

关于 $\lambda_u \in [0,1]$ 的最小值可以对每个 u 分别取，从而得到

$$\min_{\lambda_u\in[0,1]} \sum_{x\in\mathcal{X}} (P_X^{(1)}(x))^{\lambda_u}(P_X^{(2)}(x))^{1-\lambda_u} = \min_{\lambda\in[0,1]} \sum_{x\in\mathcal{X}} (P_X^{(1)}(x))^{\lambda}(P_X^{(2)}(x))^{1-\lambda} = C\big(P_X^{(1)} \| P_X^{(2)}\big),$$

因此，关于 u 的求和等于 $C(P_{(1}\|P_X^{(2)})$。

(ii) 若 $P_{(1)X} = P_{(2)X} =: P_X$，则对每个 u 和任意 $\lambda_u \in [0,1]$，

$$\begin{aligned}\sum_{x\in\mathcal{X}} \big(P_X(x)Q_U^{(1)}(u)\big)^{\lambda_u}\big(P_X(x)Q_U^{(2)}(u)\big)^{1-\lambda_u} &= \Big(Q_U^{(1)}(u)\Big)^{\lambda_u}\Big(Q_U^{(2)}(u)\Big)^{1-\lambda_u}\sum_{x\in\mathcal{X}} P_X(x) \\ &= \Big(Q_U^{(1)}(u)\Big)^{\lambda_u}\Big(Q_U^{(2)}(u)\Big)^{1-\lambda_u}.\end{aligned}$$

对 $\lambda_u \in [0,1]$ 取最小值得到 $\min\{Q_U^{(1)}(u),\ Q_U^{(2)}(u)\}$，因此

$$\sum_{u\in\mathcal{U}} \min_{\lambda_u\in[0,1]} \Big(Q_U^{(1)}(u)\Big)^{\lambda_u}\Big(Q_U^{(2)}(u)\Big)^{1-\lambda_u} = \sum_{u\in\mathcal{U}} \min\{Q_U^{(1)}(u), Q_U^{(2)}(u)\}.$$

最后，对于离散分布，$\sum_u \min\{Q^{(1)}(u),\ Q^{(2)}(u)\} = 1 - \mathrm{TV}(Q^{(1)},\ Q^{(2)})$，从而得到所需结果。

□

命题 4.2 表明，Chernoff–TV 散度自然地介于 Chernoff 信息与全变差的一个函数之间。在我们的应用中，分布 P_X 对应于 基于图的 贡献（由 SBM 所诱导的邻域统计量产生），而 Q_U 对应于 数据 （辅助信息）贡献，该贡献由节点属性承载；精确的对应关系见下一小节。

引理 4.3。设 $P_{(1)X}, P_{(2)X}$ 是定义在 $\mathcal{X}$ 上的概率质量函数，且 $Q_U^{(1)}, Q_U^{(2)}$ 是定义在 $\mathcal{U}$ 上的概率质量函数。那么

$$D_{\mathrm{CT}}\Big(P_X^{(1)}, Q_U^{(1)} \Big\| P_X^{(2)}, Q_U^{(2)}\Big) \le C\Big(P_X^{(1)} \Big\| P_X^{(2)}\Big) - \log\Big(1 - \mathrm{TV}\big(Q_U^{(1)}, Q_U^{(2)}\big)\Big).$$

证明。 固定 $u \in \mathcal{U}$。对于任意 $\lambda \in [0,1]$,

$$\sum_{x\in\mathcal{X}} \big(P_X^{(1)}(x)Q_U^{(1)}(u)\big)^{\lambda}\big(P_X^{(2)}(x)Q_U^{(2)}(u)\big)^{1-\lambda} = \big(Q_U^{(1)}(u)\big)^{\lambda}\big(Q_U^{(2)}(u)\big)^{1-\lambda}\sum_{x\in\mathcal{X}} P_X^{(1)}(x)^{\lambda}P_X^{(2)}(x)^{1-\lambda}.$$

对两边关于 λ取最小值，得到

$$\begin{aligned}
&\min_{\lambda\in[0,1]}\sum_{x\in\mathcal{X}} \big(P_X^{(1)}(x)Q_U^{(1)}(u)\big)^{\lambda}\big(P_X^{(2)}(x)Q_U^{(2)}(u)\big)^{1-\lambda}\\
&\geq \Big(\min_{\lambda\in[0,1]}\sum_{x\in\mathcal{X}} P_X^{(1)}(x)^{\lambda}P_X^{(2)}(x)^{1-\lambda}\Big)\min_{\lambda\in[0,1]}\big(Q_U^{(1)}(u)\big)^{\lambda}\big(Q_U^{(2)}(u)\big)^{1-\lambda}\\
&= \Big(\min_{\lambda\in[0,1]}\sum_{x\in\mathcal{X}} P_X^{(1)}(x)^{\lambda}P_X^{(2)}(x)^{1-\lambda}\Big)\min\Big\{Q_U^{(1)}(u), Q_U^{(2)}(u)\Big\}.
\end{aligned}$$

对 u求和，得到

$$\sum_{u\in\mathcal{U}}\min_{\lambda_u\in[0,1]}\sum_{x\in\mathcal{X}} \big(P_X^{(1)}(x)Q_U^{(1)}(u)\big)^{\lambda_u}\big(P_X^{(2)}(x)Q_U^{(2)}(u)\big)^{1-\lambda_u} \geq A\sum_{u\in\mathcal{U}}\min\{Q_U^{(1)}(u), Q_U^{(2)}(u)\},$$

其中 $A := \min_{\lambda\in[0,1]}\sum_{x\in\mathcal{X}} P_{(1)\,X}(x)^{\lambda}P_{(2)\,X}(x)^{1-\lambda} = \exp(-C(P_X^{(1)}\|P_{(2)\,X}))$。回想对于离散分布，$\sum_u \min\{Q^{(1)}(u)\ Q^{(2)}(u)\} = 1 - \mathrm{TV}(Q^{(1)}, Q^{(2)})$. 取对数即完成证明。

□

4.2 DBM的精确恢复阈值

在本节中，固定 $k \in \mathbb{N}$ 以及 $[k]$上的先验分布 $P = (p_1, \ldots, p_k)$。令$\mathbf{Q} \in \mathbb{R}_+^{k\times k}$为对称的，并考虑对数度数区间

$$\mathbf{W}^{(n)} = \frac{\log n}{n}\mathbf{Q}.$$

设 $\{\mathcal{U}^{(n)}\}_{n\geq 1}$为一列有限字母表，并设 $\{P_{(n)}{}_{U|X}\}_{n\geq 1}$为一列条件概率质量函数，使得对每个 n和每个 $x \in [k]$，$P_{(n)}{}^{U|X}(\cdot|x)$是 $\mathcal{U}^{(n)}$上的概率质量函数。对每个 $r \in [k]$，定义

$$\boldsymbol{\mu}_r := (\mathrm{diag}(P)\mathbf{Q})_r \in \mathbb{R}_+^k, \qquad \boldsymbol{\mu}_r^{(n)} := \boldsymbol{\mu}_r \log n,$$

其中，$(\mathrm{diag}(P)\mathbf{Q})_r$表示 $\mathrm{diag}(P)\mathbf{Q}$ 的第 r 列。对于 $s, t \in [k]$, $s \neq t$，定义（归一化的）Chernoff–全变差分离度

$$D_{s,t} := \liminf_{n\to\infty}\frac{1}{\log n} D_{\mathrm{CT}}\Big(\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}}, P_{U|X}^{(n)}(\cdot|s)\ \Big\|\ \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}}, P_{U|X}^{(n)}(\cdot|t)\Big), \tag{2}$$

其中，$\mathcal{P}_{\boldsymbol{\mu}_r^{(n)}}$表示定义在 $\mathbb{Z}_+^k$ 上的乘积泊松分布，其坐标相互独立，均值向量为 $\boldsymbol{\mu}_r^{(n)}$。

定理 4.4 （可实现性）. 设 $(G_n, X^n, U^n) \sim \mathrm{DBM}\Big(n, k, P, \mathbf{W}^{(n)}, P_{(n}{}_{U|X}\Big)$. 若对所有 $s, t \in [k]$, $s \neq t$,

$$D_{s,t} > 1, \tag{3}$$

那么可以实现精确恢复，并且存在一种多项式时间（事实上是近线性时间）算法，可以实现精确恢复。

定理 4.5 (逆命题)。 设$(G_n, X^n, U^n) \sim \mathrm{DBM}\Big(n, k, P, \mathbf{W}^{(n)}, P_{\substack{(n)\\U|X}}\Big)$。若存在 $s, t \in [k]$满足 $s \neq t$，使得

$$D_{s,t} < 1, \tag{4}$$

那么精确恢复是不可能的。

定理 4.4 的证明见第4.4小节，定理 4.5 的证明见第 4.5小节。

备注 (化归到经典 SBM 阈值)。 当没有顶点数据可用时（或数据不包含关于标签的信息时），定理 4.4 和 4.5 恢复了标准 SBM 的精确恢复阈值。事实上，假设各社区的属性分布相同，即

$$P^{(n)}_{U|X}(\cdot|s) = P^{(n)}_{U|X}(\cdot|t) \qquad \text{for all } s, t \in [k].$$

那么，由命题 4.2(i),

$$D_{\mathrm{CT}}\Big(\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}}, P^{(n)}_{U|X}(\cdot|s) \,\Big\|\, \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}}, P^{(n)}_{U|X}(\cdot|t)\Big) = C\Big(\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}} \Big\| \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}}\Big).$$

此外，根据 CH 散度的 Chernoff 信息表示（命题 3.5），

$$C\Big(\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}} \Big\| \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}}\Big) = D_+(\boldsymbol{\mu}_s \| \boldsymbol{\mu}_t)\,.$$

因此，定理 4.4–4.5 中的判据可化简为

$$D_+(\boldsymbol{\mu}_s \| \boldsymbol{\mu}_t) > 1 \quad \text{for all } s \neq t,$$

这正是对数度数 SBM 中精确恢复的 Abbe–Sandon 精确阈值。

备注 （与Asadi 等（2017）的充分条件比较）。 （Asadi 等（2017）采用的一种自然方法，是将基于图的统计量和顶点属性视为在每个社区下联合观测到的量，并通过标准 Chernoff 信息比较由此得到的乘积分布 $\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}} \times P_{(n)U|X}(\cdot|s)$ 和 $\mathcal{P}_{\boldsymbol{\mu}_t^{(n)}} \times P_{(n)U|X}(\cdot|t)$。由此可得精确恢复的一个显式充分条件：若这些乘积分布之间的 Chernoff 分离度足够大（在对数度数区间按适当的 $\log n$ 归一化），则可以实现精确恢复。

然而，这个乘积切尔诺夫判据通常并非必要条件。原因在于，乘积分布的 Chernoff 信息是在所有属性取值共用的单个指数参数 $\lambda \in [0, 1]$上进行优化，而 Chernoff–TV 散度 D_{CT} 则是在可能随观测到的属性取值 u变化的一组 $\{\lambda_u\}_{u \in \mathcal{U}}$ 上进行优化。因此， D_{CT} 可能严格大于相应的乘积切尔诺夫指数；即使乘积切尔诺夫充分条件不成立，仍然可能实现精确恢复。

一个自然的问题是，当图低于精确恢复阈值时，节点数据何时能够弥补这一不足。下一个定理表明，如果基于图的分离度 $D_+(\boldsymbol{\mu}_s \| \boldsymbol{\mu}_t)$ 小于 1，且数据分布的全变差距离不够大，那么精确恢复仍然不可能。

推论 4.6 (数据何时无法提供帮助)。 设 $(G_n, X^n, U^n) \sim \mathrm{DBM}\Big(n, k, P, \mathbf{W}^{(n)}, P^{(n}_{U|X}\Big)$。假设存在 $s, t \in [k]$, $s \neq t$，使得

$$D_+(\boldsymbol{\mu}_s \,\|\, \boldsymbol{\mu}_t) < 1 - \varepsilon \qquad \textit{and} \qquad \mathrm{TV}\Big(P^{(n)}_{U|X}(\cdot|s),\, P^{(n)}_{U|X}(\cdot|t)\Big) \le 1 - n^{-\varepsilon}$$

对于某个 $\varepsilon > 0$，以及所有充分大的 n，精确恢复是不可能的 .

证明。设 $s \neq t$ 满足假设。在对数度数区间内， $\boldsymbol{\mu}_s^{(n)} = (\log n)\boldsymbol{\mu}_s$且 $\boldsymbol{\mu}_t^{(n)} = (\log n)\boldsymbol{\mu}_t$，根据多元泊松分布的 CH 散度与 Chernoff 信息之间的恒等关系（命题3.5 以及缩放关系 $D_+(c\boldsymbol{a}\|c\boldsymbol{b}) = c\, D_+(\boldsymbol{a}\|\boldsymbol{b})$）可得

$$C\Big(\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}} \Big\| \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}}\Big) = D_+\Big(\boldsymbol{\mu}_s^{(n)} \Big\| \boldsymbol{\mu}_t^{(n)}\Big) = (\log n)\, D_+(\boldsymbol{\mu}_s \| \boldsymbol{\mu}_t) < (\log n)(1-\varepsilon).$$

应用引理4.3，取

$$P_X^{(1)} = \mathcal{P}_{\boldsymbol{\mu}_s^{(n)}}, \quad P_X^{(2)} = \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}}, \quad Q_U^{(1)} = P^{(n)}_{U|X}(\cdot|s), \quad Q_U^{(2)} = P^{(n)}_{U|X}(\cdot|t),$$

可得，对每个 n,

$$\begin{aligned}
&D_{\mathrm{CT}}\Big(\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}},\, P^{(n)}_{U|X}(\cdot|s) \,\Big\|\, \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}},\, P^{(n)}_{U|X}(\cdot|t)\Big) \\
&\quad \le C\Big(\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}} \Big\| \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}}\Big) - \log\Big(1 - \mathrm{TV}\big(P^{(n)}_{U|X}(\cdot|s), P^{(n)}_{U|X}(\cdot|t)\big)\Big) \\
&\quad = (\log n)\, D_+(\boldsymbol{\mu}_s\|\boldsymbol{\mu}_t) - \log\Big(1 - \mathrm{TV}\big(P^{(n)}_{U|X}(\cdot|s), P^{(n)}_{U|X}(\cdot|t)\big)\Big) \\
&\quad < (\log n)(1-\varepsilon) + \varepsilon \log n \\
&\quad = \log n.
\end{aligned}$$

因此,

$$D_{s,t} := \liminf_{n\to\infty} \frac{1}{\log n}\, D_{\mathrm{CT}}\Big(\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}},\, P^{(n)}_{U|X}(\cdot|s) \,\Big\|\, \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}},\, P^{(n)}_{U|X}(\cdot|t)\Big) < 1.$$

因此，根据定理4.5，精确恢复是不可能的。 □

一种研究广泛的顶点侧信息形式是噪声擦除标签模型，其中节点属性取值于社区字母表，但会受到干扰。具体而言，给定$X_i = x$，观测到的标签 $U^{(i)}$ 以 x 的概率趋近于 1，否则该标签会被擦除或替换为另一个社区标签。二元对称情形（$k = 2$）由Saad 和 Nosratinia于2018年进行了分析。作为定理4.4 和4.5的应用，我们推导出在多项式衰减噪声模型下，一般 $k \ge 2$ 精确恢复的尖锐充要条件。

推论 4.7 （噪声擦除标签）。考虑数据块模型

$$(G_n, X^n, U^n) \sim \mathrm{DBM}\left(n, k, P, \frac{\log n}{n}\mathbf{Q}, P^{(n)}_{U|X}\right),$$

其中 $P = (p_1, \ldots, p_k)$ 是 $[k]$ 上的先验分布，且 $\mathbf{Q} \in \mathbb{R}_+^{k\times k}$是对称的。令 $\boldsymbol{\mu}_s := (\mathrm{diag}(P)\mathbf{Q})_s \in \mathbb{R}_+^k$，并将其 ℓ个分量记为 μ_{s_ℓ} 。

假设标签字母表为 $\mathcal{U} = \{1, \dots, k, \varepsilon\}$，其中 ε 表示擦除，并且对于每个 $x \in [k]$，数据–社区信道满足

$$P_{U|X}^{(n)}(u|x) = \begin{cases} n^{-d_{u,x}}, & u \in [k],\ u \neq x, \\ n^{-d_{\varepsilon,x}}, & u = \varepsilon, \\ 1 - \sum_{\substack{u \in [k] \\ u \neq x}} n^{-d_{u,x}} - n^{-d_{\varepsilon,x}}, & u = x, \end{cases}$$

对某些常数 $d_{u,\,x} > 0$ （对于 $u \neq x$）以及 $d_{\varepsilon,\,x} > 0$。

那么，对于任意 $s \neq t$，定理 4.4 和4.5中出现的成对散度率$D_{s,\,t}$ 与以下显式变分表达式一致

$$D_{s,t} = \min_{u \in \mathcal{U}} \max_{\lambda_u \in [0,1]} \left[d_{u,s}\lambda_u + d_{u,t}(1-\lambda_u) + \sum_{\ell=1}^{k} \left(\mu_{s_\ell}\lambda_u + \mu_{t_\ell}(1-\lambda_u) - \mu_{s_\ell}^{\lambda_u}\mu_{t_\ell}^{1-\lambda_u} \right) \right],$$

约定 $d_{x,\,x} = 0$ 对于 $x \in [k]$。

因此，对于所有 $s \neq t$，若 $D_{s,\,t} > 1$，则精确恢复可解；若存在 $s \neq t$ 使得 $D_{s,\,t} < 1$，则不可解。

证明。 为求相变阈值$D_{s,\,t}$，我们求 Chernoff–TV 散度

$$D_{\mathrm{CT}}\left(\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}}, P_{U|X}^{(n)}(\cdot|s) \middle\| \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}}, P_{U|X}^{(n)}(\cdot|t) \right) =$$
$$- \log \left(\sum_{u \in \mathcal{U}} \exp \left(\log n \left(\max_{\lambda_u \in [0,1]} d_{u,s}\lambda_u + d_{u,t}(1-\lambda_u) + \sum_{\ell=1}^{k} \mu_{s_\ell}\lambda_u + \mu_{t_\ell}(1-\lambda_u) - \mu_{s_\ell}^{\lambda_u}\mu_{t_\ell}^{1-\lambda_u} \right) \right) \right).$$

定义

$$\Delta_{s,t} = \min_{u \in \mathcal{U}} \max_{\lambda_u \in [0,1]} \left[d_{u,s}\lambda_u + d_{u,t}(1-\lambda_u) + \sum_{\ell=1}^{k} \mu_{s_\ell}\lambda_u + \mu_{t_\ell}(1-\lambda_u) - \mu_{s_\ell}^{\lambda_u}\mu_{t_\ell}^{1-\lambda_u} \right].$$

那么，

$$\begin{aligned} \exp(-\Delta_{s,t} \log n) &\leq \exp\left(-D_{\mathrm{CT}}\left(\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}}, P_{U|X}^{(n)}(\cdot|s) \middle\| \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}}, P_{U|X}^{(n)}(\cdot|t) \right) \right) \\ &\leq (k+1)\exp(-\Delta_{s,t} \log n). \end{aligned}$$

因此，噪声标签问题的相变阈值为 b y

$$D_{s,t} = \liminf_{n \to \infty} -\frac{1}{\log n} \log(\exp(-\Delta_{s,t} \log n)) = \Delta_{s,t}.$$

□

例 4.8（擦除标签）。 考虑先验分布为$P = (1/2, 1/2)$、边概率为对数阶度数边概率的平衡双社区对称模型，其中

$$\mathbf{Q} = \begin{pmatrix} a & b \\ b & a \end{pmatrix}.$$

侧信息字母表为 $\mathcal{U}=\{1,2,\varepsilon\}$，其中 ε 表示擦除。在给定 $X\in\{1,2\}$的条件下，观测到的标签以概率 $1-n^{-\alpha}$ 正确，以概率 $n^{-\alpha}$被擦除：

$$P_{U|X}^{(n)}(x|x)=1-n^{-\alpha},\qquad P_{U|X}^{(n)}(\varepsilon|x)=n^{-\alpha},\qquad P_{U|X}^{(n)}(\bar{x}|x)=0,\qquad x\in\{1,2\}.$$

应用推论 4.7 可得

$$D_{1,2}=\max_{\lambda\in[0,1]}\Bigg[\alpha\lambda+\alpha(1-\lambda)+\frac{a}{2}\lambda+\frac{b}{2}(1-\lambda)-\left(\frac{a}{2}\right)^{\lambda}\left(\frac{b}{2}\right)^{1-\lambda}+\frac{b}{2}\lambda+\frac{a}{2}(1-\lambda)-\left(\frac{b}{2}\right)^{\lambda}\left(\frac{a}{2}\right)^{1-\lambda}\Bigg].$$

取 $\lambda=\frac{1}{2}$ 可得_

$$D_{1,2}=\alpha+\frac{a+b}{2}-\sqrt{ab}=\alpha+\frac{(\sqrt{a}-\sqrt{b})^2}{2}.$$

Henc，精确恢复相变与（Saad 和 Nosratinia，2018 ）.

例 4.9 （噪声标签）。考虑先验分布为$P=(1/2,\ 1/2)$且边概率为对数度形式的平衡双社区对称模型，其中

$$\mathbf{Q}=\begin{pmatrix}a & b\\ b & a\end{pmatrix}.$$

标签字母表为 $\mathcal{U}=\{1,2\}$。在给定真实社区标签 $X\in\{1,2\}$ 的条件下，观测标签以 $n^{-\alpha}$ 的概率发生翻转：

$$P_{U|X}^{(n)}(x|x)=1-n^{-\alpha},\qquad P_{U|X}^{(n)}(\bar{x}|x)=n^{-\alpha},\qquad x\in\{1,2\}.$$

应用推论 4.7 得到

$$\begin{aligned}D_{1,2}&=\max_{\lambda\in[0,1]}\left[\alpha\lambda+\frac{a}{2}\lambda+\frac{b}{2}(1-\lambda)-\left(\frac{a}{2}\right)^{\lambda}\left(\frac{b}{2}\right)^{1-\lambda}+\frac{b}{2}\lambda+\frac{a}{2}(1-\lambda)-\left(\frac{b}{2}\right)^{\lambda}\left(\frac{a}{2}\right)^{1-\lambda}\right]\\&=\frac{a+b}{2}+\max_{\lambda\in[0,1]}\left[\alpha\lambda-\left(\frac{a}{2}\right)^{\lambda}\left(\frac{b}{2}\right)^{1-\lambda}-\left(\frac{b}{2}\right)^{\lambda}\left(\frac{a}{2}\right)^{1-\lambda}\right].\end{aligned}$$

令 $T:=\log(a/b)$ 且 $\eta:=\sqrt{\alpha^2+abT^2}$。最大化点为

$$\lambda^{\star}=\frac{1}{2}+\frac{1}{2T}\log\left(\frac{\eta+\alpha}{\eta-\alpha}\right)\qquad\text{when}\quad\alpha<\frac{T(a-b)}{2},$$

an否则，最优值在边界 $\lambda^{\star}=1$ 处取得。代入后得到 s

$$D_{1,2}=\begin{cases}\dfrac{\alpha}{2}+\dfrac{a+b}{2}-\dfrac{\eta}{T}+\dfrac{\alpha}{2T}\log\left(\dfrac{\eta+\alpha}{\eta-\alpha}\right), & \alpha<\frac{T(a-b)}{2},\\ \alpha, & \alpha>\frac{T(a-b)}{2}.\end{cases}$$

因此，精确恢复阈值与（Saad 和 Nosratinia，2018）一致。

备注：在（Dreveton 等人，2023）的示例 3 中，作者研究了对称双社区 SBM 的一种半监督变体。用我们的记号表示，这对应于一个满足 $k=2$、不存在擦除，且噪声标签信道满足以下条件的 DBM：

$$\mathbb{P}\big(U_i \neq X_i \mid X_i\big) = n^{-\alpha}.$$

根据他们的分析，Dreveton 等人(2023)指出，只有在以下条件成立时才能实现精确恢复：

$$\lim_{n\to\infty} (\sqrt{a}-\sqrt{b})^2 - \frac{2}{\log n}\log\Big(2\sqrt{n^{-\alpha}(1-n^{-\alpha})}\Big) > 2,$$

这可简化为

$$\frac{\alpha}{2} + \frac{a+b}{2} - \sqrt{ab} > 1.$$

相比之下，例4.9中得到的尖锐阈值与（Saad 和 Nosratinia，2018）中的刻画一致，却与这一条件不符。

4.3 精确恢复的两阶段算法

本节介绍一种两阶段过程，该过程能够以高概率并在多项式时间内实现 DBM 中的精确恢复。其核心思路是将观测图拆分为两个独立子图。随后，我们利用其中一个子图快速进行近似聚类，并在第二个子图上对每个节点分别执行最大后验（MAP）估计，以细化聚类结果；这种方法能够自然地纳入节点度分布和辅助侧信息。具体而言，假设 σ' 是算法第一阶段近似聚类的结果，我们将每个节点的度分布设为

$$\mathbf{d}(v) = \{d_1(v), d_2(v), \ldots, d_k(v)\}$$

其中，$d_r(v)$是第二个子图中节点 v 与被 σ' 分配到社区 r的节点之间的边数。我们用 $\mu_{sr}^{(n)}$表示平均度分布向量$\boldsymbol{\mu}_s^{(n)} = (\mathrm{diag}(P)\mathbf{Q})_s \log n$的第 r个分量，即，

$$\mu_{sr}^{(n)} = (\boldsymbol{\mu}_s^{(n)})_r = p_r Q_{rs} \log n, \qquad s, r \in [k].$$

要成功执行算法，必须利用观测图结构及相关数据标签，消除模型中固有的对称性。下面的引理给出了一个充分条件：在可获取 $\Theta(n)$ 个数据样本（其中 m可能受到污染）的情况下，任意两个社区之间的对称性都能以高概率被打破。

引理 4.10 (Scheff´e 检验（存在 m污染）。设 $P_X^{(1)}$和 $P_{(2)\,X}$是定义在离散字母表 $\mathcal{X}$上的概率质量函数。设 $X_1, \ldots, X_n$为观测值，且存在一个（未知）指标集 $\mathcal{I} \subseteq [n]$，其中 $|\mathcal{I}| \geq n - m$，使得随机变量 $\{X_i : i \in \mathcal{I}\}$ 独立同分布，且服从共同分布 P_X，其中

$$P_X \in \{P_X^{(1)}, P_X^{(2)}\},$$

其余 m 个观测值 $\{X_i : i \notin \mathcal{I}\}$ 则是任意的。

考虑检验

$$H_1 : \; P_X = P_X^{(1)} \qquad vs. \qquad H_2 : \; P_X = P_X^{(2)}.$$

设经验测度为

$$P_n(A) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \in A\}, \qquad A \subseteq \mathcal{X},$$

并定义

$$A^* = \arg\max_{A\subseteq\mathcal{X}} \left|P_X^{(1)}(A) - P_X^{(2)}(A)\right|.$$

考虑 *Scheff´e* 判决规则

$$\left|P_n(A^*) - P_X^{(1)}(A^*)\right| \underset{H_1}{\overset{H_2}{\gtrless}} \left|P_n(A^*) - P_X^{(2)}(A^*)\right|.$$

If

$$\mathrm{TV}\left(P_X^{(1)}, P_X^{(2)}\right) > \frac{2m}{n},$$

则该检验的（先验分布平均）贝叶斯错误概率 P_e 满足

$$P_e \le 4\exp\left(-\frac{n}{2}\left(\mathrm{TV}\left(P_X^{(1)}, P_X^{(2)}\right) - \frac{2m}{n}\right)^2\right).$$

证明。令 $A^* \in \arg\max_{A\subseteq\mathcal{X}}\left|P_X^{(1)}(A) - P_{(2)X}(A)\right|$，于是

$$\left|P_X^{(1)}(A^*) - P_X^{(2)}(A^*)\right| = \mathrm{TV}\left(P_X^{(1)}, P_X^{(2)}\right).$$

定义 $S_n := \sum_{i=1}^{n} \mathbb{1}\{X_i \in A^*\}$，使得 $P_n(A^*) = S_n/n$。

我们首先界定 H_1 下的第一类错误（即，$P_X = P_X^{(1)}$ 在未污染样本上）。仅当经验质量 $P_n(A^*)$ 比起 $P_X^{(1)}(A^*)$ 更接近 $P_{(2)X}(A^*)$ 时，才会发生错误，即

$$\left|P_n(A^*) - P_X^{(1)}(A^*)\right| > \left|P_n(A^*) - P_X^{(2)}(A^*)\right|.$$

由三角不等式，

$$\left|P_X^{(2)}(A^*) - P_X^{(1)}(A^*)\right| \le \left|P_n(A^*) - P_X^{(1)}(A^*)\right| + \left|P_n(A^*) - P_X^{(2)}(A^*)\right|,$$

因此在误差事件上，我们必有

$$2\left|P_n(A^*) - P_X^{(1)}(A^*)\right| > \left|P_X^{(2)}(A^*) - P_X^{(1)}(A^*)\right| = \mathrm{TV}\left(P_X^{(1)}, P_X^{(2)}\right).$$

因此

$$P_e(1) \le \Pr_{H_1}\left\{2\left|P_n(A^*) - P_X^{(1)}(A^*)\right| > \mathrm{TV}\left(P_X^{(1)}, P_X^{(2)}\right)\right\}. \tag{5}$$

设 $\mathcal{I} \subseteq [n]$ 为未污染样本的（未知）指标集，且 $|\mathcal{I}| \ge n - m$。记

$$S_n = \sum_{i\in\mathcal{I}} \mathbb{1}\{X_i \in A^*\} + \sum_{i\notin\mathcal{I}} \mathbb{1}\{X_i \in A^*\} =: S_{\mathcal{I}} + S_{\mathcal{I}^c}.$$

在 H_1 条件下，变量 $\{\mathbb{1}\{X_i \in A^*\} : i \in \mathcal{I}\}$ 服从均值为 $P_X^{(1)}(A^*)$ 的独立同分布伯努利分布，而$S_{\mathcal{I}^c} \in [0, m]$ 是任意的。因此，

$$\left|S_n - (n-m)P_X^{(1)}(A^*)\right| \le \left|S_{\mathcal{I}} - (n-m)P_X^{(1)}(A^*)\right| + m,$$

并将其除以 n，得到

$$\left|P_n(A^*) - P_X^{(1)}(A^*)\right| = \frac{1}{n}\left|S_n - nP_X^{(1)}(A^*)\right| \le \frac{1}{n}\left|S_{\mathcal{I}} - (n-m)P_X^{(1)}(A^*)\right| + \frac{m}{n}.$$

将其代入（5）可得

$$P_e(1) \le \Pr_{H_1}\left\{\frac{2}{n}\left|S_{\mathcal{I}} - (n-m)P_X^{(1)}(A^*)\right| > \mathrm{TV}\left(P_X^{(1)}, P_X^{(2)}\right) - \frac{2m}{n}\right\}.$$

根据 $(n-m)$ 个独立同分布伯努利变量之和的 Hoeffding不等式,

$$\begin{aligned}\Pr_{H_1}\left\{\left|S_{\mathcal{I}} - (n-m)P_X^{(1)}(A^*)\right| > \frac{n}{2}\left(\mathrm{TV}(P_X^{(1)}, P_X^{(2)}) - \frac{2m}{n}\right)\right\} \\ \le 2\exp\left(-\frac{1}{2}n\left(\mathrm{TV}(P_X^{(1)}, P_X^{(2)}) - \frac{2m}{n}\right)^2\right),\end{aligned}$$

其中我们利用 $n-m \le n$ 来简化指数项。因此,

$$P_e(1) \le 2\exp\left(-\frac{1}{2}n\left(\mathrm{TV}(P_X^{(1)}, P_X^{(2)}) - \frac{2m}{n}\right)^2\right).$$

在 H_2 下用相同的论证可得

$$P_e(2) \le 2\exp\left(-\frac{1}{2}n\left(\mathrm{TV}(P_X^{(1)}, P_X^{(2)}) - \frac{2m}{n}\right)^2\right).$$

最后，对于 $\{H_1, H_2\}$ 上任意先验分布，贝叶斯误差满足 $P_e \le P_e(1) + P_e(2)$，因此

$$P_e \le 4\exp\left(-\frac{1}{2}n\left(\mathrm{TV}(P_X^{(1)}, P_X^{(2)}) - \frac{2m}{n}\right)^2\right).$$

□

所提出的两阶段恢复算法如下:

Exact Recovery Algorithm for Data Block Model (DBM)

输入:

- 一个图 $g([n],\ \mathcal{E})$，其顶点集为 $[n]$，边集为 $\mathcal{E}$,
- 所有 $v \in [n]$ 的数据标签 $L(v) \in \mathcal{U}$,
- DBM参数:

  – 一个 P 上的离散分布，– 一个边概率矩阵 $\mathbf{Q} \in \mathbb{R}_+^{k\times k}$，–针对 $u \in \mathcal{U}$ 的条件数据标签分布 $P_{U|X}(u|s),\ s \in [k]$，–一个拆分参数 $\gamma \in (0,\ 1)$ 以及用于 Sphere-comparison算法的误差参数 $\delta \in (0,\ 1)$。

输出: 一个社区分配 $\sigma'' \in [k]^n$，将每个节点 $v \in [n]$ 划分到一个社区$c \in [k]$。

1. **图拆分：**
   以概率 γ，将 g 中的每条边独立地加入新的边集 $\mathcal{E}'$。令 $\mathcal{E}'' = \mathcal{E} \setminus \mathcal{E}'$。定义子图 $g'([n],\ \mathcal{E}')$ 和 $g''([n],\ \mathcal{E}'')$。

2. **初始社区估计：**
   应用 SPHERE-COMPARISON（Abbe 和 Sandon，2015a）中的算法处理 g'，得到近似社区分配 $\sigma' \in [k]^n$。

3. **对称性识别：**
   定义（近似）不可区分社区对的集合：

$$\mathcal{S} = \left\{ (s,t) \in [k] \times [k] \;\middle|\; \begin{array}{l} p_s = p_t, \\ \boldsymbol{\mu}_s = \boldsymbol{\mu}_t, \\ \mathrm{TV}\left(P^{(n)}_{U|X}(\cdot|s), P^{(n)}_{U|X}(\cdot|t) \right) \le \dfrac{2\delta}{\log n} \end{array} \right\}.$$

   中的社区对 $\mathcal{S}$ 对应于具有相同连接模式的社区，其属性分布过于接近（相对于步骤2的污染水平），因此无法在置换消解步骤中可靠地区分。因此，我们的目标仅是在由 $\mathcal{S}$诱导的等价类内，恢复至标签置换意义下的标签。对于（$s,\ t$）$\notin \mathcal{S}$，引理4.10表明，s与 t之间的对称性可以以高概率打破。

4. **通过最大后验估计进行精细分类：**
   对于每个节点 $v \in [n]$，根据初步分类 σ' 计算其在 g'' 中的度分布 $\mathbf{d}(v)$。通过一系列成对最大后验假设检验 $k-1$，确定最终分配 $\sigma''(v)$：

   - 如果 $(s,t) \in \mathcal{S}$，则使用：

$$\Pr\{\mathbf{d}(v) \mid H = s\} \cdot p_s > \Pr\{\mathbf{d}(v) \mid H = t\} \cdot p_t \quad \Rightarrow \quad H \neq t$$

   - 如果 $(s,t) \notin \mathcal{S}$，则使用：

$$\Pr\{\mathbf{d}(v), L(v) \mid H = s\} \cdot p_s > \Pr\{\mathbf{d}(v), L(v) \mid H = t\} \cdot p_t \quad \Rightarrow \quad H \neq t$$

5. **返回：** 输出最终的社区分配 $\sigma'' \in [k]^n$。

在小节 4.4 中，我们首先证明，在定理 4.4 中，若 $D_{s,t} > 1$ 对所有 $s,t \in [k], s \neq t$ 均成立，则 EXACT RECOVERY ALGORITHM FOR DBM 能够以趋于零的错误概率找到正确的社区（模去对称性），其中 n 趋于无穷。随后，在小节 4.5 中，我们证明，若对于任意 $s,t \in [k], s \neq t$，都有 $D_{s,t} < 1$，那么任何试图根据 (G_n, U^n) 找出节点所属社区的算法，其错误概率都不会趋于零。

4.4 充分性证明

定理的证明 *4.4*。 存在 $N_0 \in \mathbb{N}$和 $\epsilon > 0$，使得对于任意 $n \ge N_0$和 $s,\ t \in [k],\ s \neq t$有

$$D_{\mathrm{CT}}\left(\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}}, P^{(n)}_{U|X}(\cdot|s) \middle\| \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}}, P^{(n)}_{U|X}(\cdot|t) \right) \ge (1+\epsilon)\log n. \tag{6}$$

对于定理4.4中的分布的 CT-散度，有

$$\begin{aligned}
&\exp\left(-D_{\mathrm{CT}}\left(\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}}, P_{U|X}^{(n)}(\cdot|s)\middle\|\mathcal{P}_{\boldsymbol{\mu}_t^{(n)}}, P_{U|X}^{(n)}(\cdot|t)\right)\right)\\
&\quad=\sum_{u\in\mathcal{U}}\min_{\lambda_u\in[0,1]}\Bigg\{(P_{U|X}^{(n)}(u|s))^{\lambda_u}(P_{U|X}^{(n)}(u|t))^{1-\lambda_u}\\
&\qquad\qquad\times\exp\left(\sum_{r=1}^{k}-\lambda_u\mu_{s_r}^{(n)}-(1-\lambda_u)\mu_{t_r}^{(n)}+(\mu_{s_r}^{(n)})^{\lambda_u}(\mu_{t_r}^{(n)})^{1-\lambda_u}\right)\Bigg\},\qquad(7)\\
&\quad=\sum_{u\in\mathcal{U}}(P_{U|X}^{(n)}(u|s))^{\lambda_u^*}(P_{U|X}^{(n)}(u|t))^{1-\lambda_u^*}\\
&\qquad\qquad\times\exp\left(\sum_{r=1}^{k}-\lambda_u^*\mu_{s_r}^{(n)}-(1-\lambda_u^*)\mu_{t_r}^{(n)}+(\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}\right),\qquad(8)
\end{aligned}$$

其中 λ_u^* 是使（7）式中的表达式对每个 $u\in\mathcal{U}$取最小值的最优值。 λ_u^*的值满足

$$\begin{aligned}
&\log(P_{U|X}^{(n)}(u|s))-\log(P_{U|X}^{(n)}(u|t))\\
&\quad+\sum_{r=1}^{k}\left(-\mu_{s_r}^{(n)}+\mu_{t_r}^{(n)}+(\log(\mu_{s_r}^{(n)})-\log(\mu_{t_r}^{(n)}))(\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}\right)=0,\qquad(9)
\end{aligned}$$

由此可得

$$\begin{aligned}
&P_{U|X}^{(n)}(u|s)\prod_{r=1}^{k}\exp(-\mu_{s_r}^{(n)})(\mu_{s_r}^{(n)})^{(\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}}\\
&\qquad=P_{U|X}^{(n)}(u|t)\prod_{r=1}^{k}\exp(-\mu_{t_r}^{(n)})(\mu_{t_r}^{(n)})^{(\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}}.\qquad(10)
\end{aligned}$$

由（6）和（8），我们推导出

$$\begin{aligned}
&\sum_{u\in\mathcal{U}}(P_{U|X}^{(n)}(u|s))^{\lambda_u^*}(P_{U|X}^{(n)}(u|t))^{1-\lambda_u^*}\\
&\qquad\times\exp\left(\sum_{r=1}^{k}-\lambda_u^*\mu_{s_r}^{(n)}-(1-\lambda_u^*)\mu_{t_r}^{(n)}+(\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}\right)\le n^{-1-\epsilon}.\qquad(11)
\end{aligned}$$

我们将给出该假设检验问题中错误概率的上界。

$$
\begin{aligned}
&\sum_{u\in\mathcal{U}}\sum_{\mathbf{x}\in\mathbb{Z}_+^k}\min\left\{p_s P_{U|X}^{(n)}(u|s)\prod_{r=1}^{k}\exp(-\mu_{s_r}^{(n)})\frac{(\mu_{s_r}^{(n)})^{x_r}}{x_r!}, p_t P_{U|X}^{(n)}(u|t)\prod_{r=1}^{k}\exp(-\mu_{t_r}^{(n)})\frac{(\mu_{t_r}^{(n)})^{x_r}}{x_r!}\right\}\\
&\overset{(a)}{\le}\max\{p_s,p_t\}\sum_{u\in\mathcal{U}}\sum_{\mathbf{x}\in\mathbb{Z}_+^k}(P_{U|X}^{(n)}(u|s))^{\lambda_u^*}(P_{U|X}^{(n)}(u|t))^{1-\lambda_u^*}\\
&\qquad\times\prod_{r=1}^{k}\exp(-\lambda_u^*\mu_{s_r}^{(n)}-(1-\lambda_u^*)\mu_{t_r}^{(n)})\frac{(\mu_{s_r}^{(n)})^{\lambda_u^* x_r}(\mu_{t_r}^{(n)})^{(1-\lambda_u^*)x_r}}{x_r!}\\
&=\max\{p_s,p_t\}\sum_{u\in\mathcal{U}}(P_{U|X}^{(n)}(u|s))^{\lambda_u^*}(P_{U|X}^{(n)}(u|t))^{1-\lambda_u^*}\\
&\qquad\times\exp\left(\sum_{r=1}^{k}-\lambda_u^*\mu_{s_r}^{(n)}-(1-\lambda_u^*)\mu_{t_r}^{(n)}+(\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}\right)\\
&\overset{(b)}{\le}\max\{p_s,p_t\}n^{-1-\epsilon},
\end{aligned}
$$

其中（a）由 $\min\{x,y\}\le x^\lambda y^{1-\lambda},\ \lambda\in[0,1]$ 而（b）由（11）可得。

如果精确恢复算法能够获取每个节点的精确度分布，则每个节点的误分类误差可由下式给出上界：

$$
\begin{aligned}
P_e(v)&\le\frac{C}{\sqrt{(1-\gamma)\log n}}\sum_{u\in\mathcal{U}}(P_{U|X}^{(n)}(u|s))^{\lambda_u^*}(P_{U|X}^{(n)}(u|t))^{1-\lambda_u^*}\\
&\qquad\times\exp\left((1-\gamma)\sum_{r=1}^{k}-\lambda_u^*\mu_{s_r}^{(n)}-(1-\lambda_u^*)\mu_{t_r}^{(n)}+(\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}\right)\\
&\le\frac{C\exp(D_+(\boldsymbol{\mu}_s\|\boldsymbol{\mu}_t)\gamma\log n)}{\sqrt{(1-\gamma)\log n}}\sum_{u\in\mathcal{U}}(P_{U|X}^{(n)}(u|s))^{\lambda_u^*}(P_{U|X}^{(n)}(u|t))^{1-\lambda_u^*}\\
&\qquad\times\exp\left(\sum_{r=1}^{k}-\lambda_u^*\mu_{s_r}^{(n)}-(1-\lambda_u^*)\mu_{t_r}^{(n)}+(\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}\right)\\
&\le Cn^{-1-\epsilon}\exp(D_+(\boldsymbol{\mu}_s\|\boldsymbol{\mu}_t)\gamma\log n).
\end{aligned}
$$

然而， Sphere-Comparison 第 2 步中的 Exact Recovery Algorithmfor DBM 算法的错误概率为 $\delta/\log n$。因此，度分布会发生扭曲。误分类标签的节点总数为 $\delta n(1+o(1))/\log n$。根据报告标签得到的给定节点的度分布概率与根据实际标签得到的度分布概率之比，可以界定为

$$
\begin{aligned}
&\frac{\prod_{r=1}^{k}e^{-\mu_{s_r}^{(n)}}(\mu_{s_r}^{(n)})^{d_r}/d_r!}{\prod_{r=1}^{k}e^{-\mu_{s_r}^{(n)}}(\mu_{s_r}^{(n)})^{d_r+\delta_r}/(d_r+\delta_r)!}\le\prod_{r=1}^{k}(\mu_{s_r}^{(n)})^{-\delta_r}\frac{(d_r+\delta_r)!}{d_r!}\\
&\overset{(a)}{\le}\exp\Big(\sum_{r=1}^{k}-\delta_r\log(\mu_{s_r}^{(n)})+\frac{1}{2}\log(1+\frac{\delta_r}{d_r})-d_r\log(d_r)-\delta_r+\frac{1}{12}+(d_r+\delta_r)\log(d_r+\delta_r)\Big)\\
&\overset{(b)}{\le}C_1\exp(C_2\delta\log\log n).
\end{aligned}
$$

这里， δ_r 是算法 r 在社区 SPHERE COMPARISON中误分类标签的数量，(a)由 Robbins (1955) 的界推出

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12(n+1)}} \le n! \le \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}$$

并且 (b) 可由以下事实推出： $|\delta_r| \le \max_{i,\ j} Q_{i,\ j}\delta$ 且 $c_1 \log n \le d_r \le c_3 \log n$ 对于足够大的c_3 以概率 $1 - 1/n^2$成立。因此，带有误分类标签的贝叶斯假设检验误差至多为使用实际标签时的贝叶斯假设检验误差乘以因子$C_1 \exp(C_2\delta \log\log\ n)$。结合真实度分布的界可得

$$\begin{aligned}
P_e(v) \le & \sum_{u,\mathbf{d}:P^{(n)}_{U|X}(u|s)\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}}(\mathbf{d}+\boldsymbol{\delta})<P^{(n)}_{U|X}(u|t)\mathcal{P}_{\boldsymbol{\mu}_t^{(n)}}(\mathbf{d}+\boldsymbol{\delta})} P^{(n)}_{U|X}(u|s)\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}}(\mathbf{d}) \\
& + \sum_{u,\mathbf{d}:P^{(n)}_{U|X}(u|s)\mathcal{P}_{\mu_t^{(n)}}(\mathbf{d}+\boldsymbol{\delta})>P^{(n)}_{U|X}(u|t)\mathcal{P}_{\mu_t^{(n)}}(\mathbf{d}+\boldsymbol{\delta})} P^{(n)}_{U|X}(u|t)\mathcal{P}_{\mu_t^{(n)}}(\mathbf{d}) \\
& \le C_1 \exp(C_2\delta \log\log n) \sum_{u,\mathbf{d}:P^{(n)}_{U|X}(u|s)\mathcal{P}_{\mu_s^{(n)}}(\mathbf{d}+\boldsymbol{\delta})<P^{(n)}_{U|X}(u|t)\mathcal{P}_{\mu_t^{(n)}}(\mathbf{d}+\boldsymbol{\delta})} P^{(n)}_{U|X}(u|s)\mathcal{P}_{\mu_s^{(n)}}(\mathbf{d}+\boldsymbol{\delta}) \\
& + \sum_{u,\mathbf{d}:P^{(n)}_{U|X}(u|s)\mathcal{P}_{\mu_t^{(n)}}(\mathbf{d}+\boldsymbol{\delta})>P^{(n)}_{U|X}(u|t)\mathcal{P}_{\mu_t^{(n)}}(\mathbf{d}+\boldsymbol{\delta})} P^{(n)}_{U|X}(u|t)\mathcal{P}_{\mu_t^{(n)}}(\mathbf{d}+\boldsymbol{\delta}) \\
& = C_1 \exp(C_2\delta \log\log n) \sum_{u,\mathbf{d}} \min\left\{P^{(n)}_{U|X}(u|s)\mathcal{P}_{\mu_s^{(n)}}(\mathbf{d}), P^{(n)}_{U|X}(u|t)\mathcal{P}_{\mu_t^{(n)}}(\mathbf{d})\right\} \\
& \le C_1 \exp(C_2\delta \log\log n) C n^{-1-\epsilon} \exp(D_+(\boldsymbol{\mu}_s\|\boldsymbol{\mu}_t)\gamma \log n) \\
& \le C_1 C n^{-1-\epsilon+c(\delta,\gamma)}
\end{aligned}$$

其中，向量 $\boldsymbol{\delta}$ 表示由于节点误分类而导致真实度分布发生的变化，误分类是由算法 SPHERE-COMPARISON 造成的，并且

$$c(\delta,\gamma) = C_2\delta\frac{\log\log n}{\log n} + \gamma \max_{s,t} D_+(\boldsymbol{\mu}_s\|\boldsymbol{\mu}_t).$$

如果我们选择e δ 和 γ， 使得 $c(\delta,\ \gamma) < \epsilon$，那么 $P_e(v) = o(1/n)$。证明至此完成 . □

4.5 反向结果

首先，我们证明，如果某一对社区的数据标签与度分布的乘积分布之间的全变差小于某个阈值，则精确恢复不可解。

引理 4.11。令 P 为 $[k]$上的先验分布，$\mathbf{Q} \in \mathbb{R}_+^{k\times k}$ 为对称矩阵，且$\{P_{(n)_{U|X}}\}_{n=1}^{\infty}$ 为集合 $\{\mathcal{U}^{(n)}\}_{n=1}^{\infty}$上的离散测度。对于所有整数 $n \ge 1$，令

$$(G_n, X^n, U^n) \sim \mathrm{DBM}\left(n, k, P, \frac{\log n}{n}\mathbf{Q}, P^{(n)}_{U|X}\right).$$

精确恢 只要存在一个 $\epsilon > 0$ 使得.....，则 (G_n, X^n, U^n) 的恢复就无法实现。 t

$$\mathrm{TV}\left(\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}} \times P^{(n)}_{U|X}(\cdot|s), \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}} \times P^{(n)}_{U|X}(\cdot|t)\right) \le 1 - n^{-1+\epsilon}.$$

证明。令 $\mathcal{C}_r$ 表示顶点集 $\mathcal{V}_n$ 中属于社区 r （对应$r \in [k]$）的节点集合。令 $\mathcal{H}$ 表示从属于 $\mathcal{C}_s \cup \mathcal{C}_t$ 的节点中随机选取的子集：对 $\mathcal{C}_s \cup \mathcal{C}_t$ 中的每个顶点，独立地以概率 $1_{\log^3 n}$ 选取。我们考虑一个精灵辅助问题，其中集合 $\mathcal{H}$ 之外所有节点所属的社区均已揭示，而给定的算法 $\mathscr{A}$ 用于确定 $\mathcal{H}$中节点所属的社区。

算法 $\mathscr{A}$必须根据每个节点 $v \in \mathcal{H}$观测到的二元组（u，$\mathbf{d}$）$\in \mathcal{U}^{(n)} \times \mathbb{Z}_+^k$，即数据标签 u和度分布 $\mathbf{d}$，将其分配到社区 s或 t。

设 $\mathcal{I}_s$ 表示算法 $\mathscr{A}$ 分配给社区 s的数据标签和度分布的集合， $\mathcal{I}_t$ 表示算法分配给社区 t的数据标签和度分布的集合。因此，数据标签和度分布的集合划分为 $\mathcal{I}_s \sqcup \mathcal{I}_t$。

对于任意算法 $\mathscr{A}$，二元假设检验的贝叶斯误差有如下下界：

$$
\begin{aligned}
P_e(v) &\geq \min\{p_s, p_t\} \left(\sum_{\langle u, \mathbf{d}(u)\rangle \in \mathcal{I}_s} P_{U|X}(u|t) \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}}(\mathbf{d}(u)) + \sum_{\langle u, \mathbf{d}(u)\rangle \in \mathcal{I}_t} P_{U|X}(u|s) \mathcal{P}_{\boldsymbol{\mu}_s^{(n)}}(\mathbf{d}(u)) \right) \\
&\geq \min\{p_s, p_t\} \left(\sum_{\langle u, \mathbf{d}(u)\rangle \in \mathcal{I}_s} P_{U|X}(u|t) \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}}(\mathbf{d}(u)) + 1 - \sum_{\langle u, \mathbf{d}(u)\rangle \in \mathcal{I}_s} P_{U|X}(u|s) \mathcal{P}_{\boldsymbol{\mu}_s^{(n)}}(\mathbf{d}(u)) \right) \\
&\geq \min\{p_s, p_t\} (1 - \mathrm{TV}(\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}} \times P_{U|X}(\cdot|s), \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}} \times P_{U|X}(\cdot|t))) \\
&\geq \min\{p_s, p_t\} n^{-1+\epsilon}
\end{aligned}
$$

因此，在集合 $\mathcal{H}$ 中，有 $\Omega(n \cdot n^{-1+\epsilon})$ 个顶点被算法 $\mathscr{A}$误分类。我们从算法 $\mathscr{A}$误分类的顶点集合中选取一个大小为 $\lceil \log n \rceil$ 的集合 $\mathcal{W}$。

接下来我们将证明，这些位于 $\mathcal{W}$ 中的 $\lceil \log n \rceil$ 个节点在 $\mathcal{H}$ 中没有邻居，因此，只要知道已由先知揭示的节点所属的社区，就能精确计算它们的度分布。设 $\mathcal{H}_t \subseteq \mathcal{C}_t$ 表示 $\mathcal{C}_t$ 的一个固定子集，其基数为 $\lceil \log n \rceil$。设 $\mathcal{R}$^ 表示事件 $\mathcal{H}_t$ 在 $\mathcal{H}$中没有边。我们有

$$
\begin{aligned}
\Pr(\mathcal{R}) &\geq (1 - p_{\max})^{|\hat{\mathcal{H}}_t|(|\mathcal{H}_t| - |\hat{\mathcal{H}}_t|) + \binom{|\hat{\mathcal{H}}_t|}{2}} \\
&\geq (1 - p_{\max})^{|\hat{\mathcal{H}}_t||\mathcal{H}_t|} \\
&\overset{(a)}{\geq} \exp\left(-2 p_{\max} |\hat{\mathcal{H}}_t||\mathcal{H}_t| \right) \\
&\geq 1 - 2 p_{\max} |\hat{\mathcal{H}}_t||\mathcal{H}_t| \\
&\overset{(b)}{\geq} 1 - O(1/\log n),
\end{aligned}
$$

其中 $p_{\max} = \max_{i,j} Q_{ij} \log n/n$，且 (a) 由 $\log(1 - p_{\max}) \geq -2p_{\max}$ for $0 \leq p_{\max} \leq 1/2$,^ 推出；而 (b) 则由以下事实推出：$|\mathcal{H}| = \Theta(\quad {}^{n}_{\log^3 n})(1 \pm o_n(1))$、$p_{\max} = O(\log\ n/n)$ 和 $|\mathcal{H}_t| = \lceil \log\ n \rceil$。

因此，算法 $\mathscr{A}$ 会将这些 $\lceil \log n \rceil$ 个节点在 $\mathcal{W}$ 中错误分类，这表明任何算法都不可能实现精确恢复。

□

定理的证明 *4.5*。假设对于 s，$t \in [k]$，$s \neq t$，有 $D_{s,\ t} < 1$。根据（4），存在一个 $\epsilon > 0$ 以及一个由正整数组成的无限集 $\mathcal{N}$，使得对任意 $n \in \mathcal{N}$，都有

$$
D_{\mathrm{CT}}\left(\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}}, P_{U|X}^{(n)}(\cdot|s) \middle\| \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}}, P_{U|X}^{(n)}(\cdot|t) \right) \leq (1 - \epsilon) \log n,
$$

换言之，根据（8），我们有

$$\sum_{u\in\mathcal{U}}(P_{U|X}^{(n)}(u|s))^{\lambda_u^*}(P_{U|X}^{(n)}(u|t))^{1-\lambda_u^*} \times \exp\left(\sum_{r=1}^{k}-\lambda_u^*\mu_{s_r}^{(n)}-(1-\lambda_u^*)\mu_{t_r}^{(n)}+(\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}\right)\geq n^{-1+\epsilon}. \tag{12}$$

定义集合 $\mathcal{D}_{s,t}$，其中包含数据标签与度分布对，具体如下:

$$\mathcal{D}_{s,t}=\left\{\langle u,\mathbf{d}(u)\rangle\middle|\begin{matrix}\mathbf{d}(u)=(d_1(u),d_2(u),\ldots,d_k(u)),\\ d_r(u)=\lfloor(\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}\rfloor \text{ for } r\in[k],u\in\mathcal{U}\end{matrix}\right\}$$

社区 s 中的节点具有数据标签和度分布 $\langle u,\mathbf{d}(u)\rangle$ 且属于集合 $\mathcal{D}_{s,t}$ 的概率下界如下:

$$\begin{aligned}
\Pr(\langle u,\mathbf{d}(u)\rangle) &= P_{U|X}^{(n)}(u|s)\prod_{r=1}^{k}\exp(-\mu_{s_r}^{(n)})\frac{(\mu_{s_r}^{(n)})^{d_r(u)}}{d_r(u)!}\\
&\overset{(a)}{\geq} P_{U|X}^{(n)}(u|s)\prod_{r=1}^{k}\exp(-\mu_{s_r}^{(n)})\frac{(\mu_{s_r}^{(n)})^{(\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}}}{C_1(\log n)d_r(u)!}\\
&\overset{(b)}{=} (P_{U|X}^{(n)}(u|s))^{\lambda_u^*}(P_{U|X}^{(n)}(u|t))^{1-\lambda_u^*}\\
&\quad\times\prod_{r=1}^{k}\frac{\exp(-\lambda_u^*\mu_{s_r}^{(n)}-(1-\lambda_u^*)\mu_{t_r}^{(n)})((\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*})^{(\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}}}{C_1(\log n)d_r(u)!}\\
&\overset{(c)}{\geq} (P_{U|X}^{(n)}(u|s))^{\lambda_u^*}(P_{U|X}^{(n)}(u|t))^{1-\lambda_u^*}\\
&\quad\times\prod_{r=1}^{k}\frac{\exp(-\lambda_u^*\mu_{s_r}^{(n)}-(1-\lambda_u^*)\mu_{t_r}^{(n)})((\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*})^{(\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}}}{C_2\log n\sqrt{(\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}}((\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}/e)^{(\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}}}\\
&\overset{(d)}{=}\frac{1}{C_2(\log n)^{3k/2}}(P_{U|X}^{(n)}(u|s))^{\lambda_u^*}(P_{U|X}^{(n)}(u|t))^{1-\lambda_u^*}\\
&\quad\times\prod_{r=1}^{k}\exp(-\lambda_u^*\mu_{s_r}^{(n)}-(1-\lambda_u^*)\mu_{t_r}^{(n)}+(\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*})
\end{aligned}\tag{13}$$

其中，(a)由以下事实可得： $\mu_{s_r}^{(n)}=\Theta(\log\ n)$ $\mu_{t_r}^{(n)}=\Theta(\log\ n)$，且

$$(\mu_{s_r}^{(n)})^{\lfloor(\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}\rfloor}\geq(\mu_{s_r}^{(n)})^{(\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}}/(C_1\log n),$$

(b)由式(10)可得；(c)由斯特林近似 $\lfloor x\rfloor!\ \leq C\sqrt{x}(x/e)^x$可得；$(d)$由以下事实可得： $\mu_{s_r}^{(n)}=\Theta(\log\ n)$ $\mu_{t_r}^{(n)}=\Theta(\log\ n)$。

注意，使用与上述相同的论证，我们可以证明社区 t 中的节点具有数据标签，且其度分布属于集合$\mathcal{D}_{s,\ t}$ 的概率下界为 (13)。由此，我们可以对度分布与数据标签分布之间的全变差进行界定

如下

$$
\begin{aligned}
&1-\mathrm{TV}\Big(\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}} \times P_{U|X}^{(n)}(\cdot|s), \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}} \times P_{U|X}^{(n)}(\cdot|t)\Big) \\
&\quad = \sum_{u\in\mathcal{U},\mathbf{d}\in\mathbb{Z}_+^k} \min\Big\{P_{U|X}^{(n)}(u|s)\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}}(\mathbf{d}), P_{U|X}^{(n)}(u|t)\mathcal{P}_{\boldsymbol{\mu}_t^{(n)}}(\mathbf{d})\Big\} \\
&\quad \geq \sum_{\langle u,\mathbf{d}\rangle\in\mathcal{D}_{s,t}} \min\Big\{P_{U|X}^{(n)}(u|s)\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}}(\mathbf{d}), P_{U|X}^{(n)}(u|t)\mathcal{P}_{\boldsymbol{\mu}_t^{(n)}}(\mathbf{d})\Big\} \\
&\quad \overset{(a)}{\geq} \frac{1}{C_2(\log n)^{3k/2}} \sum_{u\in\mathcal{U}} (P_{U|X}^{(n)}(u|s))^{\lambda_u^*} (P_{U|X}^{(n)}(u|t))^{1-\lambda_u^*} \\
&\qquad\quad \times \prod_{r=1}^{k} \exp\Big(-\lambda_u^*\mu_{s_r}^{(n)} - (1-\lambda_u^*)\mu_{t_r}^{(n)} + (\mu_{s_r}^{(n)})^{\lambda_u^*}(\mu_{t_r}^{(n)})^{1-\lambda_u^*}\Big) \\
&\quad \overset{(b)}{\geq} C_3 \frac{n^{-1+\epsilon}}{(\log n)^{3k/2}} = C_3 \exp((-1+\epsilon)\log n - o_n(1)),
\end{aligned}
$$

其中（a）由（13）得出，而（b）由（12）得出。由于 $(\log n)^{3k/2} = n^{o(1)}$，上述界意味着，对于所有充分大的 n，

$$
\mathrm{TV}\Big(\mathcal{P}_{\boldsymbol{\mu}_s^{(n)}} \times P_{U|X}^{(n)}(\cdot|s),\ \mathcal{P}_{\boldsymbol{\mu}_t^{(n)}} \times P_{U|X}^{(n)}(\cdot|t)\Big) \leq 1 - n^{-1+\epsilon/2}.
$$

因此，引理 4.11适用（取 $\epsilon/2$），精确恢复不可实现。

□

5 模拟

本节通过实证研究双社区 DBM 在对数度数区间内的精确恢复问题。潜在标签的取值等概率，$P = (0.5,0.5)$，边独立地以概率 $\Pr\{(i,\ j) \in E \mid X_i = s,\ X_j = r\} = (\log n/n)\ Q_{sr}$生成，其中 $Q = \begin{pmatrix} a & b \\ b & a \end{pmatrix}$ 满足 $a > b$。侧信息由擦除标签信道提供：对于每个顶点，以概率 $1 - n^{-\alpha}$ 揭示其真实标签，否则将其替换为擦除符号（见例 4.8）。在这一对称设置下，决定精确恢复的散度为

$$
D_{1,2} \ = \ \alpha \ + \ \frac{(\sqrt{a} - \sqrt{b})^2}{2},
$$

以及定理 4.4 和 4.5 指出，当且仅当 $D_{1,\ 2} > 1$ 时，才能实现精确恢复。因此，我们在所有图中标出了 ***DBM*** 阈值 $a^{\star}_{\mathrm{DBM}}(b,\ \alpha) = \big(\sqrt{b} + \sqrt{2(1-\alpha)}\big)^2$，并将其与 ***SBM*** 阈值 $a^{\star}_{\mathrm{SBM}}(b) = \big(\sqrt{b} + \sqrt{2}\big)^2$进行比较。我们通过两项互补指标量化性能：翻转不变误分类误差（误分类顶点的比例，并在所有全局标签置换中取最小值）和精确恢复概率（ERP），后者对应于定义2.5中的精确恢复概念。

基于图的算法（DBM、迭代 DBM、SBM、迭代 SBM 和谱方法）首先对邻接矩阵进行谱划分，将其分为两个簇。除谱基线（仅使用图）外，我们随后在对数度数区间内对每个顶点应用局部最大后验更新。对于候选类别 s 和节点 v，得分由对数先验、侧标签的对数似然以及

度分布的泊松对数似然 $d(v) = (d_1(v)\ d_2(v))$ 组成，其均值为 $\mu_{sr} = p_r Q_{sr} \log n$。DBM 方法在该得分中同时使用图和侧信息；SBM 基线则省略侧标签项。我们考察单轮细化（谱初始化后进行一次最大后验更新）和迭代细化（算法 1），后者重复进行局部更新，直至标签变化数少于 10^{-3} 或达到五次迭代的上限。本文报告的实验采用 不拆分变体，即谱初始化和最大后验细化均使用完整图（因此两个阶段并不独立）。为完整呈现相图，我们还展示了仅数据最大后验（仅使用侧信道）。所有评估均对标签翻转保持不变，并且每次重复实验使用独立的随机种子。

5.1 实验 1：跨越 $(a,\ \alpha)$ 的相图

我们的目标是确定精确恢复的经验转变位置，并直观展示侧信息如何使其偏移。固定 $n = 1000$ 和 $b = 10$，遍历 $a \in \{14, 15, \ldots, 23\}$ 和 $\alpha \in \{0.2, 0.4, 0.6, 0.8\}$，并针对每个 $(a,\ \alpha)$，分别对 DBM、迭代 DBM、SBM、迭代 SBM、谱方法（仅图）和仅数据运行 $M = 1000$ 次独立试验。我们展示两组截面： $1 -$ ERP 随 a变化（纵轴为线性刻度），以及平均误差随 a 变化（纵轴为对数刻度），汇总于图2和图 3。完整 $(a,\ \alpha)$ 网格上的热力图见图 4 （平均准确率）和图 5（ERP）；每幅图均标出了理论阈值：DBM 阈值曲线 $a^\star_{\mathrm{DBM}} = (\sqrt{b} + \sqrt{2(1-\alpha)})^2$，描绘了 $(a,\ \alpha)$ 平面中的边界 $D_{1,\ 2} = 1$；以及 SBM 阈值，其以 $a^\star_{\mathrm{SBM}}(b) = (\sqrt{b} + \sqrt{2})^2$ 处的一条竖线表示。

经验转变紧跟 DBM 阈值，但存在有限样本右移偏差；随着真实阈值移向更小的 a，这一偏差会增大。以 ERP 为评判标准，DBM 在 $\alpha = 0.2$ 分别为 0.4、0.6、0.8 时，达到 95% 的位置分别为 $a = 21$、19、18、17，达到 99% 的位置分别为 $a = 22$、21、20、18；见表 1–2。 $a = 22$、21、20、18 处的 99% 交点大致位于 $+1.4$、 $+0.9$、 $+1.5$ 和 $+2.6$ 的右侧$a^\star_{\mathrm{DBM}}(b,\ \alpha)$（相对偏差约为 7%、5%、9 % 和 18%），这是 $n = 1000$ 以及较低 a 下期望度数较小所导致的自然结果。相比之下，SBM、迭代 SBM 和谱方法的聚类结果都集中在 SBM 阈值附近：三者的 ERP 均需超过 95%，且（除 $\alpha = 0.2$ 外）均需$a = 23$ 才能超过 99%，这与 $a^\star_{\mathrm{SBM}}(10) \approx 20.94$ 及其有限-n 偏移相符。

仔细查看 $a^\star_{\mathrm{DBM}}$ 附近的截面，可以分离出侧信道的影响。在$\alpha = 0.4$ （$a^\star_{\mathrm{DBM}} = 18.13$）处，DBM的ERP从 $a = 18$ 处的 ≈ 0.862 跃升至 $a = 19$处的 ≈ 0.952 ，而SBM在相同横坐标处仅为 ≈ 0.135 和 ≈ 0.459 。在 $\alpha = 0.8$（$a^\star_{\mathrm{DBM}} = 14.40$）处，DBM从 $a = 14$ 处的 ≈ 0.631 增至 $a = 15$处的 ≈ 0.823 ，而SBM和谱方法在该处仍接近于零。因此，对SBM而言处于亚临界的同一图密度，一旦将侧信息纳入局部最大后验概率评分，便足以实现精确恢复。

算法 1 迭代谱初始化与MAP精化（模拟中使用的不拆分变体）

要求： 图 $G=(V,\ E)$，包含 $|V|=n$；侧信息观测 $\{U_v\}_{v\in V}$，信道为 $P_{U|X}$；类别先验 $P=(p_1,\ \ldots,\ p_k)$；亲和矩阵 $Q\in\mathbb{R}_{>0}^{k\times k}$；容差 $\varepsilon>0$；最大 MAP迭代次数$T_{\max}\in\mathbb{N}$。

确保： 估计标签 $\widehat{\sigma}\in[k]^n$。

谱初始化

$\sigma'\leftarrow$ SpectralClustering(G,k)。

标签规范化（置换消解）

定义信息锚点集合

$$S\leftarrow\Big\{v\in V:\ \exists s\neq t \text{ with } P_{U|X}(U_v\mid s)\neq P_{U|X}(U_v\mid t)\Big\}.$$

如果 $S\neq\varnothing$ **那么** ▷使用侧信息进行规范化

用 $C\in\mathbb{R}^{k\times k}$ 构造

$$C_{i,t}\leftarrow\sum_{\substack{v\in S:\\ \sigma'(v)=i}}\log P_{U|X}(U_v\mid t).$$

计算使 $\sum_{i=1}^{k}C_{i,\ \pi(i)}$ 最大化的 $\pi\in S_k$（例如，在 $-C$ 上使用匈牙利算法）。

else ▷无信息f性辅助 h者：使用 G进行规范化

对于 $i\in[k]$，令 $m_i\leftarrow|\{v:\ \sigma'(v)=i\}|$；对于 $1\le i\le j\le k$，令 $E_{ij}\leftarrow|\{(u,v)\in E:\ \sigma'(u)=i,\ \sigma'(v)=j,u<v\}|$。

任选一个确定性最大化器

$$\pi\in\arg\max_{\pi\in S_k}\left[\sum_{i=1}^{k}m_i\log p_{\pi(i)}+\sum_{1\le i\le j\le k}E_{ij}\log Q_{\pi(i),\pi(j)}\right].$$

结束 if

重新标记 $\sigma'\leftarrow\pi\circ\sigma'$ 并设置 $\widehat{\sigma}^{(0)}\leftarrow\sigma'$。

对 G进行最大后验精化

预先计算所有 $s,r\in[k]$对应的 $\mu_{sr}\leftarrow p_rQ_{sr}$log n。

对于 $t=1$ **到** $T_{\max}$**，执行**

对于所有 $v\in V$**，执行**

For each $r\in[k]$, compute $d_r(v)\leftarrow|\{u:\ (v,u)\in E,\ \widehat{\sigma}^{(t-1)}(u)=r\}|$.

对于每个 $s\in k[,\ ]$设置

$$\mathsf{s}_v(s)=\log p_s+\log P_{U|X}(U_v\mid s)+\sum_{r=1}^{k}\log\mathrm{Poisson}\big(d_r(v);\mu_{sr}\big).$$

$\widehat{\sigma}^{(t)}(v)\ \leftarrow\mathrm{argmax}^{s\in[k]}\ \mathsf{s}_v(s)$.**结束**

循环$\delta\leftarrow\frac{1}{n}\sum_{v\in V}\mathbf{1}\{\widehat{\sigma}^{(t)}(v)\ \neq\widehat{\sigma}^{(t-1)}(v)\}$.

如果 $\delta<\bar{}\varepsilon$**，则跳出循环****结束**

if **结束循环** **返回** $\widehat{\sigma}^{(t)}$。

方法	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$
DBM	21	19	18	17
迭代 DBM	21	19	18	17
SBM	22	22	22	22
迭代 SBM	22	22	22	22
谱方法	22	22	22	22
仅数据	–	–	–	–

表 1：各方法的 ERP 达到 $\geq 95\%$ 时所需的最小整数 a，随 α ($n = 1000$,$b = 10$, $M = 1000$) 变化。'–' 表示在 $a \in \{14, \ldots, 23\}$ 范围内未达到该水平。

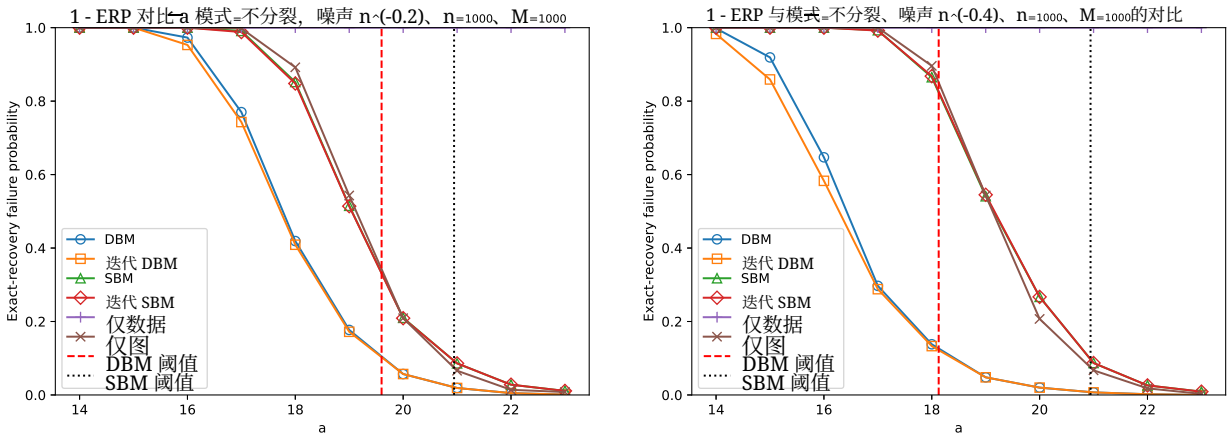
方法	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$
DBM	22	21	20	18
迭代 DBM	22	21	20	18
SBM	–	23	23	23
迭代 SBM	–	23	23	23
谱方法	23	23	23	23
仅数据	–	–	–	–

表 2：对于每种方法，ERP 达到 $\geq 99\%$ 时的最小整数 a，作为 α ($n = 1000$,$b = 10$, $M = 1000$) 的函数。'–'表示在 $a \in \{14, \ldots, 23\}$ 内未达到该水平。

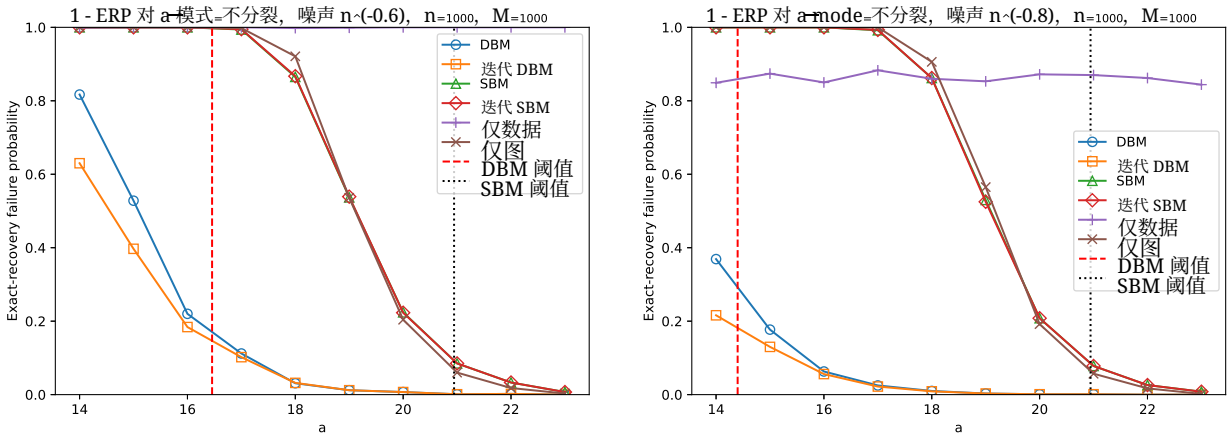
对数误差图阐明了为什么在 ERP 达到饱和之前，平均误差看起来可能很小。在 $\alpha = 0.8$和 $a = 14$ 时，DBM 的准确率已经超过 0.999，但 ERP 仅为 ≈ 0.63。这说明了趋近于零的平均误差（少数几个错误）与零错误（精确恢复）这一更严格事件之间预期存在的差距：在 $n = 1000$时，即使只有一个错误，对平均误差的贡献也仅为 10^{-3}，但会导致无法实现精确恢复。在转变附近，a 每增加一个单位，DBM 的平均误差通常会乘以约 2.5–5（在对数刻度上，每单位下降 0.4–0.7 个数量级），直到曲线在蒙特卡罗误差下限处趋于平缓。

通过算法 1 进行的迭代对平均误差影响甚微，但在侧信息充足时，于相变附近带来适度且有意义的 ERP 增益。对夹住 $a^{\star}_{\mathrm{DBM}}(b, \alpha)$ 的两个 a刻度取平均后，$\mathrm{ERP}_{\mathrm{iter}} - \mathrm{ERP}_{\mathrm{one}}$ 的增幅在 $\alpha \in \{0.2, 0.4\}$ 处为 ≈ 0.0025，在 $\alpha = 0.6$ 处升至 ≈ 0.023，并在 $\alpha = 0.8$ 处达到 ≈ 0.10，这与如下直觉相符：借助更干净的邻域分配进行第二轮打分，会翻转最后那些顽固的顶点。基线表现符合预期：仅数据 ERP 与 a 无关，并与擦除标签公式 $\mathbb{E}[2^{-K}] = (1 - n^{-\alpha}/2)^n$相符（例如，在 $n = 1000$ 处为 ≈ 0.136，$\alpha = 0.8$），而谱方法跟踪 SBM，并在 SBM 阈值附近趋近于 1。

运行时间趋势补充了统计图景。图6 显示实际运行时间随 a 在$\alpha = 0.8$ 时的变化（其他 α 呈现相同模式）。单次遍历 DBM 与 SBM 泊松评分器基本持平，接近每次重复实验 0.55 秒。迭代 DBM 与迭代 SBM 耗时 2-3 秒（当较小的 a时），并降至 ≈ 0.56 秒（随着 a 增大），反映出问题变容易后迭代次数减少；迭代 DBM 的平均迭代次数从约 2.25（$\alpha = 0.2$ 处）降至 1.62（$\alpha = 0.8$处），迭代 SBM 约为 2.6，贯穿 α。谱方法始终较快（0.15-0.20 秒），仅数据则几乎瞬时完成。



(a) $\alpha = 0.2$ (b) $\alpha = 0.4$



(c) $\alpha = 0.6$ (d) $\alpha = 0.8$

图 2：固定 α ($n = 1000$, $b = 10$, $M = 1000$) 时， $1-$ERP 与 a 的关系。垂直参考线标记$a^{\star}_{\mathrm{DBM}}(b,\ \alpha)$ 和 $a^{\star}_{\mathrm{SBM}}(b)$。

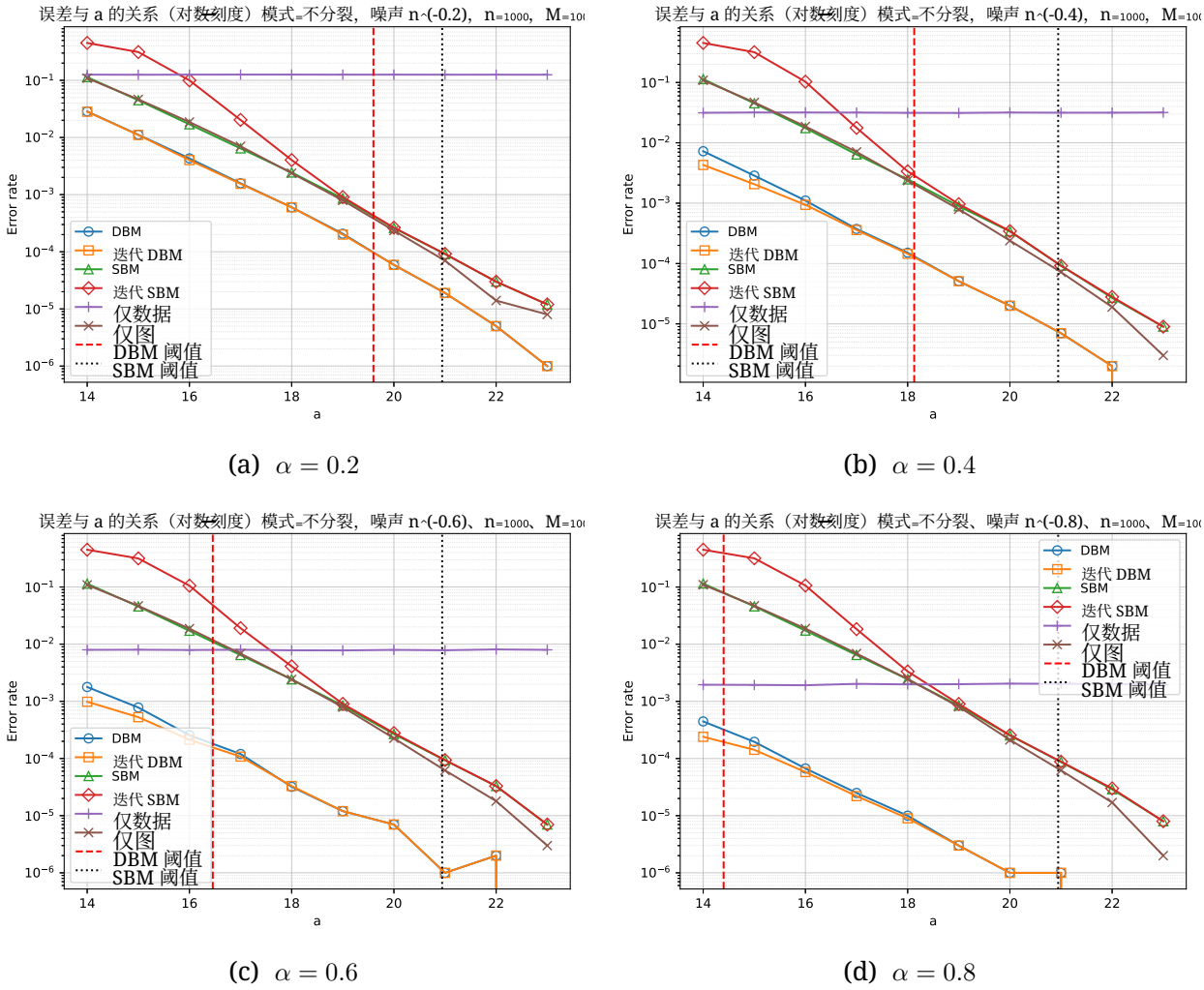
5.2 实验 2： n 中的有限尺寸标度

接下来，我们考察随着 n 增大，有限问题以多快的速度继承渐近图景。为分离该效应，我们设定（b, α） = （10，0.3）并选择

$$a \;=\; 1.10\, a^{\star}_{\mathrm{DBM}}(b,\alpha) \;=\; 1.10\big(\sqrt{10}+\sqrt{2(1-0.3)}\big)^2 \;\approx\; 20.77,$$

对于 DBM 而言，这是超临界的，且严格低于 SBM 阈值 $a^{\star}_{\mathrm{SBM}}(10) \approx 20.94$。我们改变$n \in \{10, 10^2, 10^3, 10^4\}$，并对每个 n运行 $M = 1000$次 DBM、迭代 DBM、SBM 和迭代 SBM 的独立试验。图 7a 展示失效概率 $1-$ERP 与 n 的关系；图 7b 在双对数坐标轴上展示平均误差与 n的关系；数值汇总见表 3。

这些趋势体现了阈值的有限样本特征。由于 a 被选在 DBM 阈值之上 10% 且低于 SBM 阈值，DBM 的 $1-$ERP 随着 n衰减，而 SBM 的 $1-$ERP 则上升。到 $n = 10^4$时，DBM 达到 ERP ≈ 0.992，而 SBM 下降至 ERP ≈ 0.892。与此同时，典型误差迅速缩小：DBM 的平均误差从 $n = 10$降低了几个数量级至 $n = 10^2$，并且对于 $n \geq 10^3$而言已无法与零区分，这与



(a) $\alpha = 0.2$

(b) $\alpha = 0.4$

(c) $\alpha = 0.6$

(d) $\alpha = 0.8$

图 3：平均误分类 误差（对数刻度）随 a 变化，在固定的 α ($n = 1000, b = 10, M = 1000$) 下 .

$1/n$ 基准相近；尽管 SBM 对于精确恢复处于亚临界状态，其平均误差也会衰减，这凸显了平均误差趋于零与错误数为零这一事件之间的差异。迭代（算法 1）在谱初始化良好后对平均误差影响很小，但在较大的 n 时通过修正残余误差，ERP 会略有改善。

运行时间随 n近线性增长，从 $n = 10$ 时的几毫秒上升到 $n = 10^4$ 时每次重复实验约八秒。小规模时，迭代变体比单次遍历慢一个固定倍数，但随着 n 增大，差距逐渐缩小，因为算法 1 通常在一次迭代后就趋于稳定：平均迭代次数从 约2.2（当 $n = 10$ 时）降至 1.0（当 $n = 10^4$时）。因此，在大型问题中，迭代带来的统计收益几乎无需额外成本。

6 结论

本文研究了数据块模型 （DBM）中的精确恢复问题。该模型是随机块模型的一种扩展，将图连通性与节点关联的侧信息结合起来。基于对数度数区间内随机块模型的 Chernoff–Hellinger 散度框架，我们引入了 Chernoff–TV 散度，并证明它精确刻画了在边按对数度缩放时，DBM 中精确恢复的相变——

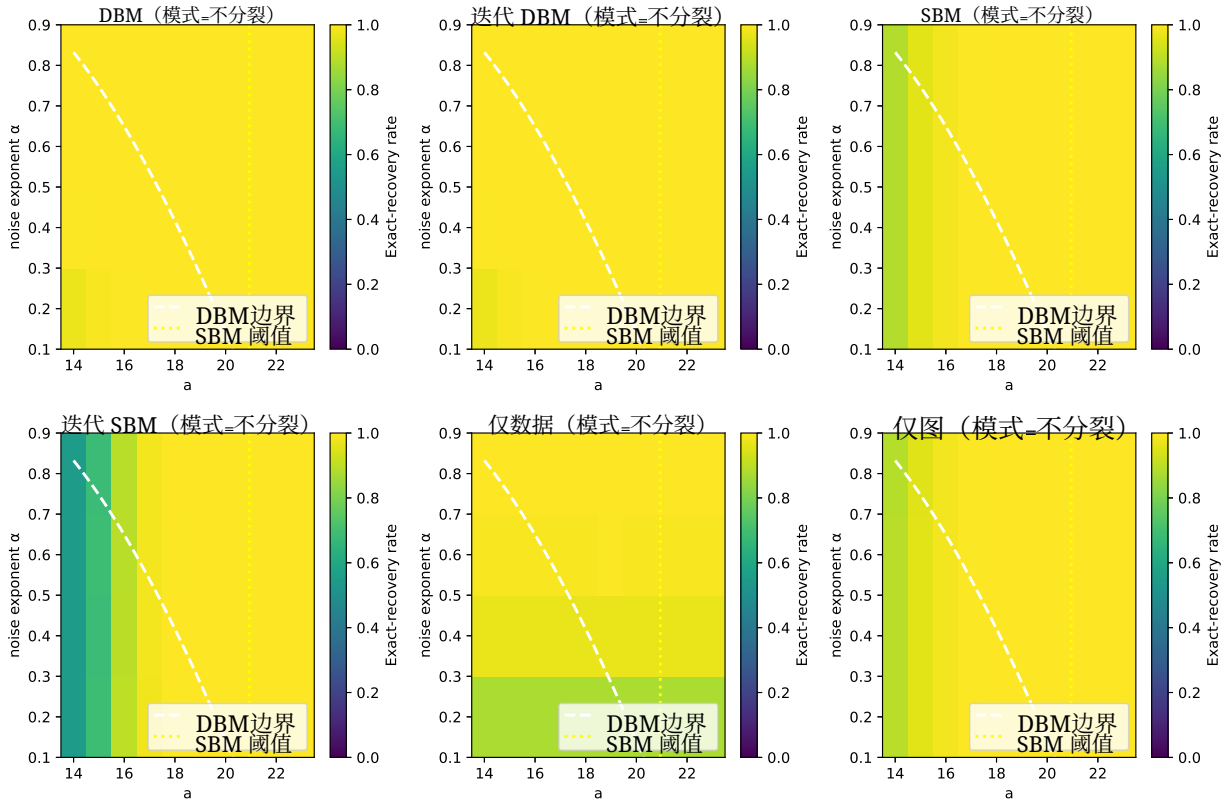


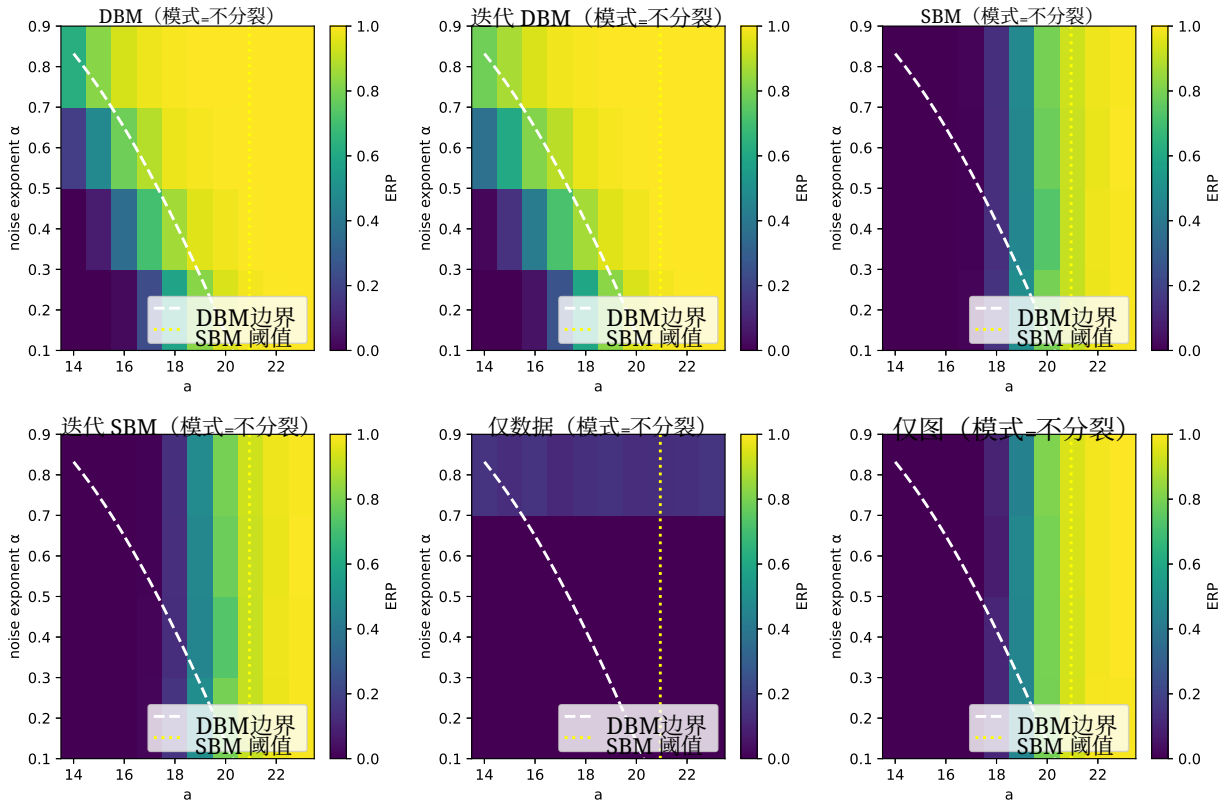
图 4：$(a,\ \alpha)$ 网格上的平均准确率热力图。DBM 阈值曲线$a^{\star}_{\mathrm{DBM}}(b,\ \alpha) = (\sqrt{b} + \sqrt{2(1-\alpha)})^2$ 与 $a^{\star}_{\mathrm{SBM}}(b)$ 处的 SBM 阈值 $= (\sqrt{b} + \sqrt{2})^2$ 叠加显示。

我们建立了匹配的可实现性结果和逆向结果，并提出了一种高效的多项式时间算法（事实上，接近线性时间），能够达到信息论阈值。在对称双社区设置下的模拟验证了预测的相变，并展示了纳入节点属性的优势。

我们的分析阐明了辅助信息何时以及如何从根本上提升可恢复性，使其超越仅凭图所能达到的水平。这些结果在统一的、基于散度的刻画框架下，整合并细化了此前研究的若干辅助信息模型，包括噪声标签、部分揭示的标签和节点特征。在受限的节点属性 SBM 场景中，我们的结果还重新审视了一项近期提出的精确恢复充分条件，并表明该条件按原表述并非必要条件；同时，我们基于 Chernoff–TV 散度给出了修正后的精确阈值。

致谢

Amir R. Asadi 获得 Leverhulme Trust 基金（ECF-2023-189）和 Isaac Newton Trust 基金（23.08(b)）的资助。



Fig图5：ERP 在（a，α）网格上的热力图，并叠加相同的阈值 s.

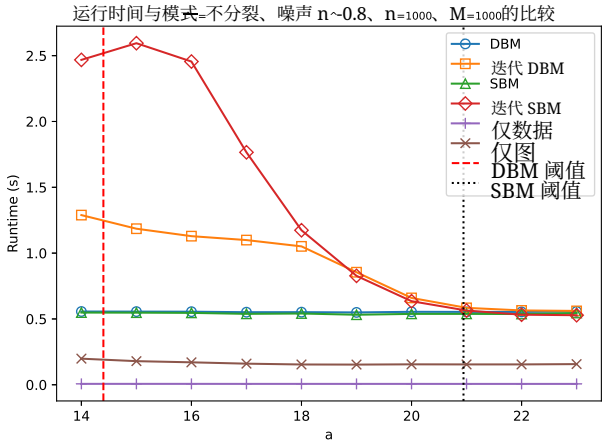
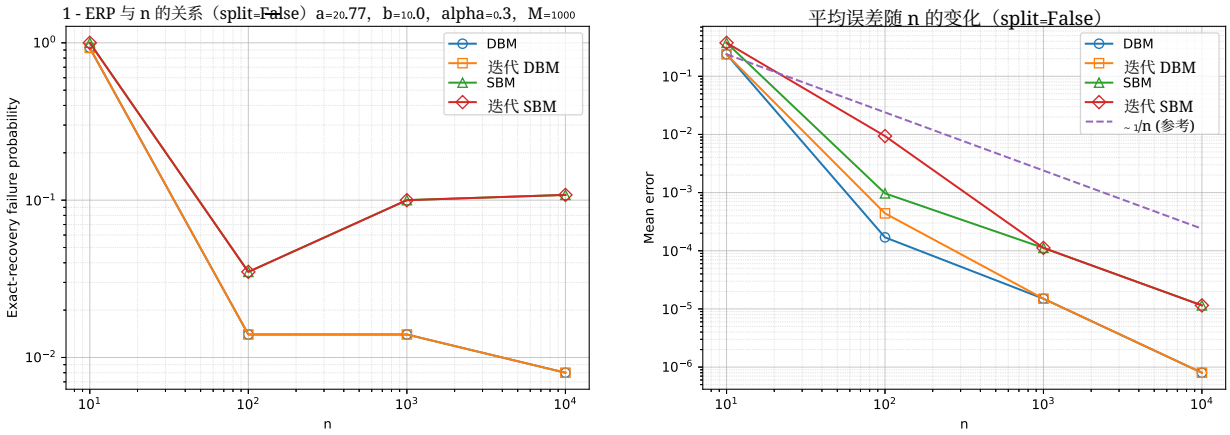


图6： a at $\alpha = 0.8(n = 1000, b = 10, M = 1000)$ 对应的运行时间。迭代方法在较小的 a 时速度较慢，并随着 a 增大而接近单遍运行时间；谱方法和仅数据方法最快。其他 α 也呈现相同的定性模式。



(a) $1-\mathrm{ERP}$ 与 n 在 $(a,\ b,\ \alpha) = (20.77,\ 10,\ 0.3)$ 时的关系。(b) 相同参数下，平均误差与 n 的关系（双对数坐标）；细线表示斜率为 -1。

图 7：有限尺寸标度。DBM 处于超临界状态，其失效概率随 n衰减；而 SBM 处于亚临界状态，其失效概率则增加。两者的典型误差都迅速衰减，但只有 DBM 的精确恢复概率呈上升趋势。

n	误差$_{\mathrm{DBM}}$	误差$_{\mathrm{SBM}}$	$\mathrm{ERP}_{\mathrm{DBM}}$	$\mathrm{ERP}_{\mathrm{SBM}}$	时间$_{\mathrm{DBM}}$	时间$_{\mathrm{SBM}}$	迭代次数$_{\mathrm{iDBM}}$	迭代次数$_{\mathrm{iSBM}}$
10	0.240	0.376	0.071	0.000	0.004	0.004	2.20	2.42
10^2	4×10^{-4}	9×10^{-3}	0.986	0.965	0.046	0.046	1.03	1.10
10^3	0	10^{-4}	0.986	0.900	0.532	0.521	1.08	1.10
10^4	0	0	0.992	0.892	7.24	7.83	1.00	1.00

表 3：在 $(a,\ b,\ \alpha) = (20.77,\ 10,\ 0.3)$ 处的缩放结果，每个 n取 $M = 1000$ 次重复实验的平均值。DBM 位于其阈值之上，而 SBM 位于其阈值之下。各项依次报告平均误差、精确恢复概率、运行时间（每次重复实验的秒数），以及迭代变体的平均迭代次数。

参考文献

Emmanuel Abbe。图压缩：簇的影响。载于*2016*年通信、控制与计算第*54*届*Allerton*年会（*Allerton*），第 1–8 页。IEEE，2016 年。Emmanuel Abbe。社区检测与随机块模型：近期进展。《机器学习研究杂志》，18(1)：6446–6531，2017 年。Emmanuel Abbe 和 Colin Sandon。一般随机块模型中的社区检测：恢复的基本极限与高效算法。载于*2015*年*IEEE*第*56*届计算机科学基础年会，第 670–688 页。IEEE，2015a。Emmanuel Abbe 和 Colin Sandon。在未知参数的情况下恢复一般随机块模型中的社区。《神经信息处理系统进展》，28，2015b。Emmanuel Abbe、Afonso S. Bandeira 和 Georgina Hall。随机块模型中的精确恢复。《IEEE信息论汇刊》，62(1)：471–487，2016 年。doi：10.1109/TIT.2015. 2490670。Amir R. Asadi、Emmanuel Abbe 和 Sergio Verd´u。带簇图上的数据压缩。载于*2017*年*IEEE*国际信息论研讨会（*ISIT*），第 1583–1587 页。IEEE，2017 年。Lars Backstrom、Dan Huttenlocher、Jon Kleinberg 和 Xiangyang Lan。大型社交网络中的群体形成：成员关系、增长与演化。载于第*12*届*ACM SIGKDD*国际知识发现与数据挖掘会议论文集，第 44–54 页，2006 年。B´ela Bollob´as、Svante Janson 和 Oliver Riordan。非齐次随机图中的相变。《随机结构与算法》，31(1)：3–122，2007 年。Guillaume Braun、Hemant Tyagi 和 Christophe Biernacki。具有最优性保证的上下文随机块模型迭代聚类算法。载于国际机器学习会议，第 2257–2291 页。PMLR，2022 年。Herman Chernoff。基于观测值之和的假设检验的渐近效率度量。《数学统计年鉴》，第 493–507 页，1952 年。Aur´elien Decelle、Florent Krzakala、Cristopher Moore 和 Lenka Zdeborov´a。模块化网络随机块模型的渐近分析及其算法应用。《物理评论E—统计、非线性与软物质物理》，84(6)：066106，2011 年。Yash Deshpande、Subhabrata Sen、Andrea Montanari 和 Elchanan Mossel。上下文随机块模型。载于神经信息处理系统进展，第 31 卷，2018 年。Maximilien Dreveton、Felipe Schreiber Fernandes 和 Daniel R. Figueiredo。节点属性随机块模型的精确恢复与 Bregman 硬聚类。载于神经信息处理系统进展（*NeurIPS*），2023 年。Mohammad Esmaeili 和 Aria Nosratinia。带次级潜变量的社区检测。载于*2020*年*IEEE*国际信息论研讨会（*ISIT*），第 1355–1360 页，2020 年。doi：10.1109/ISIT44484.2020.9174105。

Mohammad Esmaeili、Hussein Metwaly Saad 和 Aria Nosratinia。利用侧信息进行社区检测的半定规划。*IEEE* 网络科学与工程汇刊，8(2):1957–1973，2021。doi: 10.1109/TNSE.2021.3078612。Fabio Gasparetti、Giuseppe Sansonetti 和 Alessandro Micarelli。社交推荐系统中的社区检测：综述。*Applied Intelligence*，51(6):3975–3995，2021。Michelle Girvan 和 Mark EJ Newman。社交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。Anna Goldenberg、Alice X Zheng、Stephen E Fienberg、Edoardo M Airoldi 等。统计网络模型综述。*Foundations and Trends® 机器学习*，2(2):129–233，2010。Paul W Holland、Kathryn Blackmond Laskey 和 Samuel Leinhardt。随机块模型：初步研究。*Social networks*，5(2):109–137，1983。Frank McSherry。随机图的谱划分。载于第*42*届 *IEEE* 计算机科学基础研讨会论文集，第529–537页。IEEE，2001。Elchanan Mossel、Joe Neeman 和 Allan Sly。植入二分模型的一致性阈值。*Electronic Journal of Probability*，21，2016。Sara Rahiminejad、Mano R Maurya 和 Shankar Subramaniam。生物网络中社区检测算法的拓扑与功能比较。*BMC Bioinformatics*，20:1–25，2019。Herbert Robbins。关于斯特林公式的一则注记。*The American Mathematical Monthly*，62(1): 26–29，1955。Hussein Saad 和 Aria Nosratinia。利用侧信息进行社区检测：随机块模型下的精确恢复。*IEEE* 信号处理精选主题期刊，12(5): 944–958，2018。doi: 10.1109/JSTSP.2018.2834874。Hussein Saad 和 Aria Nosratinia。利用连续值侧信息进行社区检测的精确恢复。*IEEE* 信号处理快报，26(2):332–336，2019。doi: 10.1109/LSP.2018. 2889920。Avirup Saha、Shreyas Sheshadri、Samik Datta、Niloy Ganguly、Disha Makhija 和 Priyank Patel。利用部分标记的随机块模型理解基于图的半监督学习的成功原因。载于 Christian Bessiere（编），第二十九届国际人工智能联合会议论文集，*IJCAI-20*，第1345–1351页。国际人工智能联合会议组织，2020。主会场。Dhiman Sarma、Wahidul Alam、Ishita Saha、Mohammad Nazmul Alam、Mohammad Jahangir Alam 和 Sohrab Hossain。使用社区检测算法进行银行欺诈检测。载于*2020*年第二届计算应用创新研究国际会议 *(ICIRCA)*，第642–646页。IEEE，2020。Jin Sima、Feng Zhao 和 Shao-Lun Huang。利用侧信息实现平衡随机块模型的精确恢复。载于*2021*年 *IEEE* 信息论研讨会，第1–6页，2021。doi: 10.1109/ITW48936.2021.9611438。

Jisang Yoon、Kangwook Lee 和 Changho Suh。社区结构与社区特征的联合恢复。载于*2018*年第*56*届*Allerton*通信、控制与计算年会，第686–694页，2018。doi: 10.1109/ALLERTON.2018.8636058。

Qiaosheng Zhang 和 Vincent YF Tan。《一般超图随机块模型中的精确恢复》。*arXiv* 预印本 *arXiv:2105.04770*，2021。

Xiaofeng Zhao、WeiZhao 和 Mingao Yuan。《具有标签信息的超图社区检测的信息极限》。*Symmetry*, 13(11)，2021。ISSN 2073-8994。doi：10.3390/sym13112060。
URLhttps://www.mdpi.com/2073-8994/13/11/2060。