arXiv:2601.05883v1 [cs.DS] 9 Jan 2026

生日悖论时间内的谱聚类

Michael Kapralov Ekaterina Kochetkova WeronikaWrzos-Kaminska 洛桑联邦理工学院 洛桑联邦理工学院 洛桑联邦理工学院

摘要

给定一个(k, φ, ϵ)-可聚类图中的一个顶点，也就是说，该图的顶点集可以划分为若干两两不相交的 φ-扩展图，每个扩展图的大小为 $\approx n/k$ ，且外导通率不超过 ϵ，我们能否快速判断它属于哪个簇？这是一个经典问题，可追溯到 Goldreich 和 Ron'11 提出的扩张性测试问题（ $k = 2$的情形），长期以来受到文献的广泛关注。对于 $k = 2$, 从给定顶点出发进行长度为对数级的 $\approx n^{1/2+O(\epsilon/\varphi^2)}$ 随机游走并取样本，便可通过生日悖论近似确定该顶点的簇归属：随机游走样本彼此"接近"的两个顶点很可能属于同一个簇，否则很可能属于不同簇。

一般情形 $k > 2$ 的研究由 Czumaj、Peng 和 Sohler [STOC'15], 发起，而 Chiplunkar 等人 [FOCS'18], 和 Gluch 等人 [SODA'21] 的研究表明，对于一般的 k，使用 $\approx \text{poly}(k) \cdot n^{1/2+O(\epsilon/\varphi^2)}$个随机游走样本及相同的查询时间就足够了。这一结果与 $k = 2$ 的结果相比，仅相差关于 k的多项式因子，但其中存在一个概念上的不一致：如果生日悖论确实是这里查询复杂度的主导现象，那么随着簇数 k增加，查询复杂度应该 $decrease$，而不是增加！由于簇的大小为 $\approx n/k$，我们预计需要$\approx (n/k)^{1/2+O(\epsilon/\varphi^2)}$ 个随机游走样本；簇数 k 越多，所需样本就越少，当 $k \approx n$时甚至降为常数。然而，目前已知的最佳查询时间（见 Gluch 等人 [SODA'21]）却会随着k 增加，因为对随机游走样本进行线性代数后处理的计算开销很大。

在本文中，我们通过对数长度游走的样本混合，为（k、 φ、 ϵ）-可聚类图中的顶点设计了一种新颖的表示。这种表示不仅使用了最优的 $\approx (n/k)^{1/2+O(\epsilon/\varphi^2)}$ 每个顶点的游走数量，还支持快速的最近邻搜索：给定一组表示各个簇的 $\approx k$顶点以及一个查询顶点 x，我们可以在表示大小为 x 的近线性时间内找到 x 所属的簇。这构造出了一个查询时间为$\approx (n/k)^{1/2+O(\epsilon/\varphi^2)}$ 、空间复杂度为 $k \cdot (n/k)^{1/2+O(\epsilon/\varphi^2)}$ 的谱聚类预言机，达到了生日悖论界限。

目录

1 引言 **1**

2 技术概述 **4**

2.1 嵌入的构造 $\tilde{f}_x$ .6 2.2 NearestNeighbor的高效实现.8 2.3 空间与查询复杂度权衡.11 2.4 获取簇代表元（预处理）.11 2.5 以次线性时间近似求解 k .13

3 预备知识与符号 **14**

4 主要结果 **16**

4.1 快速最近邻搜索.16 4.2 预处理.18 4.3 关于 S 的良好分散集 S 和典型顶点.19 4.4 Sketch（算法 3）的分析.22 4.5 IsCovered （算法 6）和 FindNeighbors（算法 7）的分析.31 4.6 Preprocessing （算法 5）的分析.33 4.7 定理 4.1 的证明.38

定理 2.10 的证明 41

A.1 (关于 $\epsilon(x)$ 的切比雪夫逼近证明）..... 43 A.2 多项式基与定义域的变换（命题 A.3 的证明）..... 49

B 对所有 $x \in V$ 51 的 p_x^t 的 ℓ_2-范数界

C 谱嵌入的性质（引理 3.7 和引理 4.24 的证明） **53**

D 证 p良好表 p示集合和 yp典型顶点 **56**

D.1 命题 4.20 的证明.65 D.2 几乎所有簇都能被 S 良好表示（引理 4.28 的证明）.66

E 用随机游走逼近 $\left\langle f_x, \sum_y \sigma_y f_y \right\rangle$ （引理 4.21 的证明） **74**

F 定理 2.12 和定理 2.13 的证明 **78**

G 证明 Sketch o f 定理 1.3 82

1 引言

图聚类是数据分析中的核心问题，其应用遍及众多领域，从数据挖掘到社会科学、统计学等。图聚类的目标是将图的顶点集划分为互不相交的“连接良好”的子图，这些子图彼此之间的连接则较为稀疏。本文聚焦于该问题的一个热门版本：输入图具有良好的隐藏聚类，而算法的任务是高效地恢复这一聚类。形式化地说，我们假设输入图$G=(V,E)$可以将顶点集 V划分为大小为 $\approx n/k$的簇 $C_1,C_2,\ldots,C_k$，使得各个簇 C_i 诱导出连接良好的子图（即 φ-扩展图），且具有稀疏边界（即割 ∂C_i是 ϵ-稀疏的）。这类图称为 (k,φ,ϵ)-可聚类图，是随机块模型 [Abb17]的一种自然的最坏情况类比。本文的目标是设计高效的算法$\widehat{\ }$ $\widehat{\ }$ $\widehat{\ }$：给定一个 (k,φ,ϵ)-可聚类图 G，恢复一个近似聚类 $C_1,C_2,\ldots,C_k$，使其能够良好地近似 $C_1,C_2,\ldots,C_k$。形式化地说，我们希望误分类率较低，即必须存在一个置换 $\pi:[k]\to[k]$，使得

$$\sum_{i\in[k]}|C_i\triangle\widehat{C}_{\pi(i)}|\approx\epsilon'\cdot n \tag{1}$$

对于某个较小的 $\epsilon'\approx\epsilon$[1]。实现 (1) 的经典方法是谱聚类。在谱聚类中，首先对每个 $x\in V$ 计算其谱嵌入 $f_x\in\mathbb{R}^k$，从而将图聚类简化为一个更易处理的问题：对嵌入后的顶点 $\{f_x\}_{x\in V}$ 在 $\mathbb{R}^k$ 中进行聚类。事实上，可以证明，对于属于同一簇的大多数顶点对 $x,\ y\in V$，其点积 $\langle f_x,\ f_y\rangle$ 较大；而对于属于不同簇的大多数顶点对，其点积 $\langle f_x,\ f_y\rangle$ 较小（更多细节见第 2 节）。嵌入 f_x 恰好是拉普拉斯矩阵中 $k\times n$ 个最小特征向量所组成矩阵的第 x 列，因此可以通过奇异值分解求得。然而，这是一项全局计算，对于大型图而言成本高得令人难以承受。本文希望设计一种用于聚类的次线性时间算法。本文的核心目标是：

> Design a *locally computable* version $\tilde{f}_x$ of the spectral embedding f_x of $x\in V$.

$\widetilde{\ }$这种嵌入 $\{f_x\}_{x\in V}$应满足什么要求？简而言之，我们希望$\widetilde{\ }$确保 $\{f_x\}_{x\in V}$ 能够实现以下非常简单（且理想化）的聚类算法，用于处理$(k$、 φ、 $\epsilon)$-可聚类图。

算法 1 IdealizedPreprocessing(G)

$S\leftarrow$ 从每个簇中采样一个顶点

计算 $\widetilde{f}_y$ 对所有 $y\in S$

返回 $\{\ \widetilde{f}_y\}_{y\in S}$

$\widetilde{\ }$算法 2 IdealizedQuery$(x,\{f_y\}_{y\in S})$

$\widetilde{\ }$计算 f_x

返回 NearestNeighbor$(\ \widetilde{f}_x,\{\widetilde{f}_y\}_{y\in S})$

我们理想化的预处理算法（算法1）只需从每个簇中采样一个顶点[2], com-$\widetilde{\ }$为样本 S 中的每个 y 计算嵌入 f_y，并存储这些嵌入。因此，数据结构占用的总空间是 k 个单独嵌入大小。我们的理想化查询原语（算法2）给定一个查询 x，只需返回样本 S 中与之最近的点 y 作为答案。为了让查询高效，至关重要的是，我们的嵌入支持快速最近邻搜索原语

[1]由于给定的 (k,φ,ϵ)-可聚类图可能存在彼此不一致的聚类，且这些聚类在 $\Omega(\epsilon)$ 比例的顶点上存在分歧，因此误分类率不可能优于 $O(\epsilon)$。目前已知的最佳结果由 [GKL+21]给出，其误分类率为 $O(\epsilon)$，但运行时间关于 k 为指数级； [Sin16] 的工作则实现了 $O(\sqrt{\epsilon})$ 的误分类率，运行时间关于 k 和 n 为多项式级（为简便起见，我们假设 $\varphi=\Omega(1)$）。[2] 当然，实际算法无法对每个簇采样一个顶点，因为它并不知道这些簇！我们的实际预处理算法通过使用过滤步骤解决了这一问题——见算法 5。

$\sim$ $\sim$NearestNeighbor(f_x, $\{f_y\}_{y\in S}$)，它能快速返回

$$\arg\max_{y\in S} \left\langle \widetilde{f}_x, \widetilde{f}_y \right\rangle . \tag{2}$$

扩展性测试的经典研究 [KS08、NS10、GR11、CS10、KPS13]可以视为将上述框架实例化为

$$\widetilde{f}_x = \textsc{Sample}\left(M^t \mathbb{1}_x, n^{1/2+O(\epsilon/\varphi^2)}\right), \tag{3}$$

其中， M 是 G上惰性随机游走的转移矩阵，步数 t 满足 t =(C/φ^2)log n，其中 C是足够大的常数；对于分布 p和整数 s，函数Sample(p, s)输出从 p中抽取 s 个独立样本所得到的经验分布。

对于 $\widetilde{f}_x$， $k>2$的大小。 由 (3) 定义的嵌入 f_x 大小为 $n^{1/2+O(\epsilon/\varphi^2)}$。这里的 $n^{1/2}$ 因子源于生日悖论的严格限制：由于簇 C_1， C_2 的大小为 $\approx n/2$，若随机游走的数量显著小于 $\sqrt{n}/2$，那么大多数情况下只会得到 $\langle \widetilde{f}_x, \widetilde{f}_y\rangle = 0$，因而无法为大多数成对的 x， y提供任何有关簇结构的信息。对于 $k>2$ 个簇，基于生日悖论的同一论证表明，我们应当预期 $\widetilde{f}_x$ 的大小为 $\approx (n/k)^{1/2+O(\epsilon/\varphi^2)}$，其中主要因子是簇大小的平方$\sim$根。文献 [GKL$^+$21]中提出了一种嵌入 $\{f_x\}_{x\in V}$ 的构造，按照 (1)，当 ϵ 为常数时，该构造可得到常数误分类率；事实上，其大小会随k**增大**，而上述基于生日悖论的推理却表明它应该**减小**！这是由于 [GKL$^+$21] 所用技术存在一个关键的低效之处，我们将在第 2 节中详细说明。这引出了一个问题：

> Can one design embeddings $\{\widetilde{f}_x\}_{x\in V}$ of size $\approx (n/k)^{1/2+O(\epsilon/\varphi^2)}$?

虽然生日悖论表明，嵌入大小的主导因子应为 $(n/k)^{1/2}$，但根据 [CKK$^+$18]，额外的 $(n/k)^{O(\epsilon/\varphi^2)}$ 因子是必要的。 [CKK$^+$18]证明了区分扩展图（即 1-可聚类图）与 (2,Ω(1) ϵ)-可聚类图的查询复杂度下界为 $n^{1/2+\Omega(\epsilon)}$。将这一下界推广到 k 簇的情形，可得 $(n/k)^{O(\epsilon/\varphi^2)}$ ——更多细节请参见第 2.5 节。

高效的最近邻搜索? 算法 2 中的 NearestNeighbor 过程能有多高效? 假设嵌入的大小为 $\approx (n/k)^{1/2+O(\epsilon/\varphi^2)}$，计算点积 $\langle \widetilde{f}_x, \widetilde{f}_y\rangle$需要 $\approx (n/k)^{1/2+O(\epsilon/\varphi^2)}$ 时间。因此，对 S 中的点进行简单的线性扫描，会使 NearestNeighbor的运行时间达到 $k\cdot(n/k)^{1/2+O(\epsilon/\varphi^2)}$，而所需的最短时间为 $(n/k)^{1/2+O(\epsilon/\varphi^2)}$，即$\sim$表示的大小 f_x。

> 能否使 NearestNeighbor 的运行时间与 $\widetilde{f}_x$ 的大小呈线性关系，即在 $\approx (n/k)^{1/2+O(\epsilon/\varphi^2)}$ 时间内完成?

我们的贡献。 我们的主要贡献是对上述两个问题都给出了肯定的回答：我们构造了大小为 $\approx (n/k)^{1/2}\cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))}$ 的局部可计算嵌入 $\{\widetilde{f}_x\}$，并设计了一种 Near-estNeighbor 基元，其运行时间为 $\approx (n/k)^{1/2}\cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))}$，在上述框架中使用时（算法1和算法 2），其误分类率$\epsilon' = \epsilon^{\Omega(1)}$ 可达到式 (1) 所述水平。

将我们的方法以一种非常简单的方式实例化，即可得到一种近乎查询最优的算法，用于近似簇数 k：

定理 1.1（近似簇数；定理 2.12 的非正式版本）。 可 k 近似一个 (k,φ,ϵ)-可聚类图中的簇数 $(1+\epsilon^{\Omega(1)})$，所需时间为 $\approx (n/k)^{1/2}\cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))}$。

所有能够近似 k 的已有工作，即使 ϵ 只是一个较小的常数（而不是例如随 n 趋于零），也都依赖于代价高昂的线性代数处理，因此运行时间至少为 $k^2\cdot(n/k)^{1/2+\Omega(\epsilon)}$——更多详情请参见第2.1节。

此外，对于 $k \le n^{0.999}$，我们的上述定理 1.1 的结果与 $(n/k)^{1/2+\Omega(\epsilon)}$ 的下界相匹配，指数中的 $O(\log(1/\epsilon))$ 因子除外。该下界可通过应用 [CKK+18] 的结果得到——更多详情请参见第 2.5 节。

定理 1.2（聚类预言机；定理 4.1 的非正式版本）。存在一个误分类率为 $\epsilon^{\Omega(1)}$ 的聚类预言机，其

- 预处理时间和空间复杂度 $\approx (nk)^{1/2} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))}$
- 查询时间 $\approx (n/k)^{1/2} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))}$

More 通常，我们可以在查询时间和空间复杂度之间进行权衡：

定理 1.3（聚类预言机的空间/查询权衡；非正式版本）。对于每个 $\delta \in [0,1]$，都存在一个误分类率为 $\epsilon^{\Omega(1)}$ 的聚类预言机，其

- 预处理时间为 $\approx \left(n^{1/2}k^{1/2} + k \cdot \binom{n}{k}^{1-\delta}\right) \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))}$
- 计算一个大小为 $\approx k \cdot \binom{n}{k}^{1-\delta} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))}$ 的数据结构
- 查询时间为 $\approx \binom{n}{k}^{\delta} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))}$

因此，数据结构的大小与查询时间的乘积为 $\approx n^{1+O(\epsilon/\varphi^2 \log(1/\epsilon))}$，**与簇数 k无关**。

注 1.4。此前这类权衡中最优的结果由 *[GKL+21]*（定理 *3*）给出，其要求数据结构的大小与查询时间的乘积随 k 多项式增长（有关这种低效性的原因，详见第 *2* 节）。我们的权衡只要求数据结构的大小与查询时间的乘积为 $\approx n$，**且与 k无关**！

注 1.5。定理 *1.3* 以非正式的方式表述，旨在说明通过我们的方法可以实现的空间—查询权衡。我们没有给出完整的形式化表述或证明，因为其论证遵循与我们主要分析相同的原则。更多详情请参见第 *2.3* 节；有关证明 *Sketch* 及所需修改的讨论，请参见 *G* 节。

已有工作。Goldreich 和 Ron [GR11] 关于扩张性测试的经典工作主要考虑了设置 $\epsilon = 0$，并使用了（3）。后续工作 [KS08、NS10、GR11、CS10、KPS13] 产生了聚类预言机，可处理 $k = 2$ 以及使用（3）的任意小常数 $\epsilon > 0$。[CPS15] 的工作首次考虑了一般 k 的情形，但仅适用于 $\epsilon \ll {}^{1}/_{\mathrm{poly}(k)\log n}$。他们证明，只要将样本数增加一个关于 k 的多项式因子，同样的选择（3）仍然有效。一般 ϵ 和一般 k 的情形需要新的技术：当簇数 k 大于 2 [CKK+18] 时，基于嵌入（3）的点积或距离分类器无法提供足够的信息。为了解决这一困难，[GKL+21] 在随机游走分布的后处理中引入了一组线性代数工具。不幸的是，这会导致运行时间产生关于 k 的多项式级膨胀。例如，对于 $\widetilde{f}_x$GKL$_{x\in V}$ 的嵌入（[] +21]，即使计算 $\|\widetilde{f}_x\|_2^2$ 也至少需要 k 上的二次开销——更多详情请参见第 2 节。

与 [GKL+21]的比较。[GKL+21]的工作实现了以下权衡。他们证明，对于每个 $\delta \in (0, 1/2]$，都存在一个谱聚类预言机，其误分类率为 $O(\epsilon \log k)$，预处理时间为 $2^{\frac{1}{\epsilon}k^{O(1)}} n^{1-\delta+O(\epsilon/\varphi^2)}$，空间复杂度为 $\approx k^{O(1)} n^{1-\delta+O(\epsilon/\varphi^2)}$，查询时间为 $\approx k^{O(1)} n^{\delta+O(\epsilon/\varphi^2)}$。特别是，空间复杂度与查询时间的乘积为 $\approx k^{O(1)} \cdot n$。$k^{O(1)}$ 开销是他们的方法固有的，因为他们的谱嵌入是通过一次"全局"计算得到的（更多细节请参见第 2 节）。因此，我们无法使用他们的工具，需要开发一种新方法。

矩阵多项式。 我们的嵌入基于这样一种思路：将多项式作用于随机游走转移矩阵，从而将随机游走嵌入 (3) 转化为实质上的谱嵌入 f_x。我们精心选择多项式，以确保其系数较小，即 $n^{O(\epsilon/\varphi^2\log(1/\epsilon))}$，从而能够高效估算所得嵌入之间的点积。利用矩阵多项式来改变矩阵特征值的想法由来已久。然而，在数值线性代数中，通常采用多项式精度，因此系数的量级为多项式级通常就足够了[3]。不过，据我们所知，我们的工作是首次将这些技术应用于亚线性聚类算法。特别是，与典型的数值线性代数应用不同，我们的多项式系数的量级至关重要：它直接转化为相对于主要 $(n/k)^{1/2}$因子的开销，并且几乎达到了信息论最优界（指数上仅相差一个 $\log(1/\epsilon)$ 因子）。

其他相关工作。 近期的工作 [JK24], [PY23] 使用 [GKL+21] 的聚类预言机，为最大割以及更一般的约束满足问题（约束满足问题）设计了测试器。近期工作 [SP23] 将 [GKL+21] 的预处理时间从关于 k 的指数时间缩短为关于 k的多项式时间，但他们的算法仅适用于 ϵ 相对于 k为多项式级小时的情况。k,φ,ϵ-可聚类模型的鲁棒版本已在 [KPS08]中研究，近期又在 [Pen20]中得到研究。

其他相关工作。 多路切格不等式（例如， [LGT14], [CKCLL+13]）可用于以较高精度恢复各个簇，但其运行时间至少与图的规模成线性关系。另一个密切相关的领域是局部聚类。局部聚类关注的是在与簇大小成正比的时间内，找出节点 v 周围的整个簇。已有若干算法解决这一问题 [ACL08、AGPT16、OA14、ST14、ALM13]，但遗憾的是，它们无法用于解决我们的问题，因为当簇的规模与图规模成线性关系时，这些算法也需要线性时间（此外，输出的簇可能相互重叠）。本文则着重于使用严格的次线性时间来解决这一问题。

2 技术概述

我们首先描述图访问模型。

图访问模型： 我们采用有界度图模型，在该模型中，我们可以指定一个顶点$x \in V$和一个数字 $i \in [d]$，并在常数时间内访问 x 的第 i个邻居。

接下来，我们定义导通率的概念，它刻画了一个簇的质量。

定义 2.1 （内部与外部导通率）。 设 $G = (V, E)$ 为一个图。对于集合 $C \subseteq V$ 和集合$S \subseteq C$，令 $E(S, C \setminus S)$ 表示一个端点在 S 中、另一个端点在 $C \setminus S$中的边的集合。 S 在 C 中的导通率为 $\phi_C^G(S) = \frac{|E(S,\, C\setminus S\,)|}{d|S|}$。 $C \subseteq V$的内部导通率定义为

$$\phi_{in}^G(C) = \min_{S\subseteq C, 0<|S|\le \frac{|C|}{2}} \phi_C^G(S).$$

C 的外导通率定义为 $\phi_{out}^G(C) = \phi_V^G(C) = \frac{|E(C,\, V\setminus C\,)|}{d|S|}$。

注释 2.2. 我们给出针对 d-正则图的证明，但该结果对于 d-有界图同样成立，导通率的定义与定义 *2.1* 相同，即以 $d|S|$ 而非 S的体积进行归一化。这是因为我们可以通过为每个顶点 $x \in V$添加 $d - deg(x)$个自环，将 G 中的 d-有界图转换为 d-正则图 G^{reg}。因此， G 上的随机游走等价于 G^{reg}上的惰性随机游走。

我们采用以下可聚类图的概念：

[3]相反，人们通常会设计递归求值原语，并证明这些原语在数值上是稳定的。

定义 *2.3*（（k, φ, ϵ）-可聚类图）。设 $G = (V, E)$ 是一个 d-正则图。G 的一个（k, φ, ϵ）-聚类是将顶点 V 划分为互不相交的子集 $C_1, \ldots, C_k$，使得对于所有 $i \in [k]$，$\phi_{in}^G(C_i) \geq \varphi, \phi_{out}^G(C_i) \leq \epsilon$，以及所有 $i, j \in [k]$，都有 $\frac{|C_i|}{|C_j|} \in O(1)$。如果图 G 存在一个（k, φ, ϵ）-聚类，则称其为（k, φ, ϵ）-可聚类图。

给定 G 的一个聚类和一个顶点 $x \in V$，我们用 $i(x) \in [k]$ 表示满足 $x \in C_{i(x)}$ 的标签。我们将 $C_{i(x)}$ 称为 x 所属的簇。我们定义 $\eta := \max_{i,\, j} \frac{|C_i|}{|C_j|}$ 为簇大小之间的最大比值。注意，根据假设，对于所有 i，j，$\frac{|C_i|}{|C_j|} \in O(1)$，我们有 $\eta = O(1)$，并且对于所有 $i \in [k]$，都有 $\frac{1}{\eta} \cdot \frac{n}{k} \leq |C_i| \leq \eta \cdot \frac{n}{k}$。我们指出，这一假设在以往的研究中很常见，例如 [GKL+21]。

定义 2.4 (谱聚类预言机)。如果随机化算法 $\mathcal{O}$ 在获得一个 d-正则图 $G = (V, E)$ 的查询访问时，该图具有一个 (k, φ, ϵ)-聚类 $C_1, \ldots, C_k$，那么当算法 $\mathcal{O}$ 提供对 V 的划分 $P = (C_1, \ldots, C_k)$ 的一致查询访问时，它就是一个(k, φ, ϵ)-聚类预言机。并且，以至少 $9/10$ 的概率（关于 $\mathcal{O}$ 的随机比特），划分 P 具有如下性质：存在某个置换$\pi : [k] \to [k]$，

$$\sum_{i \in [k]} |C_i \triangle \widehat{C}_{\pi(i)}| \leq \left(\frac{\epsilon}{\varphi^2}\right)^{\Omega(1)} n.$$

接下来，我们正式定义谱嵌入。

定义 2.5 (谱嵌入)。令 $U_{[k]} \in \mathbb{R}^{n \times k}$ 表示一个矩阵，其列为拉普拉斯矩阵 $\mathcal{L}$的底部 k个特征向量。对于每个顶点 $x \in V$，x的谱嵌入记为 $f_x \in \mathbb{R}^k$，即 $U_{[k]}$的第x行，也就是 $f_x := U_{[k]}^\top \mathbb{1}_x$。

备注 2.6。请注意，谱嵌入 f_x 并非唯一确定。不过，对于任何满足 ϵ/φ^2 小于某个绝对常数的(k, φ, ϵ)-可聚类图，点积$\langle f_x, f_y \rangle$ 是唯一确定的。（参见备注 *3.3*）。

谱嵌入的一项关键性质是，它会很好地集中在其簇均值附近：

$$\mu_i := \frac{1}{|C_i|} \sum_{x \in C_i} f_x \tag{4}$$

对于几乎所有 $x \in V$（形式化的量化见引理 3.6），嵌入都满足以下集中界：

$$\|f_x - \mu_{i(x)}\|_2^2 \leq \left(\frac{\epsilon}{\varphi^2}\right)^{\Omega(1)} \|\mu_{i(x)}\|_2^2. \tag{5}$$

(k、φ、ϵ)-可聚类图的簇均值 (4) 具有若干有用的结构性质。值得注意的是，它们的平方 ℓ_2 范数接近簇大小的倒数，并且经过适当缩放后，在 $\mathbb{R}^k$ 中构成一组近似标准正交基：

引理 2.7（引自 [GKL+21] 的引理 7）。设 $k \geq 2$ 是一个整数，$\varphi \in (0, 1)$，且 $\epsilon \in (0, 1)$。设 $G = (V, E)$ 是一个 d-正则图，它允许 (k，φ，ϵ)-聚类 C_1、.....、C_k。则有

(1) 对所有 $i \in [k]$，

$$\left| \|\mu_i\|_2^2 - \frac{1}{|C_i|} \right| \leq \frac{4\sqrt{\epsilon}}{\varphi} \frac{1}{|C_i|}$$

(2) 对所有 $i \neq j \in [k]$，

$$|\langle \mu_i, \mu_j \rangle| \leq \frac{8\sqrt{\epsilon}}{\varphi} \frac{1}{\sqrt{|C_i||C_j|}}.$$

注 2.8. 注意，(1) 意味着，对于所有 $i \in [k]$，都有 $\|\mu_i\|_2^2 = \Theta(k/n)$。

引理 2.9 (引理 9，出自 [GKL+21])。 对于所有 $\alpha \in \mathbb{R}^k$，均有

$$\left| \alpha^\top \left(\sum_{i=1}^{k} |C_i| \mu_i \mu_i^\top - I \right) \alpha \right| \leq \frac{4\sqrt{\epsilon}}{\varphi} \|\alpha\|_2^2.$$

因此，根据公式 (5) 和引理 2.7，可以利用谱嵌入的内积来检验两个顶点是否属于同一个簇。对于属于同一簇 C_i 的顶点 x，y，我们有

$$\langle f_x, f_y \rangle \approx \langle \mu_i, \mu_i \rangle = \|\mu_i\|_2^2 \approx \frac{k}{n}. \tag{6}$$

Conv反过来，对于属于不同簇的顶点，例如 $x \in C_i$ 和 $y \in C_j$、$i \neq j$，我们有 e

$$\langle f_x, f_y \rangle \approx \langle \mu_i, \mu_j \rangle \leq \left(\frac{\epsilon}{\varphi^2} \right)^{\Omega(1)} \|\mu_i\|_2 \|\mu_j\|_2 \approx \left(\frac{\epsilon}{\varphi^2} \right)^{\Omega(1)} \frac{k}{n} \ll \frac{k}{n}. \tag{7}$$

方程 (6) 和 (7) 表明，同一簇中的顶点具有较大的点积，而不同簇中的顶点点积较小。谱嵌入的这一关键性质证明了我们理想化框架（算法 1 和 2）的合理性。

~2.1 嵌入的构造 f_x

~回顾一下，我们的目标是构造可高效计算的嵌入 f_x，以便使用 IdealizedPreprocessing （算法1）和 IdealizedQuery （算法 2）的一个变体对顶点进行分类。现在我们将介绍这一构造，并概述其分析所依据的主要思路。

我们设计嵌入，使其能够将真实谱嵌入的点积近似到加性误差 $\approx \frac{k}{n}$ 以内；也就是说，对于（几乎）所有 x、$y \in V$，它们都必须满足 –

$$\left| \langle \widetilde{f}_x, \widetilde{f}_y \rangle - \langle f_x, f_y \rangle \right| \leq \xi \cdot \frac{k}{n}, \tag{8}$$

其中某个常数 ξ 充分小。 [GKL+21] 引入了一种数据结构，能够像式 (8) 那样近似谱嵌入空间中的点积，并将其称为 谱点积预言机。 [GKL+21] 的谱点积预言机需要耗费 $k^{O(1)} n^{1/2+O(\epsilon/\varphi^2)}$ 时间和查询，才能以足以进行聚类的精度近似点积。 [GKL+21] 的方法涉及计算惰性随机游走矩阵 M 的特征值分解的代理。为此，他们抽取一个大小为 poly(k) 的集合 $S \subseteq V$，并计算从 S 出发的随机游走分布的格拉姆矩阵 $\mathcal{G}$ 的特征值分解。这必然会带来 poly(k) 的运行时间开销，因为样本大小 $|S|$ 至少必须达到 k，才能覆盖所有簇。因此，我们无法采用他们的技术，需要另辟蹊径。我们设计了新的嵌入，只需 $(n/k)^{1/2+O(\epsilon/\varphi^2)}$ 时间和查询即可构建谱点积预言机，这已接近信息论最优——详见第 2.5 节的讨论。现在我们给出嵌入 f_x 的 ~ 构造。

我们的嵌入。 与以往工作不同，以往工作使用从以 $x \in V$ 为起点的 t 步游走的分布中抽取的（少量）样本作为 x 的嵌入（见 (3)）；我们则针对一个范围内的 t，使用经过精心设计的 t 步游走分布混合。具体而言，我们选取最小游走长度 $t_{\min} \approx \frac{1}{\varphi^2} \log n$，即这些簇的混合时间，以及一个参数 $\overline{\Delta} \approx \epsilon \cdot t_{\min}$。我们的嵌入将这些分布中的样本按精心选定且尤为重要的是较小的系数加以组合。形式化地说，

$$\widetilde{f}_x = \bigoplus_{t_{\min} \leq t \leq t_{\min} + \Delta} c_t \cdot \textsc{Sample}\left(M^t \mathbb{1}_x, (n/k)^{1/2+O(\epsilon/\varphi^2 \log(1/\epsilon))} \right), \tag{9}$$

其中，c_t 是满足以下条件的系数：$|c_t| = n^{O(\epsilon/\varphi^2 \log(1/\epsilon))}$，而 $\bigoplus$ 表示长度为 n的（稀疏）向量的加法。系数 c_t 是多项式 p 的系数，使得将 p 应用于 G的惰性随机游走矩阵 M，本质上等同于将 M 投影到其对应于大于 $1-\epsilon$的特征值的特征空间上。

构造多项式 p。令 $M = U\Sigma U^\top$ 表示惰性随机游走矩阵 M的特征值分解。回顾定义 2.5， $U_{[k]}$的列是 M中对应于不小于 $1-\epsilon$的特征值的特征向量，而其余 M个特征向量对应的特征值不大于 $1-\varphi^2/4$（见备注 3.2）。因此，如果我们有一个多项式

$$p(x) = \sum c_t x^t$$

将区间 $[0, 1-\varphi^2/4]$ 映射到接近 0 的位置，将区间 $[1-\epsilon, 1]$ 映射到接近 1 的位置，则矩阵

$$p(M) = Up(\Sigma)U^\top$$

近似于投影到由 $U_{[k]}$的列张成的空间上的投影。因此，我们得到

$$\langle f_x, f_y\rangle = \mathbb{1}_x^\top U_{[k]}U_{[k]}^\top \mathbb{1}_y \approx \mathbb{1}_x^\top U p(\Sigma)^2 U^\top \mathbb{1}_y = \langle p(M)\mathbb{1}_x, p(M)\mathbb{1}_y\rangle = \left\langle \sum_t c_t M^t \mathbb{1}_x, \sum_t c_t M^t \mathbb{1}_y \right\rangle. \tag{10}$$

此外，如果 p 的系数较小，且仅当 t 大于 $\approx \frac{1}{\varphi^2}\log n$（即簇的混合时间）时才非零，那么公式（10）右侧就可以按照生日悖论的思路，利用少量随机游走高效地近似。

我们的首要贡献是构造出具有所需性质的多项式 p。

定理 2.10。对于每个满足 $\epsilon/\varphi^2\log(1/\epsilon) \le c_1$ 的 $\epsilon, \varphi \ge 0$，若 $\log(1/\epsilon)\log(\varphi^2/\epsilon) \le c_2 \log n$，其中 c_1, c_2 是充分小的绝对常数，并且对于每个 $t \ge 20\frac{\log n}{\varphi^2}$，都存在一个形式如下的多项式 p

$$p(x) = x^t \cdot q(x),$$

其中 q 是一个次数为$\deg q = O(\epsilon t + \log(\varphi^2/\epsilon))$的多项式，其系数的$(1/\epsilon)^{O(\epsilon t + \log(\varphi^2/\epsilon))}$ 绝对值有界，并且

- $|p(x) - 1| \le \epsilon/\varphi^2$ *for* $x \in [1-\epsilon, 1]$
- $|p(x)| \le n^{-4}$ *for* $x \in [0, 1-\varphi^2/4]$.

我们在附录 A 中给出定理 2.10 的证明。主要思路是从多项式 x^t入手，它在 $[0, 1-\varphi^2/4]$ 上和 $x = 1$ 处具有所需的性质，但随着 x 偏离 1 而衰减得过快，因此我们要在区间 $[1-\epsilon, 1]$上将其拉平。为此，我们将它乘以一个多项式 q，该多项式在此区间上几乎可以抵消 x^t。我们通过使用函数 $(1-\epsilon x)^{-t}$的切比雪夫逼近来构造这样的多项式。我们对 $(1-\epsilon x)^{-t}$ 的切比雪夫逼近的分析深受 [AA22], 的启发；该文作者研究了对 e^x的切比雪夫逼近，而整体上，我们的方法似乎遵循了这类切比雪夫逼近的标准做法。请注意， p 的系数必须足够小。这是因为我们要近似

$$\left\langle p(M)^\top \mathbb{1}_x, p(M)^\top \mathbb{1}_y \right\rangle = \left\langle \sum_t c_t M^t \mathbb{1}_x, \sum_t c_t M^t \mathbb{1}_y \right\rangle = \sum_{t,t'} c_t c_{t'} \mathbb{1}_x^\top M^{t+t'} \mathbb{1}_y \tag{11}$$

通过采样实现，采样率以系数大小的多项式方式依赖于系数大小。具体而言，我们的多项式 p的系数上界为 $(1/\epsilon)^{O(\epsilon t + \log(\varphi^2/\epsilon))} = n^{O(\epsilon/\varphi^2 \log(1/\epsilon))}$，适用于 $\epsilon \ge 2^{-o(\sqrt{\log n})}$。我们注意到，类似的因子也出现在 [GKL+21] 的查询复杂度中，这是因为需要稳定地求逆格拉姆矩阵。该矩阵由从图中均匀随机选取的顶点出发、经过 $\approx k$ 步随机游走所得 $t = t_{\min}$分布构成，其条件数的上界只能达到 $n^{O(\epsilon/\varphi^2)}$；而且，根据 [CKK+18]的下界，这些因子也是必要的。我们还指出，构造格拉姆矩阵并应用其逆矩阵的需求意味着， [GKL+21] 的技术无法实现快于 $k^2 \cdot \sqrt{n/k}$ 的查询时间。

我们的点积预言机的运行时间。我们证明，运行 $\approx \sqrt{n/k} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))}$ 次随机游走，就能将公式（11）中的每一项 $\mathbb{1}_x^\top M^{t+t'} \mathbb{1}_y$ 近似到加性误差 $\approx \frac{k}{n}$ 以内，这正如生日悖论所预期的那样（精确表述见引理 4.21）。由此可得到一个满足 (8) 的点积预言机，其运行时间为 $\approx \sqrt{n/k} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))}$（相差 n 的多对数因子），我们注意到，对于 $k \le n^{0.99}$，该时间至多为 $(n/k)^{1/2+O(\epsilon/\varphi^2 \log(1/\epsilon))}$。这消除了 [GKL+21]中所有关于 k 的多项式损失因子；尤其是，运行时间会随着k而下降，这与 [GKL+21]中相反的增长趋势不同。

2.2 最近邻的高效实现

给定嵌入 f_x，$\{f_y\}_{y \in S}$，理想化的查询算法（算法 2）需要一种高效的最近邻搜索操作，以便快速返回

$$\arg\max_{y \in S} \left\langle \widetilde{f_x}, \widetilde{f_y} \right\rangle.$$

解决最近邻搜索问题的一种自然方法，是在嵌入空间中实现局部敏感哈希（LSH） [IM98] 。鉴于不同簇中“典型”点之间的距离与同一簇内“典型”点之间的距离存在$\epsilon^{\Omega(1)}$ 倍的差距（见（6）和（7）），人们会预期 LSH 的运行时间约为嵌入规模的 $k^{\epsilon^{\Omega(1)}}$ 倍。然而，要在嵌入空间中高效实现局部敏感哈希似乎很困难。因此，我们采用群体测试方法，该方法依赖于我们高效计算内积的能力 $\langle f_x, f_y \rangle$（事实上，这种能力得到了显著的改进——见下文）。我们的方法使运行时间几乎与嵌入规模呈线性关系。

求解该问题的判定版本。考虑这样一个子问题：给定一个顶点 $x \in V$和一个集合 $S \subseteq V$，其中每个簇至多包含一个顶点，我们的任务是高效地确定 S是否包含一个与 x属于同一簇的顶点。为简单起见，我们先考虑一个理想化情形，即已知真实的谱嵌入 f_x 和 $\{f_y\}_{y \in S}$。我们的方法是将嵌入 f_y与独立均匀随机抽取的随机符号 $\sigma_y \sim \{-1, 1\}$ 结合起来，并考察组合点积 $\left\langle f_x, \sum_y \sigma_y f_y \right\rangle$。我们证明，对于几乎所有的 $x \in V$，关于随机性σ，以较高概率有

$$\left| \left\langle f_x, \sum_{y \in S} \sigma_y f_y \right\rangle \right| \approx \begin{cases} \frac{k}{n}, & \text{if } S \text{ contains a vertex from the cluster of } x \\ \left(\frac{\epsilon}{\varphi^2}\right)^{\Omega(1)} \frac{k}{n} \ll \frac{k}{n}, & \text{otherwise} \end{cases} \tag{12}$$

（正式陈述见引理 4.19）。特别地，这可用于确定 S 是否包含一个属于 x所在簇的顶点。现在我们概述式（12）的证明，该证明依赖于（k、φ、ϵ）-可聚类图的谱结构与生日悖论类碰撞测试技术之间的相互作用。引理 4.19 的完整证明见第 4.4 节。

令 C_i 为 x 的簇。我们首先考虑 S 不包含来自 C_i 的顶点的情况。定义随机变量 $X := \left\langle f_x, \sum_{y \in S} \sigma_y f_y \right\rangle$，其随机性来自 σ。于是 $\mathbb{E}[X] = 0$，利用公式 (5)，我们可以对方差作如下界定：

$$\mathrm{Var}[X] \le \mathbb{E}[X^2] = \sum_{y \in S} \langle f_x, f_y \rangle^2 \approx \sum_{y \in S} \langle \mu_i, \mu_{i(y)} \rangle^2 \le \sum_{j \in [k] \setminus \{i\}} \langle \mu_i, \mu_j \rangle^2,$$

其中，最后一个不等式是由以下假设推出的： S 每个簇至多包含一个顶点，并且S 不包含来自 x所在簇的任何顶点。为了给出 $\sum_{j \in [k] \setminus \{i\}} \langle \mu_i, \mu_j \rangle^2$的上界，我们利用

spe（k, φ, ϵ）-可聚类图的谱结构。利用引理 2.9和引理 2.7，我们得到 t

$$\begin{aligned}
\sum_{j\in[k]\setminus\{i\}} \langle \mu_j, \mu_i\rangle^2 &\approx \frac{k}{n}\mu_i^\top \left(\sum_{j\in[k]\setminus\{i\}} |C_j|\mu_j\mu_j^\top\right)\mu_i \\
&= \frac{k}{n}\mu_i^\top \left(\sum_{j\in[k]} |C_j|\mu_j\mu_j^\top - |C_i|\mu_i\mu_i^\top\right)\mu_i \\
&= \frac{k}{n}\mu_i^\top \left(\sum_{j\in[k]} |C_j|\mu_j\mu_j^\top - I\right)\mu_i + \frac{k}{n}\mu_i^\top\left(I - |C_i|\mu_i\mu_i^\top\right)\mu_i \\
&\lesssim \frac{\sqrt{\epsilon}}{\varphi^2}\frac{k}{n}\|\mu_i\|_2^2 + \frac{k}{n}\|\mu_i\|_2^2(1-|C_i|\|\mu_i\|_2^2) && \text{by Lemma 2.9} \\
&\approx \frac{\sqrt{\epsilon}}{\varphi^2}\frac{k^2}{n^2} && \text{by Lemma 2.7,}
\end{aligned}$$

它给出了方差的上界。因此，根据切比雪夫不等式，以较大的常数概率，我们得到

$$\left|\left\langle f_x, \sum_{y\in S}\sigma f_y\right\rangle\right| \approx \left(\frac{\epsilon}{\varphi^2}\right)^{\Omega(1)}\frac{k}{n}.$$

现在考虑 S 确实包含一个与 x 属于同一簇的顶点的情况，将该顶点记为 $x^* \in C_i$。那么$\left\langle f_x,\ \sum_y \sigma_y f_y\right\rangle$由项 $\langle f_x,\ f_{x^*}\rangle \approx \frac{k}{n}$主导。与上面相同的论证表明，以较大的常数概率，$\left|\left\langle f_x, \sum_{y\in S\setminus C_i}\sigma_y f_y\right\rangle\right| \approx \left(\frac{\epsilon}{\varphi^2}\right)^{\Omega(1)}\frac{k}{n} \ll \frac{k}{n}$，由此得到 — – –

$$\left|\left\langle f_x, \sum_{y}\sigma_y f_y\right\rangle\right| \geq \langle f_x, f_{x^*}\rangle - \left|\left\langle f_x, \sum_{y\in S\setminus C_i}\sigma_y f_y\right\rangle\right| \approx \frac{k}{n},$$

这正是公式 (12) 所要求的。

注意，随机符号 σ 至关重要。举例来说，考虑 $f_x = \mu_i + \frac{\sqrt{\epsilon}/\varphi}{\sqrt{k}}\sum_{j\in[k]\setminus\{i\}}\mu_j$和 $S = \{\mu_j : j\in[k]\setminus\{i\}\}$，这与我们对输入图所知的谱界完全一致。那么 $\|f_x - \mu_i\|_2^2 \approx \frac{\epsilon}{\varphi^2}\frac{k}{n}$，但 $\langle f_x, \sum_{y\in S} f_y\rangle \approx \sqrt{k}\frac{\sqrt{\epsilon}}{\varphi}\frac{k}{n} \gg \frac{k}{n}$。换言之，$S$ 不包含来自 x 所属簇的任何顶点，但不带随机符号的内积 $\langle f_x, \sum_{y\in S} f_y\rangle$ 无法捕捉这一点。

近似 $\left\langle f_x,\ \sum_{y\in S}\sigma_y f_y\right\rangle$。根据公式 (12)，我们的任务可归结为将 $\left\langle f_x,\ \sum_{y\in S}\sigma_y f_y\right\rangle$近似到加性误差为 $\approx \frac{k}{n}$ 的范围内。理想情况下，我们希望直接使用公式 (8) 中的谱点积预言机。然而，我们不能将其作为黑盒逐项应用，因为对于大小为 $|S| = k$的集合 S，这只能得到如下误差界：

$$\left|\left\langle \widetilde{f}_x, \sum_{y\in S}\sigma_y\widetilde{f}_y\right\rangle - \left\langle f_x, \sum_{y\in S}\sigma_y f_y\right\rangle\right| = \left|\sum_{y\in S}\sigma_y\left(\left\langle \widetilde{f}_x, \widetilde{f}_y\right\rangle - \langle f_x, f_y\rangle\right)\right| \leq \left|\sum_{y\in S}\sigma_y\right|\cdot\xi\cdot\frac{k}{n} \approx \xi\cdot\frac{k\sqrt{k}}{n},$$

这比我们所能容忍的程度多了一个因子 $\sqrt{k}$。因此，我们不再分别计算每个点积，而是一次性近似整个和。为实现这一目标，我们的核心原语是下面给出的 Sketch 过程。首先，对于每个 $y\in S$，我们从在G 中以每个 $x\in S$为起点的 t 步惰性随机游走的分布中生成样本，其中 t 取自范围 $[t_{\min}, t_{\min}+t_\Delta]$，并将这些样本与定理 2.10 中多项式的系数$c_t$ 重新组合。这本质上是一种从 $p(M)\mathbb{1}_x$ 生成样本的形式。其次，Sketch 将这些关于 $x\in S$ 的经验分布与随机符号相结合，其中每个 $x\in S$对应的符号都是独立选取的。

算法 3 SKETCH(S)

$\sigma \sim$ 均匀分布 $(\{-1,1\}^S)$
$t_{\min} \leftarrow$ 20对数 n/φ^2 ▷最短游走长度
$t_\Delta \leftarrow 20 \frac{\epsilon}{\varphi^2} \cdot$ 对数 $n +$ 2对数(φ^2/ϵ) ▷ 不同游走长度的数量
$r \leftarrow O^*\left(\sqrt{\frac{n}{k}} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$ ▷游走的数量
当 $x \in S$ **时**
  对 $t = t_{\min}$ **到** $t_{\min} + t_\Delta$**，执行**
    $\widehat{p}_x^t \leftarrow$ RANDOMWALKS$(r,\ t,\ x)$ ▷ 向量 $\widehat{p}_x^t$ 的支撑集至多为 r
  结束循环
结束循环 ▷ c_t 是定理 2.10 中多项式 $p(x)$ 里的 x^t 的系数

RANDOMWALKS 原语只是从以输入顶点 x为起点的 t 步惰性随机游走的分布中生成样本：

算法 4 RANDOMWALKS(r, t, x)

从 x 出发，运行 r 次长度为 t 的惰性随机游走。令 $\widehat{p}_x(y)$ 为
最终到达 y 的随机游走所占的比例 ▷ 向量 $\widehat{p}_x$ 的支撑集大小至多为 r
返回 $\widehat{p}_x$

我们证明，对于几乎所有顶点 $x \in V$，都可以使用 SKETCH 过程正确判断 S 是否包含来自 x的簇中的一个顶点。

引理 2. **11.** 过程 SKETCH （算法 *3*）具有以下性质。其运行时间为

$$O^*\left(|S| \cdot \sqrt{\frac{n}{k}} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right),$$

并输出一个支撑大小至多为

$$O^*\left(|S| \cdot \sqrt{\frac{n}{k}} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right).$$

此外，对于每个集合 $S \subseteq V$ 以及每个相对于 S 而言是典型的顶点 $x \in V \setminus B_\delta$ （见定义 ***4.13***），以下结论成立。

- 若 $C_{i(x)} \cap S \neq \emptyset$，则

$$\Pr\left[\left|\langle \text{SKETCH}(x), \text{SKETCH}(S)\rangle\right| \geq 0.8 \left|\langle \text{SKETCH}(x), \text{SKETCH}(x)\rangle\right|\right] \geq 0.4.$$

- 若 $C_{i(x)} \cap S = \emptyset$，则

$$\Pr\left[\left|\langle \text{SKETCH}(x), \text{SKETCH}(S)\rangle\right| \leq 0.2 \left|\langle \text{SKETCH}(x), \text{SKETCH}(x)\rangle\right|\right] \geq 0.999.$$

在这两种情况下，概率是针对 SKETCH 的内部随机性而言的，并且 SKETCH 的所有调用都是相互独立的。

这里，我们使用 O^* 记法来省略 poly(φ^2/ϵ) 和 polylog n 因子。我们在第 3 节中正式定义集合 B_δ 并界定其大小，在第 4.3 节中正式定义典型顶点 的概念，并证明几乎所有顶点都是典型顶点。我们在第 4.4 节中给出引理 2.11 的证明。

寻找最近邻。借助 Sketch 原语，我们可以高效地实现最近邻搜索。给定一个顶点 x 和一个集合 S，我们只需通过查询 $\langle \textsc{Sketch}(x)\ \textsc{Sketch}(S)\rangle$（并对二分搜索过程中遇到的子集 $S' \subseteq S$ 查询 $\langle \textsc{Sketch}(x)\ \textsc{Sketch}(S')\rangle$），在 S 上执行二分搜索。我们证明，对于每个良态集合 S，除至多 $\approx (\epsilon/\varphi)^{1/3}$ 比例的顶点 x 外，我们的最近邻搜索（算法 7）都能正确返回 S 中属于 x 的簇的顶点集合（精确保证见引理 4.26）。我们的最近邻搜索可实现为：查询时间为 $\approx \sqrt{n/k} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))}$，空间复杂度为 $\approx \sqrt{nk} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))}$，预处理时间为 $\approx \sqrt{nk} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))}$；实现方法是预先计算并存储向量 $\textsc{Sketch}(S)$（以及二叉搜索树中子集 $S' \subseteq S$ 对应的 $\textsc{Sketch}(S')$）。值得注意的是，我们的最近邻搜索适用于 k 的所有取值。如果 $k \le n^{0.99}$，那么查询时间的上界为 $(n/k)^{1/2+O(\epsilon/\varphi^2 \log(1/\epsilon))}$。我们在第 4.1 节中正式描述该算法（算法 7），并在第 4.5 节中证明其保证（引理 4.26）。

2.3 空间与查询复杂度的权衡

我们的生日悖论式碰撞计数分析允许对随机游走的数量进行权衡（见定理 1.3）。分析表明，只要从查询顶点 x 出发的随机游走数量与从每个 $y \in S$ 出发的随机游走数量的乘积至少为 $\approx \frac{n}{k}$（见引理 4.21），内积 $\langle \textsc{Sketch}(x)\ \textsc{Sketch}(S)\rangle$ 就能近似所需的内积 $\langle f_x,\ \sum_{y\in S} \sigma_y f_y\rangle$。具体而言，对于任意 $\delta \in [0,1]$，我们可以从 x 出发进行 $\approx \left(\frac{n}{k}\right)^{\delta}$ 次随机游走，并从每个 $y \in S$ 出发进行 $\approx \left(\frac{n}{k}\right)^{1-\delta}$ 次随机游走，从而得到查询时间 $\approx \left(\frac{n}{k}\right)^{\delta}$ 和大小为 $\approx |S| \cdot \left(\frac{n}{k}\right)^{1-\delta} \approx k \cdot \left(\frac{n}{k}\right)^{1-\delta}$ 的数据结构，这就实现了定理 1.3 中所述的权衡。

请注意，当 $\delta \le \frac{1}{2}$ 时，仅从查询顶点 x 运行 $\approx \left(\frac{n}{k}\right)^{\delta}$ 次随机游走，不足以以足够高的精度近似 $\langle \textsc{Sketch}(x)\ \textsc{Sketch}(x)\rangle$。因此，在这种情况下，算法不应计算内积 $\langle \textsc{Sketch}(x)\ \textsc{Sketch}(x)\rangle$，而应将 $|\langle \textsc{Sketch}(x)\ \textsc{Sketch}(S)\rangle|$ 与阈值 $\frac{1}{\eta} \cdot \frac{k}{n}$ 进行比较。为简洁起见，我们省略细节，因为这与主要分析所依据的原理相同。证明概要以及所需修改的讨论见 G 节。

2.4 获取簇代表元（预处理）

到目前为止，我们假设可以访问一个集合 S，其中每个簇都包含一个代表顶点。我们使用过程 Preprocessing（算法 5）来创建这样一个集合。

算法概述：我们首先抽取一个顶点集合 S，其中可能包含来自同一簇的多个顶点。使用最近邻算法（算法 7），我们在 S 上构建一个有向相似图（算法 5 第 5–8 行）。随后，我们将其转换为无向相似图 H（算法 5 第 11 行）：只有当对应的两条有向边都存在时，才添加一条无向边。这样可以避免纳入因某一端点的最近邻搜索失败而错误构建的边。

理想情况下，我们期望从每个簇中采样的顶点形成一个连通分量。因此，我们通过从 H 中的每个连通分量各选取一个代表顶点，创建一个代表顶点的初始候选集，记为 R_{cand}（第 16–20 行）。

然而，候选集 R_{cand} 可能过大，因此我们需要进一步细化。例如，我们可能有多个没有近邻的离群顶点，每个顶点各自形成一个单例分量。为细化代表顶点集合，我们随机抽取一个顶点集合 S_{test}，并对 S_{test} 中的每个顶点执行最近邻搜索，以找出 R_{cand} 中最常被选为最近邻的顶点（第 24–32 行）。最终的代表顶点集合记为 R，由 R_{cand} 中至少一次被选为最近邻的顶点组成（第 34 行）。我们在第 4.6 节中对 Preprocessing（算法 5）进行正式分析。

算法 5 PREPROCESSING($\hat{k}$)
$\hat{k}$ 是簇数的常数因子近似

$S \leftarrow$ （多重）集合，其中包含 O 个顶点（$k \log(\varphi^2/\epsilon)$），这些顶点是独立均匀随机采样得到的
$\mathcal{T}_S \leftarrow$由 S 构成的草图树 ▷（见定义 4.5）

▷基于非对称相似性测试构造有向图
$A \leftarrow \emptyset$
对 $x \in S$ **执行**
  $C \leftarrow$ FINDNEIGHBORS(x, $\mathcal{T}_S$.根节点) ▷ 使用算法 7查找 $S \cap C_{i(x)}$
  $A \leftarrow A \cup \{(x, y)\colon\ y \in C\}$
</style}结束对

▷ 在 S样本上创建无向相关图 H
$H = (S,\ \{\{x, y\}\colon (x,\ y) \in A$和$(y,\ x) \in A\})$ ▷如果两条有向边都存在，则添加无向边
$\mathcal{C} \leftarrow H$的连通分量集合

▷ 从 H的每个连通分量中选取一个候选代表顶点
$R_{\text{cand}} \leftarrow \emptyset$ ▷存储代表顶点的候选项
对于 $\widehat{C}_i \in \mathcal{C}$ **执行**
  选择任意一个 $x_i \in \widehat{C}_i$
  $R_{\text{cand}} \leftarrow R_{\text{cand}} \cup \{x_i\}$
结束循环
$\mathcal{T}_{\text{cand}} \leftarrow R_{\text{cand}}$的草图树 ▷ （见定义 4.5）

▷减少代表顶点的数量
$S_{\text{test}} \leftarrow$ 个（多重）集合，包含 $O(k \log(\varphi^2/\epsilon))$ 个独立均匀随机采样的顶点
$c_y \leftarrow 0$ 对于每个 $y \in R_{\text{cand}}$ ▷ 记录 FINDNEIGHBORS 返回每个 $y \in R_{\text{cand}}$ 的次数的计数器
</style}对 $x \in S_{\text{test}}$ **执行**
  $N \leftarrow$ FINDNEIGHBORS(x, $\mathcal{T}_{\text{cand}}$.根节点) ▷ 使用算法 7 查找 $x$in R_{cand} 的所有邻居
  如果 $|N| = 1$，**则** ▷ 如果恰好找到一个邻居，则递增计数器
    令 y 为 N的唯一元素
    $c_y \leftarrow c_y + 1$ ▷ 计数器以 R_{cand}的元素为索引
  结束 if
结束 for

▷ 保留作为某个 $x \in S_{\text{test}}$ 的唯一最近邻的顶点
$R \leftarrow \{y \in R_{\text{cand}}\colon\ c_y > 0\}$
返回 R 的草图树 ▷（见定义 4.5）

获取候选代表 R_{cand}：我们通过抽取一个包含 O（$k\log(\varphi^2/\epsilon)$) 个顶点的集合 S 来构建相似性图，以确保 S 与除 $O(\epsilon/\varphi^2)$比例的簇以外的所有簇都相交。对于每个 $x \in S$，我们调用最近邻搜索 FINDNEIGHBORS (x, S)（算法 7）。由此得到一个有向相似图 A。如果最近邻搜索成功，那么 FINDNEIGHBORS(x, S)会正确识别 S 中与 x属于同一簇的所有顶点，并将边 $\{(x, y)\colon y \in C_{i(x)} \cap S\}$ 加入有向相似图 A。因此， x 在 A 中的出邻域恰好是 $N^+A(x) = S \cap C_{i(x)}$。所以，如果对每个 $x \in S \cap C_i$， FINDNEIGHBORS(x, S) 都成功，那么 $S \cap C_i$ 就在算法 5 第 11 行构造的无向相似图中形成自己的连通分量。若出现这种情况，簇 C_i 中恰好有一个顶点被纳入 R_{cand}，这正是我们想要的。由于 FINDNEIGHBORS 对除 $\approx(\epsilon/\varphi^2)^{1/3}$ 比例以外的所有顶点都能成功，因此我们可以证明，至多有 $\approx (\epsilon/\varphi^2)^{1/3}k$ 个簇包含一个顶点 x ，使得 FINDNEIGHBORS(x, S) 失败；而其余所有与样本 S有非空交集的簇，在 R_{cand}中都有唯一的簇代表。

精化代表顶点集合：对于 FindNeighbors未能成功的簇，可能会在相似性图中分裂成多个连通分量，从而在 R_{cand}中产生多个代表顶点。为了移除冗余代表，我们再抽取另一个包含 $O(k\log(\varphi^2/\epsilon))$个顶点的集合 S_{test}，并针对每个 $x \in S_{\text{test}}$运行最近邻搜索 FindNeighbors(x, R_{cand})（算法 7）。

如果 FindNeighbors(x, R_{cand})对顶点 $x \in C_i \cap S_{\text{test}}$成功，且 C_i是 $\approx k(1-(\epsilon/\varphi^2)^{1/3})$个簇中具有唯一代表 y_i（位于 R_{cand}中）的一个，那么 FindNeighbors(x, R_{cand})恰好会返回集合 $\{y_i\}$，此时 y_i 会被纳入精化集合 R。因此，我们可以证明，除 $\approx(\epsilon/\varphi^2)^{1/3}k$ 个簇外，其余簇在精化集合 R中仍有唯一的簇代表。另一方面，由于 FindNeighbors(x, R_{cand})至多对 $|S_{\text{test}}|$中的 $\approx(\epsilon/\varphi^2)^{1/3}|S_{\text{test}}| \approx (\epsilon/\varphi^2)^{1/3}k$ 个顶点失败，因此至多有 $\approx (\epsilon/\varphi^2)^{1/3}k$ 个来自 R_{cand} 的额外顶点被纳入 R。

2.5 在线性时间内近似 k

作为我们技术的一个简单应用，我们给出一种几乎查询最优的算法，可对 k 达到 $(1+\varepsilon)$-近似，其中 $\varepsilon = \left(\frac{\epsilon}{\varphi^2}\right)^{\Omega(1)}$。更正式地说，假设我们得到 k 的常数因子近似，并将其细化为 $(1+\varepsilon)$-近似。

定理 2.12. 令 C 为一个充分大的常数。假设 ϵ、φ 满足注 *4.2* 中的标准假设。对于每个 $\varepsilon \geq C \cdot \left(\epsilon/\varphi^2\right)^{1/4}$、ApproxK$(\varepsilon)$（算法 *8*）的运行时间为

$$O^*\left(\left(\frac{n}{k}\right)^{1/2+O(\epsilon/\varphi^2\log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))} \cdot \frac{1}{\varepsilon^2}\right)$$

并以至少 0.99 的概率找到 k 的$(1+\varepsilon)$-乘法近似。

证明概要：算法 ApproxK 背后的思路如下。假设我们可以访问簇大小倒数预言机，它对每个 $x \in V$ 返回值 $\frac{1}{|C_{i(x)}|}$。那么，我们只需从 V 中均匀随机抽取一个集合$F := \{x_1, \ldots, x_L\}$，得到数值$\frac{1}{|C_{i(x_1)}|}, \ldots, \frac{1}{|C_{i(x_L)}|}$，并用所得数值的平均值来近似 $\frac{k}{n}$。由于，

$$\mathbb{E}\left[\frac{1}{L}\sum_{l\in[L]}\frac{1}{|C_{i(x_l)}|}\right] = \sum_{i\in k}\cdot\frac{|C_i|}{n}\cdot\frac{1}{|C_i|} = \frac{k}{n},$$

我们有，$\sum_{l\in[L]}\frac{1}{|C_{i(x_l)}|}$将是 $\frac{k}{n}$ 的无偏估计量。此外，由于对于所有 $i \in [k]$，都有$|C_i| = \Theta(k/n)$，根据 Hoeffding 不等式，

$$\Pr\left[\left|\frac{1}{L}\sum_{l\in[L]}\frac{1}{|C_{x_l}|} - \frac{k}{n}\right| \geq \varepsilon\frac{k}{n}\right] \leq e^{-\Omega(\varepsilon^2\cdot L)},$$

因此，只需取 $L = O\left(\frac{1}{\varepsilon^2}\right)$。如果 $\varepsilon = \text{poly}\left(\frac{\epsilon}{\varphi^2}\right)$，这只会给访问逆簇大小预言机所需时间带来 $O^*(1)$ 的开销。

尽管我们无法访问逆簇大小预言机，但我们有一个过程，可以对几乎所有 $v \in V$ – Sketch 进行近似 $\frac{1}{|C_{i(v)}|}$。更具体地说，请注意，对于几乎所有 $v \in V$，量 $\|f_v\|_2^2$ 接近于$\frac{1}{|C_{i(v)}|}$（形式化表述见引理 3.7 和注记 3.8），并且可以使用 Sketch 来获得 $\|f_v\|_2^2$ 的近似值，除了 $v \in V$ 中的一小部分之外，对所有情况均适用（形式化表述见引理 4.24）。这些思想在 F 节的引理 F.3 和推论 F.4 中得到了形式化。

下界在 [CKK+18] 中的一个应用表明，对于常数 φ 和$\epsilon = 2^{-o(\sqrt{\log n})}$，这是近乎最优的：

定理 2.13。 任何一种算法，在能够访问一个(k, φ, ϵ)-可聚类图的情况下，若以至少 $2/3$ 的概率将 k 近似到一个 $2-\Omega(1)$ 因子以内，则至少需要进行 $\binom{n}{k}^{1/2+\Omega(\epsilon)}$ 次查询。–

3 预备知识与符号

设 $G=(V,E)$ 是一个 d-正则图，并设 $n:=|V|$。给定一个顶点 $x\in V$，我们将 $\mathbb{1}_x\in\mathbb{R}^n$ 记作 x 的指示向量，即在索引 x 处取值为 1、其他位置均为 0 的向量。我们用符号 $\mathbb{1}\in\mathbb{R}^n$ 表示所有条目均等于 1 的向量。

给定一个聚类 C_1、……、C_k 以及一个顶点 G 和顶点 $x\in V$，我们用 $i(x)\in[k]$ 表示满足 $x\in C_{i(x)}$ 的标签。我们将 $C_{i(x)}$ 称为 x 所在的簇。我们定义 $\eta:=\max_{i,j}|C_i|/|C_j|$ 为簇大小之间的最大比值。注意，根据假设，对于所有 i、j 都有 $|C_i|/|C_j|\in O(1)$，因此我们有 $\eta=O(1)$；并且对于所有 $i\in[k]$，都有 $\frac{1}{\eta}\cdot\frac{n}{k}\le|C_i|\le\eta\cdot\frac{n}{k}$。

我们令 A 表示 G 的邻接矩阵，并令 $\mathcal{L}:=I-\frac{1}{d}A$ 表示 G 的归一化拉普拉斯矩阵。我们将 $\mathcal{L}$ 的特征值记为 $0\le\lambda_1\le\dots\lambda_n\le 2$，并将相应的特征向量记为 u_1、……、$u_n\in\mathbb{R}^n$。此外，令 $U\in\mathbb{R}^{n\times n}$ 表示这样一个矩阵，其 i 列为 i 个特征向量 u_i，该特征向量对应于 $\mathcal{L}$。我们用 M 表示惰性随机游走矩阵 $M:=\frac{1}{2}I+\frac{1}{2d}A$。注意，对于所有 $i\in[n]$，u_i 是 M 的特征向量，其特征值为 $\left(1-\frac{\lambda_i}{2}\right)\in[0,1]$。我们令 Σ 表示 M 的特征值按降序排列后构成的对角矩阵。于是 M 具有特征值分解 $M=U\Sigma U^\top$。

众所周知，$(k,\ \varphi,\ \epsilon)$-可聚类图的拉普拉斯矩阵在其第 k 个与第 $(k+1)$ 个特征值之间存在特征值间隙 [LGT14] [CKK+18]:

引理 3.1（引自 [GKL+21]的引理 3）。设 $G=(V,E)$ 是一个 d-正则图，且存在一个 (k,φ,ϵ)-聚类。那么 $\lambda_k\le 2\epsilon$ 和 $\lambda_{k+1}\ge\frac{\varphi^2}{2}$。

备注 3.2。作为一个推论，惰性随机游走矩阵 M 的特征值 $\left(1-\frac{\lambda_i}{2}\right)$ 满足 $\left(1-\frac{\lambda_k}{2}\right)\ge 1-\epsilon$ 和 $\left(1-\frac{\lambda_{k+1}}{2}\right)\le 1-\frac{\varphi^2}{4}$.

备注 3.3。如果 G 是 (k,φ,ϵ)-可聚类的，且 ϵ/φ^2 小于某个常数，那么由引理 *3.1* 可知，$\mathcal{L}$ 的底部 k 个特征向量所张成的空间是唯一确定的，即 $U_{[k]}$ 的选取在右乘一个正交归一矩阵 $R\in\mathbb{R}^{k\times k}$ 的意义下是唯一的。我们注意到，尽管 $x\in V$ 对应的 f_x 的选取并不唯一，但 $x\in V$ 和 $y\in V$ 的谱嵌入之间的点积是定义良好的，因为对于任意正交归一矩阵 $R\in\mathbb{R}^{k\times k}$，都有

$$\langle Rf_x,Rf_y\rangle=(Rf_x)^\top(Rf_y)=(f_x)^\top(R^\top R)(f_y)=(f_x)^\top(f_y).$$

对于簇 C_i，我们将其谱嵌入 μ_i 定义为

$$\mu_i:=\frac{1}{|C_i|}\sum_{x\in C_i}f_x.$$

$(k,\ \varphi,\ \epsilon)$-可聚类图的谱嵌入的一个关键性质是，它们会很好地集中在各自的簇均值周围。更具体地说，我们在下文中界定偏离其簇均值达到非平凡程度的顶点 $x\in V$ 的数量。

定义 3.4. 给定 $\delta>0$，定义

$$B_\delta:=\left\{x\in V:\|f_x-\mu_{i(x)}\|_2^2>\frac{\delta}{|C_{i(x)}|}\right\}$$

为谱嵌入偏离其簇均值的顶点集合。我们有时将顶点 $x\in B_\delta$ 称为坏顶点。

不难证明，B_δ 至多占顶点的 $O\left(\frac{\epsilon}{\varphi^2}\cdot\frac{1}{\delta}\right)$ 比例。证明很简单，但这是一个我们会频繁用到的关键事实，因此在此列出。证明依赖于 [GKL+21] 中的以下引理。

引理 3.5 (引理 6，出自 [GKL+21]；“方差界”)。设 $k\ge 2$ 为整数，$\varphi\in(0,1)$，且 $\epsilon\in(0,1)$。设 $G=(V,E)$ 是一个 d-正则图，且存在 (k,φ,ϵ)-聚类 $C_1,\dots,C_k$。那么对于所有满足 $\|\alpha\|_2=1$ 的 $\alpha\in\mathbb{R}^k$，我们有

$$\sum_{i=1}^k\sum_{x\in C_i}\langle f_x-\mu_i,\alpha\rangle^2\le\frac{4\epsilon}{\varphi^2}.$$

引理 3.6。 对于每个 $\delta > 0$，都有

$$|B_\delta| \le O\left(n \cdot \frac{\epsilon}{\varphi^2} \cdot \frac{1}{\delta}\right).$$

证明。 令 α_1、……、 $\alpha_k \in \mathbb{R}^k$ 为 $\mathbb{R}^k$的一个标准正交基。将引理 3.5 应用于 α_1、……、 α_k 并求和，可得

$$\sum_{i=1}^{k} \sum_{x \in C_i} \|f_x - \mu_i\|_2^2 = \sum_{j=1}^{k} \sum_{i=1}^{k} \sum_{x \in C_i} \langle f_x - \mu_i, \alpha_j \rangle^2 \le \frac{4\epsilon k}{\varphi^2}. \tag{13}$$

另一方面，对 $x \in B_\delta$ 求和并利用 B_δ的定义，可得

$$\sum_{x \in B_\delta} \|f_x - \mu_{i(x)}\|_2^2 \ge \sum_{x \in B_\delta} \frac{\delta}{|C_{i(x)}|} \ge |B_\delta| \cdot \frac{\delta}{\max_{i \in [k]} |C_i|} \ge |B_\delta| \cdot \frac{\delta}{\eta} \cdot \frac{k}{n}, \tag{14}$$

其中，最后一个不等式由定义 $\eta = \max_{i,\, j} |C_i|/|C_j|$可得。结合方程 (13) 和 (14)，可得

$$|B_\delta| \cdot \frac{\delta}{\eta} \cdot \frac{k}{n} \le \sum_{i=1}^{k} \sum_{x \in C_i} \|f_x - \mu_i\|_2^2 \le \frac{4\epsilon k}{\varphi^2},$$

整理可得

$$|B_\delta| \le \frac{4\epsilon}{\varphi^2} \cdot \frac{\eta}{\delta} \cdot n = O\left(n \cdot \frac{\epsilon}{\varphi^2} \cdot \frac{1}{\delta}\right).$$

□

关于集合 $V \setminus B_\delta$，我们的第一个观察是：属于同一个簇的任意两个点 x、 $y \in V \setminus B_\delta$ 彼此接近。形式化地说，有

引理 3.7。 对于所有 x、 $y \in V \setminus B_\delta$，如果 x 和 y 属于同一个簇 C_i，那么

$$\left|\langle f_x, f_y \rangle - \frac{1}{|C_i|}\right| \le \left(\frac{4\sqrt{\epsilon}}{\varphi} + 4\sqrt{\delta}\right) \frac{1}{|C_i|}.$$

注记 3.8. 引理 *3.7* 表明，对于所有 $x \in V \setminus B_\delta$，通过选择 $x = y$，都有 $\left|\|f_x\|_2^2 - \frac{1}{|C_{i(x)}|}\right| \le \left(\frac{4\sqrt{\epsilon}}{\varphi} + 4\sqrt{\delta}\right) \cdot \frac{1}{|C_{i(x)}|}$。特别地，如果 δ 和 ϵ/φ^2 都小于某个充分小的常数，那么

$$\|f_x\|_2^2 = \Theta\left(\frac{1}{|C_{i(x)}|}\right) = \Theta\left(\frac{k}{n}\right)$$

对于所有 $x \in V \setminus B_\delta$，都有。我们将反复使用这一事实。

引理 3.7 的证明见 C 节。

记号。 我们使用 O^*，Ω*-记号来省略 poly(φ^2/ϵ) 和 polylog n 因子。
本文中，提及多重集的大小时，均按元素的重数计数。对于整数 l，我们用 I_l 表示 $\mathbb{R}^l$ 上的单位矩阵。若 l 可由上下文明确，则省略下标。
对于矩阵 A、 B，我们将弗罗贝尼乌斯内积 $A \bullet B = \sum_{ij} A_{ij} B_{ij}$记作 $A \bullet B$。我们常将形如 $\langle f_x, f_y \rangle^2$的表达式写成 $\langle f_x, f_y \rangle^2 = f_x f_x^\top \bullet f_y f_y^\top$.。这是因为我们的许多证明都使用了关于 f_x 的平均论证；在这种情况下，将这些表达式视为 f_x的二阶张量幂的线性函数，而非 f_x 的二次函数，会更加方便。

4 主要结果

本节中，我们证明聚类预言机所提供的保证。

定理 4.1（定理 1.2 的正式版本）。 对于任意 ϵ、 $\varphi > 0$，以及任意 $k \geq 1$，都存在一个（k, φ, ϵ）聚类预言机，其误分类率为 $O\left((\epsilon/\varphi^2)^{1/3}\right.$ 对数^2（φ^2/ϵ）$\left.\right)$，并且具有

- 预处理时间和空间复杂度 $O^*\left(\sqrt{nk} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$
- 查询时间 $O^*\left(\sqrt{\frac{n}{k}} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$。

注 4.2（关于参数假设）。 我们在定理 *2.10* 中构造多项式，以及由此得到的聚类预言机保证，都依赖于参数假设

$$\frac{\epsilon}{\varphi^2} \log(1/\epsilon) \leq c_1 \quad and \quad \log(1/\epsilon)\, \log(\varphi^2/\epsilon) \leq c_2 \log n$$

对于足够小的绝对常数 c_1、 c_2。这些条件确保分析中使用的多项式近似具有次线性（关于 n）系数，这对于实现次线性时间和空间复杂度至关重要。

在 $\frac{\epsilon}{\varphi^2} \log(1/\epsilon) > c_1$ 或 $\log(1/\epsilon)\log(\varphi^2/\epsilon) > c_2 \log n$ 中任一条件成立的情形下，我们的运行时间界中出现的因子 $n^{O(\epsilon/\varphi^2 \log(1/\epsilon))}$ 或$(1/\epsilon)^{O(\log(\varphi^2/\epsilon))}$ 会关于 n 呈多项式增长，因此总体运行时间不再是次线性的。由于已知存在多项式时间聚类预言机 *[Sin16]*，，在这种情形下可以改用 *[Sin16]* 中的算法。

在本节余下部分中，我们假设 $G = (V, E)$ 具有一个 (k, φ, ϵ)-聚类$C_1, \ldots C_k$，其中 ϵ, φ 满足上述注 4.2 中的假设。此外，我们设定以下参数。

定义 4.3（参数设置）。在本节中，我们固定以下参数。令

$$\delta = c \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{2/3},$$

其中 $c > 0$ 是一个足够小的绝对常数。因此，每当我们写 B_δ时，均指采用此固定的 δ取值时定义 *3.4* 中定义的集合。我们还设定 $t_{\min} = 20 \log n/\varphi^2$（算法 *3* 中的最小随机游走长度）、 $t_\Delta = 20 \cdot \frac{\epsilon}{\varphi^2} \cdot \log n + 2\log(\varphi^2/\epsilon)$（算法 *3* 中不同的游走长度数量），以及 $J = 6 \cdot 10^4$（为草图树中的每个节点 u 计算的独立 Sketch（S_u）向量的数量；该草图树在定义 *4.5* 中定义，并由算法 *5* 计算）。

备注 4.4。引理 *3.6* 和引理 *3.7* 对任意 $\delta > 0$均成立。我们采用上述 δ 取值，以优化最终的误分类率。

本节其余部分安排如下。第 4.1 节介绍我们的最近邻过程 FindNeighbors （算法 7）。接着，第 4.2 节介绍预处理过程Preprocessing（算法 5），并陈述其主要保证。第 4.3 节引入分析所需的若干关键定义。在第 4.4 节中，我们证明 Sketch （算法 3）的正确性。然后，在第 4.5 节中，我们证明最近邻过程 FindNeighbors （算法 7）的正确性。第 4.6 节证明 Preprocessing （算法 5）的正确性。最后，在第 4.7 节中，我们将所有内容整合起来，完成定理 4.1 的证明。

4.1 快速最近邻搜索

我们的 Sketch算法（算法 3）使我们能够高效地判断一个集合 S 是否包含来自 x所在簇的顶点。我们将使用这一基本操作来获得一种快速最近邻搜索算法。

首先，我们提高 Sketch 的成功概率。这通过以下过程IsCovered （算法 6）实现：该过程比较 O（log n）次独立试验，试验对象为 $|\langle$Sketch（x） Sketch（S）$\rangle|$

将 O 替换为 $|\langle \textsc{Sketch}(x)\ \ \textsc{Sketch}(x)\rangle|$ 的 $(\log n)$ 次独立试验，并返回（有偏的）多数票。这样，如果 S 包含来自 x 所属簇的一个顶点， IsCovered 就会以高概率返回 True ；否则， IsCovered 会以高概率返回 False 。

为确保查询时间高效， IsCovered 不会在每次查询时计算 Sketch(S)。相反，它以列表 $\{\textsc{Sk}_j\}_{j\in[J]}$ 作为输入，其中包含先前计算好的 Sketch(S)向量，共有 $J = O(\log n)$个。这样就避免了每次调用 IsCovered 时重新计算 Sketch(S)所带来的计算开销。根据引理 2.11，从头计算Sketch(S)需要时间 $\approx |S|\cdot\sqrt{n/k}$；当 $|S| \approx k$时，该时间为 $\approx \sqrt{nk}$，因此对每次查询来说开销过大。为此，这些草图会在预处理过程中计算一次（算法 5）。因此，查询时 IsCovered 只需计算 Sketch(x)向量并求内积，运行时间为 $\approx \sqrt{n/k}$。

使用过程 IsCovered，算法 FindNeighbors （算法 7）会在集合 S 上执行二分搜索（见第 8–13 行），以确定 S 中哪些元素是 x的近邻，也就是谱嵌入与 f_x的内积较大的元素。为了高效地支持这次二分搜索，我们引入数据结构草图树，它为二叉搜索树中出现的所有子集$S' \subseteq S$ 存储预先计算好的 Sketch(S')向量。这确保了查询时 IsCovered所需的所有 Sketch 向量都已就绪。

定义 4.5 (S 的草图树)。 设 $S = (s_1, \ldots, s_{|S|})$ 是以有序列表表示的集合。 S 的草图树是一棵完全二叉树，其中每个节点 u 都关联一个子集 $S_u \subseteq S$。根节点关联整个集合 S。对于关联子集 S_u的任意内部节点 u，其左子节点 $u.$*left* 关联 S_u 的前半部分，右子节点 $u.$*right* 关联 S_u的后半部分，划分依据是 S的顺序。树的叶节点对应单元素子集 $\{s_i\}$，因此叶节点与 S的元素之间存在一一对应关系。

每个节点 u 存储两个属性： u.标签 $= S_u$和 u.草图，草图由针对子集 S_u预先计算的 $O(\log n)$ 个独立Sketch(S_u) 向量组成。

该树还存储一个指向其根节点的指针 *root*。

为确保运行时间足够短，我们将过程 Find-Neighbors （算法 7）允许访问的叶节点最大数量限制为 $O^*(1)$（见算法 7 第 2 行）。在分析中，我们证明这不会影响误分类率（见引理 4.26）。

算法 6 IsCovered(x, $\{\textsc{Sk}_j\}_{j\in[J]}$)
输入： 顶点 x和 J预计算草图 $\textsc{Sk}_1$, ..., $\textsc{Sk}_1$属于同一个集合 S)。
输出： True 如果 S 包含来自 $C_{i(x)}$的顶点，则为 False

```
c ← 0                                              ▷ 投票数计数器
对于 j = 1 到 J，执行                                  ▷次 Sketch独立试验
    若 |⟨Sketch(x) Sk_j⟩| ≥ 0.5 |⟨Sketch(x) Sketch(x)⟩| 则
        c ← c + 1                                  ▷若 x与 S 的估计相关性较高，则递增
    结束 if
结束 for
若 c ≥ 0.3J 则
    返回 True
否则
    返回 False
结束 if
```

算法 7 FindNeighbors(x, r)
输入： 一个顶点 x 属于 G，以及树 $\mathcal{T}$中的一个节点 r （该树由 S 的草图组成，参见定义 4.5）
输出： S 中属于 $C_{i(x)}$的所有顶点

$N \leftarrow \emptyset$
如果 访问的叶节点数 $> O^*(1)$ **则** ▷ 将访问的叶节点数上限设为 $O^*(1)$
  返回 $\emptyset$
结束 if
▷ 对 S进行二分搜索
如果 r 是 $\mathcal{T}$的叶节点**则**
  返回 r.标签 ▷ 返回与叶节点 r关联的单元素集合
否则
  如果 IsCovered(x, r.左.草图) **则** ▷ 测试左子树是否包含来自 $C_{i(x)}$的顶点
    $N \leftarrow N \cup$ FindNeighbors(x, r.左) ▷递归处理左子树
  结束条件判断
  如果 IsCovered(x, r.右。草图) **则** ▷ 测试右子树是否包含来自 $C_{i(x)}$ 的顶点
    $N \leftarrow N \cup$ FindNeighbors(x, r.右) ▷递归处理右子树
  结束如果
  返回 N
结束 if

我们陈述并证明 IsCovered 的保证（见引理 4.25 和 FindNeighbors （见引理 4.26），详见第 4.5 节

4.2 预处理

在上一节中，我们介绍了 FindNeighbors 算法，它可以找出集合 S中顶点 x 的所有近邻。若给定一组代表顶点 $R = \{y_1, \ldots, y_k\}$，其中每个$i \in [k]$都对应一个 $y_i \in C_i \setminus B_\delta$，以及关于 R 的草图树 $\mathcal{T}$（如定义 4.5 所述），其根节点为 r，那么 FindNeighbors($\cdot$, r)自然就构成了一个聚类预言机。具体来说，给定一个查询顶点 x，我们在关于 R 的草图树 $\mathcal{T}$ 上运行 FindNeighbors(x, r)，并将 x 分配到与返回的代表顶点对应的簇中。这与理想化的查询过程类似，该过程在算法 2 中得到形式化描述。

现在我们将形式化定义代表顶点的概念。

定义 4.6（代表顶点）。给定一个集合 $R \subseteq V$，如果 $y \in C_i$，且 R 中没有其他元素属于 C_i，则称顶点 $y \in R$ 为 R 中簇 C_i 的代表顶点。

$We\ s$该簇 C_i 在 R 中有代表顶点，当 R 包含 C的代表顶点时 $_i$.

理想情况下，我们希望计算一个集合 R，其中包含每个簇的一个代表顶点。然而，我们不知道如何在次线性时间内实现这一目标。因此，我们的预处理算法将返回一个集合 R，并满足一个较宽松的保证：几乎所有顶点都属于在 R中有代表顶点的簇；对于这些顶点，可以使用 FindNeighbors恢复相应的代表顶点。

下述定理将这一保证形式化。

ˆ**定理 4.7 (** Preprocessing的正确性**)**。过程 Preprocessing(k) （算法 *5*）的运行时间为

$$O^*\left(\sqrt{nk} \cdot n^{O\left(\frac{\epsilon}{\varphi^2}\log\frac{1}{\epsilon}\right)} \cdot \left(\frac{1}{\epsilon}\right)^{O\left(\log\frac{\varphi^2}{\epsilon}\right)}\right)$$

并输出一棵基于集合 $R \subseteq V$构造的草图树 $\mathcal{T}$。

在 Preprocessing的内部随机性下，以至少0.999 的概率，存在一个大小为 $B \subseteq V$的集合

$$|B| = O\left(n \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log\left(\frac{\varphi^2}{\epsilon}\right)\right)$$

使得以下条件成立。

对于每个顶点 $x \in V \setminus B$， x 的簇 $C_{i(x)}$ 都有唯一的代表顶点 $y_{i(x)} \in R$，且 FindNeighbors（x，$\mathcal{T}$.根节点）以至少 $1 - n^{-50}$ 的概率返回 $y_{i(x)}$，即，

$$\Pr\big[\textsc{FindNeighbors}(x, \mathcal{T}.\text{root}) = y_{i(x)}\big] \geq 1 - n^{-50},$$

其中，概率是针对 FindNeighbors的内部随机性而言的。

ˆ**注 4.8**。 Preprocessing(k) 返回的集合 R 隐式定义了一个聚类预言机：给定查询顶点 x，我们将 x 分配到与 FindNeighbors(x、$\mathcal{T}$.根节点)返回的代表顶点相对应的簇中。根据定理 *4.7*，该预言机最多会将

$$O\left(\left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log^2\left(\frac{\varphi^2}{\epsilon}\right)\right)$$

比例的顶点分类错误。

我们将在第 4.6 节证明定理 4.7。现在简要介绍 Preprocessing 算法。该算法分为两个主要阶段。

阶段 1：寻找代表顶点的候选项。 在此阶段（算法 5 的第 1–21 行），我们计算一组（可能过大的）候选代表顶点，记为 R_{cand}。我们对随机顶点集合 S 进行采样，并使用 FindNeighbors 在 S 上构建相似性图（第 1–11 行）。这一思路与 [CPS15]类似。不过请注意，在 [CPS15], 中，相似性图基于欧几里得距离这一对称的相似性概念，而我们的相似性概念是不对称的。对于顶点 x，我们向 FindNeighbors($x, \mathcal{T}_S$.根节点) 返回的顶点添加有向边，其中 $\mathcal{T}_S$ 是 S的草图树（根据定义 4.5）。由此得到一个有向图，之后我们将其转换为无向相似性图 H （算法 5 第 11 行）：只有当对应的两条有向边都存在时，才添加一条无向边。这样可以确保鲁棒性。如果 FindNeighbors($x, \mathcal{T}_S$.根节点)对顶点 x执行成功，那么 x 在 H 仅 包含 x的真实近邻，即使 FindNeighbors($y, \mathcal{T}_S$.根节点)对其他某个顶点 $y \in S$的返回结果有误，也依然如此。

得到相似性图 H 后，我们从 H 的每个连通分量中选取一个候选代表顶点（第 12–21 行）。直观来说，同一簇中的顶点应当在相似性图中彼此连通，因此每个连通分量都应对应一个簇。因此，获取代表顶点候选项的一种自然方法是从每个连通分量中选取一个顶点。唯一的问题是，得到的候选代表顶点集合可能过大。例如，采样得到的坏顶点 $x \in S \cap B_\delta$ 可能没有近邻，各自形成一个单例分量，这种情况下，每个坏顶点都会对应一个代表顶点。我们在算法的阶段 2 中处理这个问题。

阶段 2：精炼代表顶点集合。在第 23–35 行中，我们从 R_{cand} 中筛除不必要的代表顶点。我们再采样一组顶点 S_{test}，并对每个 $x \in S_{\text{test}}$ 运行 FindNeighbors(x， $\mathcal{T}_{\text{cand}}$.根节点)（其中 $\mathcal{T}_{\text{cand}}$. 根节点表示定义 4.5 中 R_{cand} 上草图树的根节点）。这使我们能够找出在 R_{cand} 中最常作为最近邻出现的顶点。我们的最终代表顶点集合记为 R，由 R_{cand} 中至少一次作为 S_{test} 的近邻出现的顶点组成（参见算法 5 第 34 行）。这一步背后的直觉如下：即使某个顶点是 $V \setminus B_\delta$ 中的“坏”顶点，它也可能进入 R_{cand}，例如，它没有其他近邻，因此成为一个孤立的单例分量。然而，当我们重新采样一组 S_{test} 时，这类顶点不太可能成为 S_{test} 中任何顶点的最近邻。我们只保留 R_{cand} 中那些在 S_{test} 中某个顶点上被 FindNeighbors 返回的顶点，从而有效地剔除这些不必要的代表顶点。

4.3 良好分散集 S 以及相对于 S 的典型顶点

之前 为了证明 Sketch （引理 2.11）的保证，我们需要先给出几个定义 .

对于 $S \subseteq V$ 和整数 $t \geq 0$，定义

$$p_S^t = \mathbb{E}_{y\sim S}[M^t \mathbb{1}_y] \tag{15}$$

表示从 S 中均匀随机选取的顶点出发进行 t 步行走后所得的分布。

碰撞概率 p_S^t 是衡量集合 S 分布是否均匀的一项指标（见定义 4.9）。这是一个相当自然的概念。为便于理解，考虑 $\epsilon = 0$ 的情形。在此设置下，图 G 由 k 个彼此不相交的扩展图组成，每个扩展图的大小为 $\approx \frac{n}{k}$。如果 S 由来自同一个簇的 k 个顶点组成，那么 $\|p_S^t\|_2^2 \approx \frac{k}{n}$。另一方面，如果 S 由来自不同簇的 k 个顶点组成，那么 $\|p_S^t\|_2^2 \approx \frac{1}{n}$。

我们用 $(p_S^t)^2$ 表示如下向量：

$$(p_S^t)^2(v) \coloneqq (p_S^t(v))^2$$

对所有 $v \in V$，其中 $p_S^t(v)$ 表示 p_S^t 的第 v^{th} 个分量。向量 $(p_S^t)^2$ 在支撑生日悖论的方差计算中发挥着关键作用。它出现在定义 4.13 中；其中我们称一个顶点是典型的，如果它满足方差条件 $\left\langle p_x^t,\ (p_S^l)^2 \right\rangle \leq O^*\left(\frac{1}{n^{3/2}} \cdot \frac{k^{3/2}}{|S|^2}\right)$ 和 $\left\langle (p_x^t)^2,\ p_S^l \right\rangle \leq O^*\left(\frac{1}{n^{3/2}} \cdot \frac{k^{3/2}}{|S|}\right)$。

我们的最近邻搜索过程，即 Sketch、IsCovered 和 FindNeighbors，只有在能够很好地体现簇结构的候选顶点集合 S 上才能表现良好——我们将这些集合称为良好分散集。下面定义 4.9 中给出的条件包括一些基本性质，例如与图 G 中离群顶点的重叠较少（下文条件**(1)** ）、对 p_S^t 的自碰撞概率设置上界（下文条件**(2)** ），以及类似于 S 中顶点的谱嵌入在谱意义上接近恒等式的谱界（条件**(3)**）。

正如我们下文将证明的（见引理 4.12），随机集合 S 以高概率满足这些条件。

定义 4.9 (良好分散集 S)。若大小为 $O(k\log(\varphi^2/\epsilon))$ 的集合 $S \subseteq V$ 满足以下性质，则称其为良好分散集：

(1)（与 B_δ 重叠较少）。S 与 B_δ（根据定义 *3.4*）的交集大小不超过 $|S \cap B_\delta| = O\left(\frac{\epsilon/\varphi^2}{\delta} \cdot k \cdot \log(\varphi^2/\epsilon)\right)$。

(2)（低自碰撞概率）自碰撞概率 p_S^t 满足 $\|p_S^t\|_2^2 = O^*\left(\frac{1}{n} \cdot \frac{k^2}{|S|^2}\right)$ 对所有 $t \in [t_{\min}, t_{\min} + t_\Delta]$ 均成立；也就是说，它与均匀分布的自碰撞概率相当。

(3)（嵌入空间中近似各向同性）$y \in S$ 的谱嵌入 f_y 满足 $\sum_{y\in S} \|f_y\|_2^2 \leq O\left(\frac{k}{n} \cdot k\log(\varphi^2/\epsilon)\right)$。此外，除 $O((\epsilon/\varphi^2)^{1/3}\log(\varphi^2/\epsilon)k)$ 个簇外，均有

$$\mu_i\mu_i^\top \bullet \sum_{y\in S\setminus C_i} f_y f_y^\top \leq 10^{-10} \cdot \|\mu_i\|_2^4$$

and

$$\mu_i\mu_i^\top \bullet \sum_{y\in S} f_y f_y^\top \leq O\left((\varphi^2/\epsilon)^{1/3}\right)\|\mu_i\|_2^4.$$

我们将违反此条件的簇称为相对于 S 的坏簇。注意，此条件与要求 $\sum_{y\in S} f_y f_y^\top$ 在谱意义上近似于 I 非常相似。不过，我们的条件**(3)** 明显更弱（参见注释 *4.10*）。

注记 4.10。如果我们采样一个包含 $\Omega(k\log k)$ 个顶点的集合 S，那么我们可以要求 $\sum_{y\in S} f_y f_y^\top$ 在谱意义上近似于一个适当缩放的单位矩阵。然而，如果这样做，我们的分析会得到误分类率 $(\epsilon/\varphi^2)^{1/3}\operatorname{polylog} k$，而不是 $O\left((\epsilon/\varphi^2)^{1/3}\operatorname{polylog}(\varphi^2/\epsilon)\right)$。不过，这需要 $\epsilon < 1/\log k$，对于较大的 k，它可能小至 $1/\log n$。因此，我们将无法处理常数 ϵ，而这正是本文的核心内容之一。

很容易验证，如果 S 是一个良好分散集，那么它的每个子集 $S' \subseteq S$ 也都是良好分散集。为完整起见，我们在 D 节给出证明。

Lemma 4.11. 若 S 是良好分散集，则其每个子集 $S' \subseteq S$ 也都是良好分散集 *t.*

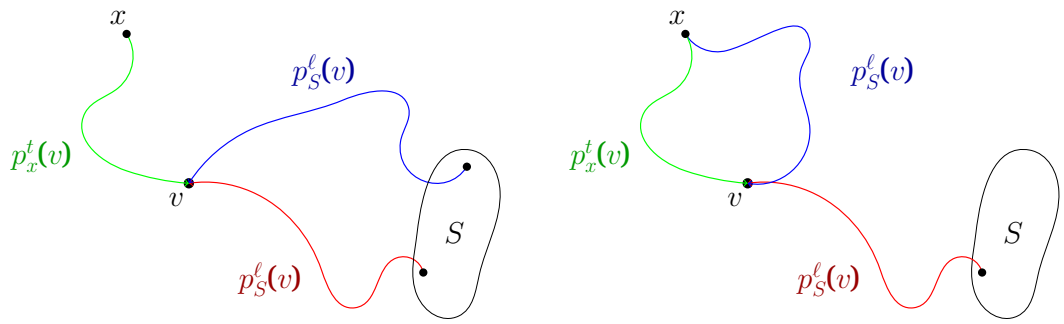

图 1：$\left\langle p_x^t,\ \left(p_S^\ell\right)^2\right\rangle$（左）和 $\left\langle\ \left(p_x^t\right)^2,\ p_S^\ell\right\rangle$（右）的示意图。直观地说，我们可以将 $\left\langle p_x^t,\ \left(p_S^\ell\right)^2\right\rangle$和 $\left\langle\ \left(p_x^t\right)^2,\ p_S^\ell\right\rangle$ 视为从 x 出发的随机游走与从 S 出发的随机游走之间的期望三方碰撞率。

大小与 k 成线性关系的随机集合 S 以高概率是良好分散集：

引理 4.12。如果 S 是一个大小至多为 $O\left(k\log(\varphi^2/\epsilon)\right)$、由独立均匀随机采样得到的顶点（多重）集合，那么该集合 S 以至少 0.9999 的概率是良好分散的。

引理 4.12 可由多次应用马尔可夫不等式得出。证明见附录 D。

定义 4.9 确保了 S 能够很好地表示簇结构。我们的过程 SKETCH、ISCOVERED和 FINDNEIGHBORS还要求，被查询的顶点 x与 S 的交互方式能够反映其与 V的交互。我们在下面的定义 4.13 中将其形式化：我们要求 x 的簇在 S 中包含的代表元不能太多（条件（**2**）），这些代表元不是坏点（条件（**1**）），从 x 出发的随机游走分布与从S 出发的分布之间的相关性不能太高（条件（**3**）和（**4**）），最后，谱嵌入 f_x 偏离簇均值$\mu_{i(x)}$ 的方向，与 S中各顶点的谱嵌入之间不能过于一致。我们将满足这些条件的顶点称为相对于 S 而言的典型顶点（条件（**5**））。

我们将在引理 4.17 中证明，对于良好分散集 S，几乎所有的 $x \in V$ 都满足这些条件。

定义 4.13（相对于 S而言的典型顶点）。给定集合 $S \subseteq V$和顶点 $x \in V$，如果 x 满足以下条件，则称其相对于 S 而言是典型的：

(1)（避开坏点）集合 S 不包含来自 x 的簇中的任何坏点，即

$$S \cap B_\delta \cap C_{i(x)} = \emptyset.$$

(2)（在 S 中未被过度代表）来自 x 的簇、属于 S 的顶点数量有上界

$$|S \cap C_{i(x)}| \le O\left(\left(\frac{\varphi^2}{\epsilon}\right)^{1/3}\right).$$

(3)（在分布 p_S^t 下典型）在 p_S^t 下分配给 x 的概率质量满足

$$p_S^t(x) \le O^*\left(\frac{1}{n}\cdot\frac{k}{|S|}\right)$$

对所有 $t \in [2t_{\min}, 2t_{\min} + 2t_\Delta]$。

(4)（*Lo* 与 p_S^t 的相关性） x 的 t 步分布与 p_S^t 弱相关，即

$$\left\langle p_x^t, \left(p_S^l\right)^2 \right\rangle \le O^*\left(\frac{1}{|S|^2}\cdot\frac{k^{3/2}}{n^{3/2}}\right) \text{ and } \left\langle \left(p_x^t\right)^2, p_S^l \right\rangle \le O^*\left(\frac{1}{|S|}\cdot\frac{k^{3/2}}{n^{3/2}}\right)$$

对所有 $t, l \in [t_{\min}, t_{\min}+t_\Delta]$（关于 t_{min} 和 t_Δ，请参见定义 *4.3*）。直观地说，可以将 $\left\langle p_x^t, \left(p_S^l\right)^2 \right\rangle$ 和 $\left\langle \left(p_x^t\right)^2, p_S^l \right\rangle$ 视为从 x 出发的随机游走与从集合 S 出发的随机游走之间的预期三方碰撞率（示意见图 *1*）。

(5)（不属于坏簇且与均值良好对齐）顶点 x 不属于任何坏簇（根据定义 *4.9*，**(3)**）。此外，它与簇均值的偏差满足

$(f_x - \mu_{i(x)})(f_x - \mu_{i(x)})^\top \bullet \sum_{y\in S} f_y f_y^\top \le 10^{-10}\|f_x\|_2^4.$

注 4.14。 条件 **(5)** 也可以等价地表示为 $\sum_{y\in S}\left\langle f_y, f_x - \mu_{i(x)}\right\rangle^2 \le 10^{-10}\|f_x\|_2^4$。然而，我们证明大多数顶点 x 相对于 S 是典型的（见下文引理 *4.17*）时，会对 $f_x - \mu_{i(x)}$使用平均论证。因此，相较于将这些表达式视为 $f_x - \mu_{i(x)}$的二次函数，将其视为 $f_x - \mu_{i(x)}$二阶张量幂的线性函数更为方便。

此外，我们引入强典型顶点这一概念。强典型顶点是典型顶点，并且还满足（**4**）中的一个更强的上界。虽然这个定义对证明主要结果定理 4.1 没有帮助，但对于定理 1.3 的权衡结果很重要。更多详情请参见 G 节。

定义 4.15（强典型顶点）。 给定一个集合 $S \subseteq V$ 和一个顶点 $x \in V$，若 x 相对于 S 是强典型的，即它相对于 S是典型的，并且还满足（**4**）的更强版本，则称之为：

$$\left\langle p_x^t, \left(p_S^l\right)^2 \right\rangle \le O^*\left(\frac{1}{|S|^2}\cdot\frac{k^2}{n^2}\right) \text{ and } \left\langle \left(p_x^t\right)^2, p_S^l \right\rangle \le O^*\left(\frac{1}{|S|}\cdot\frac{k^2}{n^2}\right)$$

对所有 $t, l \in [t_{\min}, t_{\min}+t_\Delta]$.

注 4.16。 容易验证，若 x 相对于集合 S是典型的（强典型的），则 x 相对于每个子集 $S' \subseteq S$都是典型的（强典型的）。

下面的引理表明，对于任意良好分散集 S，几乎所有 $x \in V$ 相对于S都是典型的。

引理 4.17。 对于每个良好分散集 S，都有

$$|\{x \in V : x \text{ is not } \textit{strongly typical with respect to } S\}| \le O\left(n\cdot\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\cdot\log(\varphi^2/\epsilon)\right).$$

引理 4.17 的表述强于证明定理 4.1 所需的结论。对于主定理，只需界定相对于 S而言并非典型 的顶点数量即可。这个更强的结论界定了并非强典型的顶点数量，用于获得定理 1.3 中的权衡结果。该引理通过多次应用马尔可夫不等式得到。证明见 D 节。

4.4 Sketch 分析（算法 3）

本节证明 SKETCH 算法（算法 3）的正确性保证，下面重新给出该算法，以方便读者阅读。

引理 2.11. 过程 SKETCH （算法 *3*）具有以下性质。其运行时间为

$$O^*\left(|S|\cdot\sqrt{\frac{n}{k}}\cdot n^{O(\epsilon/\varphi^2\log(1/\epsilon))}\cdot(1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right),$$

并输出一个支撑大小至多为

$$O^*\left(|S|\cdot\sqrt{\frac{n}{k}}\cdot n^{O(\epsilon/\varphi^2\log(1/\epsilon))}\cdot(1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right).$$

此外，对于每个集合 $S\subseteq V$，以及每个相对于 S 而言是典型的顶点 $x\in V\setminus B_\delta$（见定义 ***4.13***），以下结论成立。

- 若 $C_{i(x)}\cap S\neq\emptyset$，则

$$\Pr\left[\left|\langle\textsc{Sketch}(x),\textsc{Sketch}(S)\rangle\right|\ \geq\ 0.8\left|\langle\textsc{Sketch}(x),\textsc{Sketch}(x)\rangle\right|\right]\ \geq\ 0.4.$$

- 若 $C_{i(x)}\cap S=\emptyset$，则

$$\Pr\left[\left|\langle\textsc{Sketch}(x),\textsc{Sketch}(S)\rangle\right|\ \leq\ 0.2\left|\langle\textsc{Sketch}(x),\textsc{Sketch}(x)\rangle\right|\right]\ \geq\ 0.999.$$

在两种情况下，概率均是针对 Sketch 的内部随机性而言的，且 Sketch 的所有调用彼此独立。

证明分为两部分：首先，我们证明量 $\left\langle f_x,\ \sum_{y\in S}\sigma_y f_y\right\rangle$ 能够正确判断 S 是否包含来自 x 所属簇的顶点。这由引理 4.19 得出。其次，我们证明 $|\langle\textsc{Sketch}(S)\ \textsc{Sketch}(x)\rangle|$ 能够近似 $\left\langle f_x,\ \sum_{y\in S}\sigma_y f_y\right\rangle$。这由引理 4.22 得出。

我们先处理第一部分。更具体地说，我们证明，以高常数概率（关于 σ 的随机性），当 S 包含来自 $C_{i(x)}$ 的顶点时，内积 $\left|\left\langle f_x,\ \sum_{y\in S}\sigma_y f_y\right\rangle\right|$ 很大；否则则很小。我们需要以下技术引理，它将有助于界定方差 $\mathrm{Var}_\sigma\left[\left\langle f_x,\sum_{y\in S}\sigma_y f_y\right\rangle\right]$。

引理 4.18。 设 $S\subseteq V$ 是一个集合，并假设 x 关于 S 是典型的（参见定义 ***4.13***）。那么

$$f_xf_x^\top\bullet\sum_{y\in S\setminus C_{i(x)}}f_yf_y^\top\leq 10^{-6}\cdot\|f_x\|_2^4,$$

and

$$f_xf_x^\top\bullet\sum_{y\in S}f_yf_y^\top\leq(\varphi^2/\epsilon)^{1/3}\cdot\|f_x\|_2^4.$$

换句话说， f_x 与来自其他簇的谱嵌入几乎没有相关性，而与整个集合 S 的相关性有界。

引理 4.18 的证明由定义 4.13、**(3)** 推出，详见附录 D。

现在我们证明，以高常数概率（关于 σ 的随机性），内积 $\left|\left\langle f_x,\ \sum_{y\in S}\sigma_y f_y\right\rangle\right|$ 在 S 至少包含一个来自 $C_{i(x)}$ 的顶点时较大，否则较小。 ct

引理 4.19。 设 $S\subseteq V$，且 $x\in V\setminus B_\delta$ 关于 S 是典型的（依据定义 ***4.13***）。令$\sigma\sim Unif(\{-1,1\}^S)$。则有如下结论。若 $C_{i(x)}\cap S\neq\emptyset$，则

$$\Pr_\sigma\left[\left|\left\langle f_x,\sum_{y\in S}\sigma_y f_y\right\rangle\right|\geq 0.89\|f_x\|_2^2\right]\geq 0.4999.$$

否则，若 $C_{i(x)}\cap S=\emptyset$，则

$$\Pr_\sigma\left[\left|\left\langle f_x,\sum_{y\in S}\sigma_y f_y\right\rangle\right|\leq 0.1\|f_x\|_2^2\right]\geq 0.9999.$$

证明。 我们首先考虑 $C_{i(x)}\cap S=\emptyset$ 的情况，然后考虑 $C_i(x)\cap S\neq\emptyset$ 的情况。

情况 1： $C_{i(x)} \cap S = \emptyset$。在这种情况下，我们有 $S = S \setminus C_{i(x)}$。我们将证明，对于所有 $x \in V \setminus B_\delta$，都有

$$\Pr_\sigma \left[\left| \left\langle f_x, \sum_{y \in S \setminus C_{i(x)}} \sigma_y f_y \right\rangle \right| \le 0.1 \|f_x\|_2^2 \right] \ge 0.9999.$$

定义随机变量 $X \coloneqq \left\langle f_x, \sum_{y \in S \setminus C_{i(x)}} \sigma_y f_y \right\rangle$ 关于 σ 的随机性。我们有

$$\mathbb{E}[X] = 0$$

and

$$\mathrm{Var}[X] \le \mathbb{E}[X^2] = \sum_{y \in S \setminus C_{i(x)}} \langle f_x, f_y \rangle^2 \le 10^{-6} \|f_x\|_2^4,$$

其中，最后一个不等式由引理 4.18 得出。因此，根据切比雪夫不等式，我们有

$$\Pr_\sigma \left[\left| \left\langle f_x, \sum_{y \in S \setminus C_{i(x)}} \sigma_y f_y \right\rangle \right| > 0.1 \|f_x\|_2^2 \right] = \Pr_\sigma [|X| > 0.1 \|f_x\|_2^2] \le \frac{10^{-6} \|f_x\|_2^4}{0.01 \|f_x\|_2^4} \le 0.0001,$$

由此得出情况 1。

情况 2： $C_{i(x)} \cap S \neq \emptyset$。根据三角不等式，我们有

$$\left| \left\langle f_x, \sum_{y \in S} \sigma_y f_y \right\rangle \right| \ge \left| \left\langle f_x, \sum_{y \in S \cap C_{i(x)}} \sigma_y f_y \right\rangle \right| - \left| \left\langle f_x, \sum_{y \in S \setminus C_{i(x)}} \sigma_y f_y \right\rangle \right|.$$

我们分别对这两项进行界定。首先，我们对第二项给出上界。通过与情况 1 相同的证明，我们有

$$\Pr_\sigma \left[\left| \left\langle f_x, \sum_{y \in S \setminus C_{i(x)}} \sigma_y f_y \right\rangle \right| \le 0.1 \|f_x\|_2^2 \right] \ge 0.9999. \tag{16}$$

这给出了第二项的上界。

现在我们只需对第一项给出下界。首先证明，对于所有 $y \in S \cap C_{i(x)}$，$\langle f_x, f_y \rangle$ 都很大。给定 $y \in C_i(x) \cap S$，有 $y \in C_{i(x)} \setminus B_\delta$。这是因为 x 对于 S 而言是典型的（根据定义 4.13），因此根据定义 4.13、**(1)**，集合 S 不包含来自簇 $C_{i(x)}$ 的坏点，即 $S \cap B_\delta \cap C_{i(x)} = \emptyset$

因此，由于 x 和 y 都属于 $C_{i(x)} \setminus B_\delta$，我们可以将引理 3.7 和注记 3.8 应用于 x 和 y，从而得到

$$\begin{aligned} \langle f_x, f_y \rangle &\ge \left(1 - \frac{4\sqrt{\epsilon}}{\varphi} - 4\sqrt{\delta}\right) \frac{1}{|C_{i(x)}|} \ge \left(1 - \frac{12\sqrt{\epsilon}}{\varphi} - 12\sqrt{\delta}\right) \|f_x\|_2^2, \\ \langle f_x, f_y \rangle &\le \left(1 + \frac{4\sqrt{\epsilon}}{\varphi} + 4\sqrt{\delta}\right) \frac{1}{|C_{i(x)}|} \le \left(1 + \frac{12\sqrt{\epsilon}}{\varphi} + 12\sqrt{\delta}\right) \|f_x\|_2^2, \end{aligned} \tag{17}$$

其中，最后几个不等式由假设 $\delta = O\left((\epsilon/\varphi^2)^{2/3}\right)$ 根据定义 4.3 成立。于是，$\sum_{y \in C_{i(x)} \cap S} \sigma_y \langle f_x, f_y \rangle$ 中的每一项都具有相对较大的绝对值。接下来，我们证明，以足够高的概率，该总和不会相互抵消。直观地说，这是因为只有当 σ_y 个量中恰好一半为正、一半为负时才会发生抵消，而这种情况发生的概率不会太高。形式化地，给定 $\sigma \in \{-1, 1\}^S$，令

$$t_+ \coloneqq |\{y \in S \cap C_{i(x)} : \sigma_y = 1\}|$$

表示分配给 y 在 $S \cap C_{i(x)}$ 中的正号数量，并令

$$t_- \coloneqq |\{y \in S \cap C_{i(x)} : \sigma_y = -1\}|$$

表示 nu 分配给 $S \cap C_{i(x)}$ 中 y 的负号数量。令 $\mathcal{E}_{\text{equal}}$ 为事件，即 $t_+ = t_-$.

命题 4.20。概率$_\sigma$ $[\mathcal{E}_{\text{equal}}] \le 0.5$.

证明由斯特林公式可得，并收录于附录 D。现在证明，在给定事件 $t_+ \neq t_-$ （即 $\mathcal{E}_{equal}$）的条件下，该和不会相互抵消。不失一般性，假设$t_+ > t_-$ （ $t_+ < t_-$ 时的情形类似）。那么

$$\begin{aligned}
\left|\left\langle f_x, \sum_{y\in C_{i(x)}\cap S} \sigma_y f_y \right\rangle\right| &\ge \left\langle f_x, \sum_{y\in C_{i(x)}\cap S} \sigma_y f_y \right\rangle \\
&\ge \left(1 - \frac{12\sqrt{\epsilon}}{\varphi} - 12\sqrt{\delta}\right)\|f_x\|_2^2 \cdot t_+ - \left(1 + \frac{12\sqrt{\epsilon}}{\varphi} + 12\sqrt{\delta}\right)\|f_x\|_2^2 \cdot t_- \\
&= (t_+ - t_-)\cdot\|f_x\|_2^2 - (t_+ + t_-)\left(\frac{12\sqrt{\epsilon}}{\varphi} + 12\sqrt{\delta}\right)\|f_x\|_2^2 \\
&\ge \|f_x\|_2^2 - |S\cap C_{i(x)}|\left(\frac{12\sqrt{\epsilon}}{\varphi} + 12\sqrt{\delta}\right)\|f_x\|_2^2 \\
&\ge 0.99\|f_x\|_2^2.
\end{aligned}$$

这里，第二步变换由公式（17）得出。第三步变换只是将$\|f_x\|_2^2$归为一组，并将 $\left(\frac{12\sqrt{\epsilon}}{\varphi} + 12\sqrt{\delta}\right)\|f_x\|_2^2$项归为一组。第四步变换利用了假设 $t_+ > t_-$，尤其是 $t_+ - t_- \ge 1$，同时也利用了根据定义， t_+ 和 t_-的 $t_+ + t_- = |S\cap C_{i(x)}|$。最后一步变换成立是因为 $|S\cap C_{i(x)}| \le O\left((\varphi^2/\epsilon)^{1/3}\right)$，这是 x 相对于 S 作为典型点的性质（定义 4.13，(2)），再结合引理陈述中的假设 $\delta \le c\cdot(\epsilon/\varphi^2)^{2/3}$。

特别地，我们有

$$\Pr_\sigma\left[\left|\left\langle f_x, \sum_{y\in S\cap C_{i(x)}} \sigma_y f_y \right\rangle\right| \ge 0.99\|f_x\|_2^2\right] \ge 1 - \Pr[\mathcal{E}_{\text{equal}}] \ge 0.5. \tag{18}$$

最后，为了完成 $S\cap C_{i(x)} \neq \emptyset$ 的情形，对方程（16）和公式（18）中的失败事件应用并集界，可得以至少 $1 - 0.5 - 0.0001 = 0.4999$ 的概率，有

$$\left|\left\langle f_x, \sum_{y\in S} \sigma_y f_y \right\rangle\right| \ge \left|\left\langle f_x, \sum_{y\in S\cap C_{i(x)}} \sigma_y f_y \right\rangle\right| - \left|\left\langle f_x, \sum_{y\in S\setminus C_{i(x)}} \sigma_y f_y \right\rangle\right| \ge 0.99\|f_x\|_2^2 - 0.1\|f_x\|_2^2 = 0.89\|f_x\|_2^2,$$

由此即可得到结论。 □

引理 4.19 表明，我们可以利用谱嵌入 $\left\langle f_x, \sum_{y\in S}\sigma_y f_y\right\rangle$ 的内积来检验 S 是否包含来自 x 所属簇的顶点。接下来，我们证明 $\langle$SKETCH(x) SKETCH(S)$\rangle$近似于 $\left\langle f_x, \sum_{y\in S}\sigma_y f_y\right\rangle$。回想一下， SKETCH($S$)（算法 3）输出 $\sum_{y\in S}\sigma_y\sum_t c_t\cdot\widehat{p}_y^t$，其中 $\widehat{p}_y^t$ 表示从 y 出发的 t 步惰性随机游走的经验分布，而 c_t 是定理 2.10 中多项式 $p(x)$ 里 x^t 的系数。下面的引理表明，如果从每个顶点出发运行 $\approx \sqrt{\frac{n}{k}}$ 次惰性随机游走，那么 $\left\langle \widehat{p}_x^{t_1}, \sum_{y\in S}\sigma_y\widehat{p}_y^{t_2}\right\rangle$ 会很好地集中在其均值 $\left\langle M^{t_1}\mathbb{1}_x, \sum_{y\in S}\sigma_y M^{t_2}\mathbb{1}_y\right\rangle$ 附近。

假设我们从 x 出发进行 Q 次长度为 t_1 的惰性随机游走，并从每个 $y\in S$ 出发进行 R 次长度为 t_2 的惰性随机游走。令 $\widehat{p}_x^{t_1}$ 表示从 x 出发的随机游走的经验分布，并且对于每个 $y\in S$，令 $\widehat{p}_y^{t_2}$ 表示从 y 出发的随机游走的经验分布。于是我们有：

引理 4.21 (碰撞计数)。设 β、$\gamma \ge 1$ 为参数，设 $x\in V$、 $S\subseteq V$ 和 t_1、$t_2 \ge 1$。假设以下条件成立：

(1) $p_S^{t_1+t_2}(x) \le \beta\cdot\frac{1}{n}\cdot\frac{k}{|S|}$,

(2) $\left\langle p_x^{t_1}, \left(p_S^{t_2}\right)^2 \right\rangle \le \gamma \cdot \frac{1}{n^2} \cdot \frac{k^2}{|S|^2}$ *and* $\left\langle \left(p_x^{t_1}\right)^2, p_S^{t_2} \right\rangle \le \gamma \cdot \frac{1}{n^2} \cdot \frac{k^2}{|S|}$.

对于每个 $\rho > 0$ *(d* 期望失败概率），$\xi > 0$ （期望精度），$\sigma \in \{-1,1\}^S$ （符号赋值），*if*

$$Q \cdot R \ge \frac{7\beta}{\rho \cdot \xi^2} \cdot \frac{n}{k} \text{ and } Q, R \ge \frac{7(\gamma + \beta^2)}{\rho \cdot \xi^2}$$

那么，

$$\Pr_{random\ walks}\left[\left|\left\langle \widehat{p}_x^{t_1}, \sum_{y\in S} \sigma_y \widehat{p}_y^{t_2} \right\rangle - \left\langle M^{t_1}\mathbb{1}_x, \sum_{y\in S} \sigma_y M^{t_2}\mathbb{1}_y \right\rangle\right| \le \xi \cdot \frac{k}{n}\right] \ge 1-\rho.$$

该证明通过仔细界定随机游走的方差得出，详见 E节。

引理 4.22。假设 $x \in V \setminus B_\delta$ 关于 S 是典型的（依照定义 *4.13*）。设 $\sigma \sim Unif\left(\{-1,1\}^S\right)$。那么存在 $r = O^*\left(\sqrt{\frac{n}{k}} \cdot \left(\frac{1}{\epsilon}\right)^{O(t_\Delta)}\right)$，使得若 $\widehat{p}_x^t = \overline{\text{RandomWalks}}(r,t,x)$，则

$$\Pr_{\substack{\sigma,\\ random\ walks}}\left[\left|\left\langle \sum_t c_t \cdot \widehat{p}_x^t, \sum_{y\in S}\sigma_y \sum_t c_t \cdot \widehat{p}_y^t \right\rangle - \left\langle f_x, \sum_{y\in S}\sigma_y f_y \right\rangle\right| \ge 10^{-2}\|f_x\|_2^2\right] \le 10^{-5}, \tag{19}$$

其中，c_t 是定理 *2.10* 中多项式 $p(x)$ 中 x^t 的系数。

证明。根据三角不等式，我们有

$$\left|\left\langle \sum_t c_t \cdot \widehat{p}_x^t, \sum_{y\in S, t\ge 0} c_t \cdot \sigma_y \widehat{p}_y^t \right\rangle - \left\langle f_x, \sum_{y\in S}\sigma_y f_y \right\rangle\right| \le A_1 + A_2 \tag{20}$$

其中

$$A_1 \coloneqq \left|\left\langle \sum_t c_t \cdot \widehat{p}_x^t, \sum_{y\in S, t\ge 0} c_t \cdot \sigma_y \widehat{p}_y^t \right\rangle - \left\langle p(M)\mathbb{1}_x, \sum_{y\in S}\sigma_y p(M)\mathbb{1}_y \right\rangle\right|$$

and

$$A_2 \coloneqq \left|\left\langle p(M)\mathbb{1}_x, \sum_{y\in S}\sigma_y p(M)\mathbb{1}_y \right\rangle - \left\langle f_x, \sum_{y\in S}\sigma_y f_y \right\rangle\right|.$$

请注意，A_1 反映了从随机游走分布中抽取有限样本所引入的误差，而A_2 反映了多项式 p 在近似谱嵌入时引入的误差。现在我们对 A_1 和 A_2进行界定。

界定 A_1**：**我们有

$$\begin{aligned}
A_1 &= \left|\left\langle \sum_t c_t \cdot \widehat{p}_x^t, \sum_{y\in S, t\ge 0} c_t \cdot \sigma_y \widehat{p}_y^t \right\rangle - \left\langle p(M)\mathbb{1}_x, \sum_{y\in S}\sigma_y p(M)\mathbb{1}_y \right\rangle\right| \\
&= \left|\left\langle \sum_t c_t \cdot \widehat{p}_x^t, \sum_{y\in S, t\ge 0} c_t \cdot \sigma_y \widehat{p}_y^t \right\rangle - \left\langle \sum_t c_t \cdot M^t\mathbb{1}_x, \sum_{y\in S, t\ge 0} c_t \cdot \sigma_y M^t\mathbb{1}_y \right\rangle\right| \\
&\le \sum_{t_1,t_2\in[t_{\min}, t_{\min}+t_\Delta]} |c_{t_1}||c_{t_2}| \left|\left\langle \widehat{p}_x^{t_1}, \sum_{y\in S}\sigma_y \widehat{p}_y^{t_2} \right\rangle - \left\langle M^{t_1}\mathbb{1}_x, \sum_{y\in S}\sigma_y M^{t_2}\mathbb{1}_y \right\rangle\right|,
\end{aligned}$$

其中最后一个不等式由三角不等式得出。

对于每个 $t_1, t_2 \in [t_{\min}, t_{\min} + t_\Delta]$。由于 x 相对于 S 是典型的（根据定义 4.13），我们有 $p_S^{t_1+t_2}(x) \le O^*\left(\frac{1}{n} \cdot \frac{k}{|S|}\right)$，$\left\langle p_x^{t_1}, \left(p_S^{t_2}\right)^2 \right\rangle \le O^*\left(\frac{1}{|S|^2} \cdot \frac{k^{3/2}}{n^{3/2}}\right)$，并且 $\left\langle \left(p_x^{t_1}\right)^2, p_S^{t_2} \right\rangle \le O^*\left(\frac{1}{|S|} \cdot \frac{k^{3/2}}{n^{3/2}}\right)$，由定义 4.13、(3) 和 (4) 得出。因此，我们可以在 $\beta = O^*(1)$、$\gamma = O^*(\sqrt{n/k})$、$\rho = 10^{-6}$ ^{1} $(t_\Delta+1)^2$ $= \Omega^*(1)$ 以及 $\xi = \frac{5 \cdot 10^{-3}}{\eta \cdot |c_{t_1}||c_{t_2}|} (t_\Delta+1)^2 = \Omega^*\left(\frac{1}{1/\epsilon}(\,)^{t_\Delta}\right)$ 的条件下应用引理 4.21。这里，最后一个等式由定理 2.10 中关于 p 的系数界得出。在引理 4.21 中令 $Q, R = \frac{7}{\rho \cdot \xi^2}(\gamma+\beta^2 = \overline{O^*}\left(\sqrt[-]{\frac{n}{k}} \cdot \binom{1}{\epsilon}^{O(t_\Delta)}\right)\right)$，并利用对所有 $x \in V \setminus B_\delta$ 都有 $\|f_x\|_2^2 = \Theta(\frac{k}{n})$ 这一事实（见注记 3.8），我们得到

$$\Pr_{\text{random walks}}\left[\left|\left\langle \widehat{p}_x^{t_1}, \sum_{y\in S} \sigma_y \widehat{p}_y^{t_2} \right\rangle - \left\langle M^{t_1}\mathbb{1}_x, \sum_{y\in S} \sigma_y M^{t_2}\mathbb{1}_y \right\rangle\right| \le \frac{5\cdot 10^{-3}}{|c_{t_1}||c_{t_2}|(t_\Delta+1)^2} \cdot \|f_x\|_2^2\right] \ge 1 - 10^{-6}\frac{1}{(t_\Delta+1)^2}.$$

对 $(t_\Delta+1)^2$ 个可能的 t_1，$t_2 \in [t_{\min}, t_{\min}+t_\Delta]$ 对应用并集界，可得

$$\begin{aligned}
&\Pr_{\text{random walks}}\left[A_1 > 5\cdot 10^{-3}\|f_x\|_2^2\right] \\
&= \Pr_{\text{random walks}}\left[\left|\left\langle \sum_t c_t \cdot \widehat{p}_x^t, \sum_{y\in S, t\ge 0} c_t \cdot \sigma_y \widehat{p}_y^t \right\rangle - \left\langle \sum_t c_t \cdot M^t\mathbb{1}_x, \sum_{y\in S, t\ge 0} c_t \cdot \sigma_y M^t \mathbb{1}_y \right\rangle\right| > 5\cdot 10^{-3}\|f_x\|_2^2\right] \\
&\le 10^{-6}.
\end{aligned} \tag{21}$$

这就完成了对 A_1 的界定。

界定 A_2： 现在我们转而界定 A_2。令 $U\Sigma U^\top$ 表示 M 的特征值分解，其中 Σ 是 M 的特征值构成的对角矩阵，特征值按降序排列。我们有 $\sum_t c_t M^t = p(M) = Up(\Sigma)U^\top$。回顾一下，$U_{[k]}$ 是 U 的子矩阵，由其前 k 列组成，并且对所有 $x \in V$，$f_x = U_{[k]}^\top \mathbb{1}_x$。于是，

$$\left\langle f_x, \sum_{y\in S}\sigma_y f_y \right\rangle = \mathbb{1}_x^\top U_{[k]} U_{[k]}^\top \left(\sum_{y\in S}\sigma_y \mathbb{1}_y\right)$$

令 $\Sigma_{[k]}$ 表示 Σ 的前 k 个元素构成的 $k\times k$ 对角矩阵，其对角元素与这些元素相等；令 $\Sigma_{[-k]}$ 表示一个 $(n-k)\times(n-k)$ 矩阵，其对角元素与 Σ 的后 $n-k$ 个元素相等。此外，记 $U_{[-k]}$ 为由 U 的最后 $n-k$ 列组成的矩阵。于是

$$Up(\Sigma)^2U^\top = U_{[k]}p(\Sigma_{[k]})^2U_{[k]}^\top + U_{[-k]}p(\Sigma_{[-k]})^2U_{[-k]}^\top. \tag{22}$$

因此，我们可以将其改写为 A_2

$$\begin{aligned}
A_2 &= \left|\left\langle p(M)\mathbb{1}_x, p(M)\sum_{y\in S}\sigma_y\mathbb{1}_y\right\rangle - \left\langle f_x, \sum_{y\in S}\sigma_y f_y\right\rangle\right| \\
&= \left|\mathbb{1}_x^\top Up(\Sigma)^2U^\top \sum_{y\in S}\sigma_y\mathbb{1}_y - \mathbb{1}_x^\top U_{[k]}U_{[k]}^\top \sum_{y\in S}\sigma_y\mathbb{1}_y\right| \\
&= \left|\mathbb{1}_x^\top U_{[k]}\left(p(\Sigma_{[k]})^2 - I_k\right)U_{[k]}^\top \sum_{y\in S}\sigma_y\mathbb{1}_y + \mathbb{1}_x^\top U_{[-k]}p(\Sigma_{[-k]})^2U_{[-k]}^\top \sum_{y\in S}\sigma_y\mathbb{1}_y\right|, \qquad \text{by Equation(22)} \\
&\le \left|f_x^\top\left(p(\Sigma_{[k]})^2 - I_k\right)\sum_{y\in S}\sigma_y f_y\right| + \left|\mathbb{1}_x^\top U_{[-k]}p(\Sigma_{[-k]})^2U_{[-k]}^\top \sum_{y\in S}\sigma_y\mathbb{1}_y\right|, \text{ by triangle inequality and } f_x = U_{[k]}^\top\mathbb{1}_x.
\end{aligned} \tag{23}$$

我们先来界定第二项。由于矩阵 $U_{[-k]}p(\Sigma_{[-k]})^2U_{[-k]}^\top$ 是半正定的，因此有

$$\begin{aligned}\left|\mathbb{1}_x^\top U_{[-k]}p(\Sigma_{[-k]})^2U_{[-k]}^\top\sum_{y\in S}\sigma_y\mathbb{1}_y\right| &\le \mathbb{1}_x^\top U_{[-k]}p(\Sigma_{[-k]})^2U_{[-k]}^\top\mathbb{1}_x+\left(\sum_{y\in S}\sigma_y\mathbb{1}_y\right)^\top U_{[-k]}p(\Sigma_{[-k]})^2U_{[-k]}^\top\left(\sum_{y\in S}\sigma_y\mathbb{1}_y\right)\\ &\le \max_{z\in[0,1-\varphi^2/4]}p(z)^2\left\|U_{[-k]}^\top\mathbb{1}_x\right\|_2^2+\max_{z\in[0,1-\varphi^2/4]}p(z)^2\left\|U_{[-k]}^\top\left(\sum_{y\in S}\sigma_y\mathbb{1}_y\right)\right\|_2^2,\end{aligned}$$

其中，最后一行是由底部对角线 $n-k$ 元素 Σ 有界于$1-\varphi^2/4$ 这一事实得出的（参见备注 3.2）。根据定理 2.10，我们有 $\max_{z\in[0,1-\varphi^2/4]}p(z)\le n^{-4}$。因此，我们可以将公式 (23) 中的第二项上界为

$$\begin{aligned}\left|\mathbb{1}_x^\top U_{[-k]}p(\Sigma_{[-k]})^2U_{[-k]}^\top\sum_{y\in S}\sigma_y\mathbb{1}_y\right| &\le \max_{z\in[0,1-\varphi^2/4]}p(z)^2\left\|U_{[-k]}^\top\mathbb{1}_x\right\|_2^2+\max_{z\in[0,1-\varphi^2/4]}p(z)^2\left\|U_{[-k]}^\top\left(\sum_{y\in S}\sigma_y\mathbb{1}_y\right)\right\|_2^2\\ &\le n^{-8}\left\|U_{[-k]}^\top\mathbb{1}_x\right\|_2^2+n^{-8}\left\|\left(U_{[-k]}^\top\sum_{y\in S}\sigma_y\mathbb{1}_y\right)\right\|_2^2\\ &\le n^{-8}\left\|\mathbb{1}_x\right\|_2^2+n^{-8}\left\|\sum_{y\in S}\sigma_y\mathbb{1}_y\right\|_2^2\\ &\le n^{-8}\cdot 1+n^{-8}\cdot k^2\\ &\le 10^{-3}\|f_x\|_2^2,\end{aligned}\tag{24}$$

其中，当 n 充分大时，利用 $1\le k\le n$ 和 $\|f_x\|_2^2=\Omega(k/n)$这一事实，该不等式成立；参见注记 3.8。

剩下的任务是对公式 (23) 中的第一项进行上界估计。我们将利用 σ的随机性。定义随机变量

$$X\coloneqq f_x^\top(p(\Sigma_{[k]})^2-I_k)\sum_{y\in S}\sigma_y f_y$$

关于 σ. 的随机性，我们有

$$\mathbb{E}[X]=0$$

and

$$\begin{aligned}\mathrm{Var}[X]&\le\mathbb{E}[X^2]\\ &=\sum_{y\in S}\left\langle(p(\Sigma_{[k]})^2-I_k)f_x,f_y\right\rangle^2\\ &\le O\left((\epsilon/\varphi^2)^2\right)\sum_{y\in S}\langle f_x,f_y\rangle^2 && \text{by Theorem 2.10}\\ &=O\left((\epsilon/\varphi^2)^2\right)\cdot f_xf_x^\top\bullet\sum_{y\in S}f_yf_y^\top\\ &\le O\left((\epsilon/\varphi^2)^2\cdot(\varphi^2/\epsilon)^{1/3}\right)\cdot\|f_x\|_2^4 && \text{by Lemma 4.18, since } x \text{ is typical with respect to } S\\ &\le 16\cdot 10^{-12}\|f_x\|_2^4 && \text{since } \epsilon/\varphi^2 \text{ is less than a small constant by Remark 4.2.}\end{aligned}$$

因此，根据切比雪夫不等式，我们得到

$$\Pr_\sigma\left[\left|f_x^\top(p(\Sigma_{[k]})^2-I_k)\sum_{y\in S}\sigma_yf_y\right|\ge 4\cdot 10^{-3}\|f_x\|_2^2\right]=\Pr_\sigma\left[|X|\ge 4\cdot 10^{-3}\|f_x\|_2^2)\right]\le 10^{-6}.\tag{25}$$

结合公式 (23)、公式 (24) 和公式 (25)，我们得到

$$\Pr_{\sigma}\left[A_2 \geq 5\cdot 10^{-3}\|f_x\|_2^2\right] = \Pr_{\sigma}\left[\left|\left\langle p(M)\mathbb{1}_x, p(M)\sum_{y\in S}\sigma_y\mathbb{1}_y\right\rangle - \left\langle f_x, \sum_{y\in S}\sigma_y f_y\right\rangle\right| \geq 5\cdot 10^{-3}\|f_x\|_2^2\right] \leq 10^{-6}. \tag{26}$$

现在，只需对公式 (21) 和公式 (26) 应用并集界，即可分别界定公式 (20) 中的两个项，从而得到该引理。

□

注 4.23 (强典型顶点的权衡)。 我们可以如下修改引理 *4.22*，以实现定理 *1.3* 中的权衡。如果将引理 *4.22* 中的假设加强为：假设 x 相对于 S 是强典型 的（根据定义 *4.15*），而不仅仅是典型的，那么我们就可以将引理 *4.22* 的结论加强为：

对于每个 $\delta' \in [0,1]$，存在 $q = O^*\left(\binom{n}{k}^{\delta'}\cdot\binom{1}{\epsilon}^{O(t_\Delta)}\right)$ 和 $r = O^*\left(\binom{n}{k}^{1-\delta'}\cdot\binom{1}{\epsilon}^{O(t_\Delta)}\right)$，使得如果对每个 $y\in S$ 都有 $\widehat{p}_x^t = \textsc{RandomWalks}(q,t,x)$ 和 $\widehat{p}_y^t = \textsc{RandomWalks}(r,t,y)$，则公式 (19) 成立。

这是因为，如果 x 相对于 S是强典型的，那么引理 *4.21* 中的两个条件对于 $\beta = O^*(1)$ 和 $\gamma = O^*(1)$都成立。因此，我们可以将引理 *4.21* 应用于

$$Q = \left(\frac{n}{k}\right)^{\delta'}\cdot\frac{7(\gamma+\beta^2)}{\rho\cdot\xi^2} = O^*\left(\left(\frac{n}{k}\right)^{\delta'}\cdot\left(\frac{1}{\epsilon}\right)^{O(t_\Delta)}\right)$$

and

$$R = \left(\frac{n}{k}\right)^{1-\delta'}\cdot\frac{7(\gamma+\beta^2)}{\rho\cdot\xi^2} = O^*\left(\left(\frac{n}{k}\right)^{1-\delta'}\cdot\left(\frac{1}{\epsilon}\right)^{O(t_\Delta)}\right)$$

在界定 t 引理 *4.22* 的证明中的项 A_1 时。对 A_2 的界及其证明保持不变 .

类似地，我们可以证明，$|\langle \textsc{Sketch}(x)\ \textsc{Sketch}(x)\rangle|$ 近似于 $\|f_x\|_2^2$。

引理 4.24. 对于每个顶点 $x\in V\setminus B_\delta$，都有

$$\Pr\left[\left|\langle \textsc{Sketch}(x), \textsc{Sketch}(x)\rangle - \|f_x\|_2^2\right| \leq 0.01\|f_x\|_2^2\right] \leq 10^{-6}.$$

上述概率是针对 Sketch的内部随机性而言的，并且对 Sketch的两次调用是独立的。

证明与引理 4.22 类似，详见附录 C。现在我们可以证明 Sketch 的主要保证（引理 2.11），为方便读者，现重述如下。

引理 2.11。 过程 Sketch （算法 *3*）具有以下性质。其运行时间为

$$O^*\left(|S|\cdot\sqrt{\frac{n}{k}}\cdot n^{O(\epsilon/\varphi^2\log(1/\epsilon))}\cdot(1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right),$$

并且输出一个支撑大小至多为

$$O^*\left(|S|\cdot\sqrt{\frac{n}{k}}\cdot n^{O(\epsilon/\varphi^2\log(1/\epsilon))}\cdot(1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right).$$

此外，对于每个集合 $S\subseteq V$ 和每个相对于 S 具有典型性的顶点 $x\in V\setminus B_\delta$ （见定义 *4.13*），以下结论成立。

- 如果 $C_{i\ (x)}\cap S\neq\emptyset$，那么

$$\Pr\left[\left|\langle \textsc{Sketch}(x), \textsc{Sketch}(S)\rangle\right| \geq 0.8\left|\langle \textsc{Sketch}(x), \textsc{Sketch}(x)\rangle\right|\right] \geq 0.4.$$

- 如果 $C_{i\ (x)}\cap S=\emptyset$，那么

$$\Pr\left[\left|\langle \textsc{Sketch}(x), \textsc{Sketch}(S)\rangle\right| \leq 0.2\left|\langle \textsc{Sketch}(x), \textsc{Sketch}(x)\rangle\right|\right] \geq 0.999.$$

在这两种情况下，概率均取自 SKETCH 的内部随机性，并且 SKETCH 的所有调用都是独立的。

证明。我们按如下步骤进行。首先，将引理 4.19 和引理 4.22 应用于集合 S，以证明 $|\langle \textsc{Sketch}(S)\ \textsc{Sketch}(x)\rangle|$ 与真实范数 $\|f_x\|_2^2$ 的比值能够正确指示 S是否包含来自 $C_{i(x)}$的顶点。然后，我们利用引理 4.24 论证 $|\langle \textsc{Sketch}(x)\ \textsc{Sketch}(x)\rangle|$ 能够很好地近似 $|f_x|_2^2$。

现在我们将 $|\langle \textsc{Sketch}(S)\ \textsc{Sketch}(x)\rangle|$ 与 $\|f_x\|_2^2$ 联系起来。回顾一下，$\textsc{Sketch}(S)$输出 $\sum_{y\in S}\sigma_y\sum_t c_t\cdot\widehat{p}_y^t$，而 $\textsc{Sketch}(x)$输出 $\sigma_x\sum_t c_t\cdot\widehat{p}_x^t$。

假设 $C_{i(x)}\cap S\neq\emptyset$。那么，通过对引理 4.19 和引理 4.22 中的失败事件应用并集界，以至少 $1-10^{-5}-0.4999\geq 0.49$的概率，有

$$\begin{aligned}
\langle \textsc{Sketch}(x), \textsc{Sketch}(S)\rangle| &= \left|\left\langle \sum_t c_t\cdot\widehat{p}_x^t, \sum_{y\in S}\sigma_y\sum_t c_t\cdot\widehat{p}_y^t\right\rangle\right| \\
&\geq \left|\left\langle f_x, \sum_{y\in S}\sigma_y f_y\right\rangle\right| - \left|\left\langle \sum_t c_t\cdot\widehat{p}_x^t, \sum_{y\in S}\sigma_y\sum_t c_t\cdot\widehat{p}_y^t\right\rangle - \left\langle f_x, \sum_{y\in S}\sigma_y f_y\right\rangle\right| \qquad (27)\\
&\geq 0.89\|f_x\|_2^2 - 10^{-2}\|f_x\|_2^2 \qquad \text{by Lemma 4.19 and Lemma 4.22}\\
&\geq 0.88\|f_x\|_2^2.
\end{aligned}$$

类似地，如果 $C_{i(x)}\cap S=\emptyset$，那么对引理 4.22 和引理 4.19 中的失败事件应用并集界可得，以至少 $1-10^{-5}-0.0001\geq 0.9998$ 的概率，有

$$\begin{aligned}
|\langle \textsc{Sketch}(x), \textsc{Sketch}(S)\rangle| &= \left|\left\langle \sum_t c_t\cdot\widehat{p}_x^t, \sum_{y\in S}\sigma_y\sum_t c_t\cdot\widehat{p}_y^t\right\rangle\right| \\
&\leq \left|\left\langle f_x, \sum_{y\in S}\sigma_y f_y\right\rangle\right| + \left|\left\langle \sum_t c_t\cdot\widehat{p}_x^t, \sum_{y\in S}\sigma_y\sum_t c_t\cdot\widehat{p}_y^t\right\rangle - \left\langle f_x, \sum_{y\in S}\sigma_y f_y\right\rangle\right| \qquad (28)\\
&\leq 0.1\|f_x\|_2^2 + 10^{-2}\|f_x\|_2^2 \qquad \text{by Lemma 4.19 and Lemma 4.22}\\
&\leq 0.11\|f_x\|_2^2.
\end{aligned}$$

最后，根据引理 4.24，以至少 $1-10^{-6}$ 的概率，有

$$0.99\|f_x\|_2^2 \leq |\langle \textsc{Sketch}(x), \textsc{Sketch}(x)\rangle| \leq 1.01\|f_x\|_2^2. \qquad (29)$$

结合公式 (27) 和公式 (29)，应用并集界可知，如果 $C_{i(x)}\cap S\neq\emptyset$，那么以至少 $1-0.51-10^{-6}>0.4$ 的概率，有

$$\begin{aligned}
\langle \textsc{Sketch}(x), \textsc{Sketch}(S)\rangle| &\geq 0.84\|f_x\|_2^2\\
&\geq \frac{0.84}{0.99}|\langle \textsc{Sketch}(x), \textsc{Sketch}(x)\rangle|\\
&> 0.8|\langle \textsc{Sketch}(x), \textsc{Sketch}(x)\rangle|.
\end{aligned}$$

类似地，结合公式 (28) 和公式 (29)，根据并集界，若 $C_{i(x)}\cap S\neq\emptyset$，则以至少 $1-0.0002-10^{-6}>0.999$ 的概率，有

$$\begin{aligned}
\langle \textsc{Sketch}(x), \textsc{Sketch}(S)\rangle| &\leq 0.11\|f_x\|_2^2\\
&\leq \frac{0.11}{1.001}|\langle \textsc{Sketch}(x), \textsc{Sketch}(x)\rangle|\\
&< 0.2|\langle \textsc{Sketch}(x), \textsc{Sketch}(x)\rangle|.
\end{aligned}$$

运行时间和稀疏度： Sketch 的运行时间主要由对 RandomWalks(r, t, x)（算法 4）的 $|S|$ 次调用决定，每次调用的运行时间为 $O^*(r)$，并返回一个支撑集大小至多为 r 的向量 $\widehat{p}_x^t$。因此， Sketch 的运行时间为 $O^*(|S| \cdot r) = O^*\left(|S|\sqrt{\frac{n}{k}} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$。 Sketch 的输出是由 $\overline{t}_\Delta + 1$ 个向量 $\widehat{p}_x^t$ 的线性组合，其中 $|S|$，因此其支撑集

$$O(|S| \cdot (t_\Delta + 1) \cdot r) = O^*\left(|S|\sqrt{\frac{n}{k}} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right).$$

□

4.5 IsCovered（算法 6）和 FindNeighbors（算法 7）的分析

本节将证明 IsCovered 和 FindNeighbors的正确性保证，具体陈述如下。

引理 4.25 （ IsCovered的正确性）。 过程 IsCovered （算法 *6*）的运行时间为

$$O^*\left(\sqrt{\frac{n}{k}} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right).$$

对于每个集合 $S \subseteq V$ 以及每个相对于 S （根据定义 *4.13*）而言是典型的 $x \in V \setminus B_\delta$，以下保证成立：设 $\{\text{Sk}_j\}_{j\in[J]}$ 是 Sketch 在输入 S上独立执行 $J \geq 6 \cdot 10^4$ 次所得到的输出。那么，以至少 $1 - n^{-102}$ 的概率（关于草图的随机性以及 IsCovered的内部随机性），

- 如果 $C_{i(x)} \cap S \neq \emptyset$，那么 IsCovered($x, \{\text{Sk}_j\}_{j\in[J]}$) 返回 True。
- 如果 $C_{i(x)} \cap S = \emptyset$，那么 IsCovered($x, \{\text{Sk}_j\}_{j\in[J]}$) 返回 False。

因此， IsCovered 可以正确判断 S 是否包含来自簇 $C_{i(x)}$的顶点。

引理 4.26 （ FindNeighbors的正确性）。 过程 FindNeighbors （算法 *7*）的运行时间为

$$O^*\left(\sqrt{\frac{n}{k}} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$$

并返回一个大小至多为 $O^*(1)$ 的集合。

对于每个集合 $S \subseteq V$ 以及相对于 S 而言是典型的任意 $x \in V \setminus B_\delta$ （根据定义 *4.13*），均有以下保证：令 r 为 S 的草图树的根节点（见定义 *4.5*）。那么，以至少 $1 - n^{-100}$ 的概率（取自 S 的草图树的随机性以及 FindNeighbors的内部随机性）， FindNeighbors（x、r）的输出恰好为

$$\text{FindNeighbors}(x, r) = S \cap C_{i(x)}.$$

我们首先证明引理 4.25。

引理 *4.25* $x \in V \setminus B_\delta$ S IsCovered x $\{\text{Sk}_j\}_{j\in[J]}$ $J \geq 6 \cdot$ 的证明。假设对于而言是典型的。10^4 n（、，）执行 X_j 的独立试验。定义 j为$|\langle$Sketch的第x次试验指示 $\text{Sk}_j\rangle| \geq 0.5|\langle$Sketch（$x$） Sketch（$x$）$\rangle|$.的事件

假设 $C_{i(x)} \cap S \neq \emptyset$。那么，根据引理 2.11,

$$\Pr[X_j = 1] \geq \Pr\left[|\langle \text{Sketch}(x), \text{Sk}_j\rangle| \geq 0.8\langle \text{Sketch}(x), \text{Sketch}(x)\rangle\right] \geq 0.4$$

其中，概率取自 IsCovered 的内部随机性和 Sk_j 的随机性。因此，根据切尔诺夫界

$$\Pr\left[\sum_{i=j}^{J} X_j < 0.3J\right] \leq e^{-0.002J} \leq n^{-102}$$

其中 最后一个不等式由 J 的选取方式得出。最后，根据该算法的定义

$$\Pr\left[\textsc{IsCovered}(x, \{\textsc{Sk}_j\}_{\{j\in[J]\}}) = \textsc{True}\right] = \Pr\left[\sum_{j=1}^{J} X_j \geq 0.3T\right] \geq 1 - n^{-102}.$$

反之，假设 $C_{i(x)} \cap S = \emptyset$。定义 Y_j 为 Sketch 的第 j 次试验表明$|\langle \textsc{Sketch}(x)\ \textsc{Sk}_j\rangle| < 0.5|\langle \textsc{Sketch}(x)\ \textsc{Sketch}(x)\rangle|$. 这一事件。根据引理 2.11

$$\Pr[Y_j = 1] \geq \Pr\left[|\langle \textsc{Sketch}(x), \textsc{Sk}_j\rangle| \leq 0.2\langle \textsc{Sketch}(x), \textsc{Sketch}(x)\rangle\right] \geq 0.999$$

其中，概率是对 IsCovered的内部随机性和 $\textsc{Sk}_j$的随机性而言的。应用切尔诺夫界可得，

$$\Pr\left[\sum_{j=1}^{J} Y_j \leq 0.7J\right] \leq e^{-0.002J} \leq n^{-102}$$

其中，最后一个不等式由 J的选取方式成立。因此，

$$\Pr\left[\textsc{IsCovered}(x, \{\textsc{Sk}_j\}_{\{j\in[J]\}}) = \textsc{False}\right] = \Pr\left[\sum_{j=1}^{J} Y_j > 0.7J\right] \geq 1 - n^{-102}.$$

运行时间：根据引理 2.11， $\textsc{Sketch}(x)$ 的运行时间为 $O^*\left(\sqrt{\frac{n}{k}} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$。我们调用它 $3J = O^*(1)$ 次，因此调用 Sketch 的总运行时间为 $O^*\left(\sqrt{\frac{n}{k}} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$。此外，根据引理 2.11， $\textsc{Sketch}(x)$ 是一个支撑集大小为 $O^*\left(\sqrt{\frac{n}{k}} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$ 的向量，因此我们可以在 $O(|\text{support}(\textsc{Sketch}(x)|) = O^*\left(\sqrt{\frac{n}{k}} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$ 时间内计算每个内积 $\langle \textsc{Sketch}(x)\ \textsc{Sketch}(x)\rangle$和 $\langle \textsc{Sketch}(x)\ \textsc{Sk}_j\rangle$。我们总共计算 $O^*(1)$ 个内积，因此所有内积计算的总运行时间为 $O^*\left(\sqrt{\frac{n}{k}} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$。

□

现在我们证明 FindNeighbors 的主要正确性保证（引理 4.26）。

引理 4.26 的证明。 假设 x 关于 S 是典型的。那么根据注 4.16，对于每个子集 $S' \subseteq S$， x 也都是典型的；特别地，对于草图树中的每个节点 u， x 关于集合 $u$label 都是典型的。因此，根据引理 4.25，在 FindNeighbors 中对 IsCovered 的每次调用都以至少 $1 - n^{-102}$ 的概率返回正确答案。由于递归树中至多有 $|S| \leq n$ 条不同路径，且每条路径的长度至多为 $\log n$，我们可以对 IsCovered 的所有递归调用应用并集界，从而以至少 $1 - n^{-102} \cdot n\log\ n \geq 1 - n^{-100}$ 的概率，所有递归调用都输出正确答案。

限定在 IsCovered 的所有递归调用都输出正确答案这一事件上。此时，我们只会对对应于集合 $S_u \subseteq S$ 且至少包含 $S \cap C_{i(x)}$ 的一个元素的节点 u 调用 FindNeighbors(x， u)。由于这同样适用于叶节点（对于这些节点， S_u 是单元素集合）， FindNeighbors(x， r) 只能返回 $S \cap C_{i(x)}$ 中的顶点。因此， FindNeighbors(x， r) $\subseteq S \cap C_{i(x)}$。

现在我们论证， FindNeighbors(x， r) 会返回 $S \cap C_{i(x)}$ 中的所有顶点，也就是说， FindNeighbors(x， r) $\supseteq S \cap C_{i(x)}$。由于相对于 S 而言， x 是典型的，根据定义 4.27 (2)，我们有 $|S \cap C_{i(x)}| \leq O^*(1)$，因此访问的叶节点数不会超过最大值。因此，在所有对 IsCovered 的调用均成功的条件下，过程 FindNeighbors(x， r) 会针对二叉搜索树中与集合 $S_u \subseteq$ 相对应的所有节点 u 调用 FindNeighbors(x， u)，其中这些集合至少包含 $S \cap C_{i(x)}$ 的一个元素。特别地， FindNeighbors(x， r) 会输出 $S \cap C_i$ 的所有元素。

因此，以至少 $1 - n^{-100}$ 的概率， FindNeighbors(x， r) 会准确返回集合 $S \cap C_{i(x)}$，符合要求。

运行时间和输出大小：算法的运行时间主要由对 IsCovered. 的调用决定。根据引理 4.25，每次调用 IsCovered 的耗时为 $O^*\left(\sqrt{\frac{n}{k}}\cdot n^{O(\epsilon/\varphi^2\log(1/\epsilon))}\cdot(1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$。由于该算法在草图树中最多访问 $O^*(1)$ 个叶节点，这意味着它总共最多访问 $O^*(1)\cdot\log n = O^*(1)$ 个节点。因此， FindNeighbors 的总运行时间为 $O^*\left(\sqrt{\frac{n}{k}}\cdot n^{O(\epsilon/\varphi^2\log(1/\epsilon))}\cdot(1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$。输出大小的保证显然成立，因为根据算法的构造，该算法在草图树中最多访问 $O^*(1)$ 个叶节点。

□

4.6 预处理分析（算法 5）

I本节证明定理 4.7，现为方便读者将其重述如下 .

ˆ**定理 4.7** （ Preprocessing的正确性）。过程 Preprocessing（k）（算法 *5*）的运行时间为

$$O^*\left(\sqrt{nk}\cdot n^{O\left(\frac{\epsilon}{\varphi^2}\log\frac{1}{\epsilon}\right)}\cdot\left(\frac{1}{\epsilon}\right)^{O\left(\log\frac{\varphi^2}{\epsilon}\right)}\right)$$

并输出一个在集合 $R\subseteq V$ 上构建的草图树 $\mathcal{T}$。

以至少 0.999 的概率，在 Preprocessing 的内部随机性下，存在一个大小为 $B\subseteq V$ 的集合

$$|B| = O\left(n\cdot\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\cdot\log\left(\frac{\varphi^2}{\epsilon}\right)\right)$$

使得以下结论成立。

对于每个顶点 $x\in V\setminus B$， x 的簇 $C_{i(x)}$ 都有唯一的代表顶点 $y_{i(x)}\in R$，并且 FindNeighbors(x,$\mathcal{T}$.根节点) 以至少 $1-n^{-50}$ 的概率返回 $y_{i(x)}$，即，

$$\Pr\left[\textsc{FindNeighbors}(x,\mathcal{T}.\text{root}) = y_{i(x)}\right]\geq 1-n^{-50},$$

其中，概率是针对 FindNeighbors 的内部随机性而言的。

ˆ**注 4.8**。 Preprocessing(k) 返回的集合 R 隐式定义了一个聚类预言机：给定一个查询顶点 x，我们将 x 分配到与 FindNeighbors(x,$\mathcal{T}$.根节点) 返回的代表顶点相对应的簇中。根据定理 *4.7*，该预言机至多会误分类

$$O\left(\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\cdot\log^2\left(\frac{\varphi^2}{\epsilon}\right)\right)$$

比例的顶点。

首先，我们需要定义由 S 良好表示的簇这一概念。这些簇是指被样本 S“发现”的簇，并且其中所有采样顶点相对于 S 都是典型的。形式化地说，

定义 4.27 （由 S 表示良好）。 给定集合 $S\subseteq V$，若簇 C_i 满足以下条件，则称其由 S 良好表示：

(1) $|S\cap C_i|\neq\emptyset$

(2) 每个 $x\in S\cap C_i$ 相对于 S 都是典型的（根据定义 *4.13*）。

对于随机集合 S，几乎所有簇都由 S 良好表示。

引理 4.28。 若 S 是从 V 中独立均匀采样的 $O(k\log(\varphi^2/\epsilon))$ 个（多重）顶点集合，则其成立的概率至少为 0。9998，即

$$|\{i\in[k] : C_i \textit{ is } \text{not } \textit{well-represented by } S\}|\leq O\left(k\cdot\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\cdot\log(\varphi^2/\epsilon)\right).$$

证明可由马尔可夫界推出，详见 D.2节。我们需要过程 PREPROCESSING （算法 5）中计算得到的集合 S 和 R_{cand} 满足以下保证。

引理 4.29. 令 S 为算法 *7* 第 *1* 行采样得到的顶点集合，令 R_{cand} 为算法 *7* 第 *16* 行生成的集合。对于每个 $i \in [k]$ ，若 C_i 由 S 良好代表（根据定义 *4.27*），则以至少 $1 - n^{-99}$ 的概率（关于 PREPROCESSING的内部随机性），集合 R_{cand} 恰好包含来自簇 C_i 的一个 顶点，即 $|R_{\text{cand}} \cap C_i| = 1$。

证明。令 $i \in [k]$ 满足 C_i 由 S 良好代表（根据定义 4.27）。根据定义 4.27，**(1)**,$S \cap C_i \neq \emptyset$。此外，根据定义 4.27，**(2)** 每个 $x \in S \cap C_i$ 相对于 S 都是典型的（根据定义 4.27）。根据定义 4.13，**(1)**，这意味着 $S \cap C_i \cap B_\delta = \emptyset$. 因此，每个 $x \in C_i \cap S$ 都属于$V \setminus B_\delta$。

因此，对于每个 $x \in S \cap C_i$， x 相对于 S 和 $x \notin B_\delta$ 都是典型的。因此，根据引理 4.26，以至少 $1-n^{-100}$ 的概率，算法 5 第 6 行中对 FINDNEIGHBORS$(x, \mathcal{T}_S$.根节点) 的调用恰好返回集合 $S \cap C_i$。特别地，算法 5 第 7 行中添加到 A 的有向边集合恰好是 $\{(x, y) : y \in S \cap C_i\}$。

通过对所有 $x \in S \cap C_i$ 使用并集界，以至少 $1 - n^{-99}$ 的概率，集合 A 包含 $C_i \cap S$ 中顶点对之间的所有边，且不包含从 $S \cap C_i$ 出发的任何其他边。

在此条件下， $S \cap C_i$ 在算法 5 第 11 行计算得到的图 H 中形成一个团，且该团与 H 的其余部分不连通。特别地， $S \cap C_i$ 在 H 中形成一个连通分量。因此，在第 17 至 20 行中，恰好选中簇 C_i 中的一个顶点。所以 R_{cand} 恰好包含来自 C_i. 的一个顶点。 □

引理 4.28 表明，大多数簇都由 S 良好表示（根据定义 4.27），而引理 4.29 表明，如果簇 C_i 由 S 良好表示，那么 R_{cand} 中就包含一个唯一的簇代表元，用于表示 C_i。这些正是我们能够保证其误分类率较低的簇。因此，至关重要的是，代表这些簇的簇代表元应被纳入由 PREPROCESSING（算法 5）生成的精化集合 R 中。下面的引理 4.30 表明，几乎所有由 S 良好表示的簇的簇代表元都会被纳入 R。它还表明，其他被纳入的顶点数量不会太多。

引理 4.30。设 R_{cand} 和 R 分别为算法 *5* 第 *16* 行和第 *34* 行计算得到的集合。对于 PREPROCESSING(k) 的内部随机性，以至少 *0.994* 的ˆ概率，有

(1) $\left|\left\{y \in R_{\text{cand}} : \ C_{i(y)} \text{ 由 } S \text{ 和 } c_y > 0\right\}\right| \geq k - O(k \cdot (\epsilon/\varphi)^{1/3}\log(\varphi^2/\epsilon$良好表示))

(2) $\left|\left\{y \in R_{\text{cand}} : \ C_{i(y)} \text{ 未由 } S \text{ 和 } c_y > 0\right\}\right| \leq (k \cdot (\epsilon/\varphi)^{1/3}\log^2(\varphi^2/\epsilon))$良好表示。

特别地， $\{y \in R : \ C_{i(y)} \text{ 由 } S\} \geq k - O(k \cdot (\epsilon/\varphi)^{1/3}\log(\varphi^2/\epsilon))$良好表示。

作为一个直接推论，引理 4.30 表明， PREPROCESSING 可用作一种近似簇数的算法。

备注 4.31 (近似 k ，耗时 $\sqrt{nk}$)。令 k 为簇的真实数量，令ˆk 为 k 的常数ˆ倍近似，并令 R 为在算法 *5* 的第 *34* 行计算出的集合 PREPROCESSING(k) 。那么，以至少0.994 的概率（关于 PREPROCESSING 的内部随机性），有

$$|k - |R|| \leq O\left(k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \log^2(\varphi^2/\epsilon)\right).$$

因此， PREPROCESSING作为副产品，可在时间 $O^*\left(\sqrt{nk} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$内，计算簇数量 k 的 $poly\left(\frac{\epsilon}{\varphi^2}\right)$ 乘法近似。我们指出，本文开发的方法可用于获得一种算法：假设能够访问下界 k （关于 k），该算法可在时间 $O^*\left(\sqrt{\frac{n}{k}} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$内计算一个。更详细的讨论请参见 *F* 节。

引理 *4.30* 的证明：我们首先证明，第一个条件成立的概率至少为 0.997；然后证明，第二个条件成立的概率也至少为 0.997。由并集界可得引理的主要结论，而“特别地”这一结论则可由**(1)**和 R 的定义直接推出。

证明 (1): 设 $y \in R_{\text{cand}}$ 满足 $C_{i(y)}$ 能由 S 良好表示（按照定义 4.27）。根据引理 4.29，由于 $C_{i(y)}$ 能由 S 良好表示，因此以至少 $1 - n^{-99}$ 的概率，有 $|C_{i(y)} \cap R_{\text{cand}}| = 1$，即 y 是 R_{cand} 中属于 $C_{i(y)}$ 的唯一元素。特别地，如果 S_{test} 至少包含一个来自 $x \in C_{i(y)} \setminus B_\delta$ 的顶点，那么根据引理 4.26，以至少 $1 - n^{-100}$ 的概率，FindNeighbors($x, \mathcal{T}_{\text{cand}}$.根节点) 返回 y，于是 $c_y > 0$。

因此，我们的任务归结为证明

$$|\{i \in [k] : C_i \text{ is well-represented by } S \text{ and } (S_{\text{test}} \cap C_i \setminus B_\delta) \neq \emptyset\}| \geq k - O(k \cdot (\epsilon/\varphi)^{1/3} \log(\varphi^2/\epsilon)),$$

或者等价地，证明

$$|\{i \in [k] : \ C_i \text{未被 } S\} \cup \{i \in [k] : \ S_{\text{test}} \cap C_i \setminus B_\delta = \emptyset\}| \leq O(k \cdot (\epsilon/\varphi)^{1/3} \log(\varphi^2/\epsilon)) \text{良好表示。} \tag{30}$$

根据引理 4.28，以至少 0.9998 的概率，有 $|\{i \in [k] : \ C_i \text{未被 } S\}| \leq O(k \cdot (\epsilon/\varphi)^{1/3} \log(\varphi^2/\epsilon))$ 良好表示。这就给出了公式 (30) 中第一个集合的大小上界。

现在我们来界定公式 (30) 中第二个集合的大小，即 $\{i \in [k] : \ S_{\text{test}} \cap C_i \setminus B_\delta = \emptyset\}$。我们有

$$|\{i \in [k] : S_{\text{test}} \cap C_i \setminus B_\delta = \emptyset\}| \leq |\{i \in [k] : S_{\text{test}} \cap C_i = \emptyset\}| + |S_{\text{test}} \cap B_\delta|.$$

令 α 为第 24 行中 O记号里隐藏的常数，因此 $|S_{\text{test}}| = \alpha k \log(\varphi^2/\epsilon)$。固定一个簇$C_i$。那么，对于每个簇 C_i，$\mathbb{E}[|C_i \cap S_{\text{test}}|] = \frac{|C_i|}{n} \alpha k \log(\varphi^2/\epsilon) \geq \frac{\alpha}{\eta} \log(\varphi^2/\epsilon)$。因此，由切尔诺夫界，我们得到 $\Pr_{S_{\text{test}}}[|S_{\text{test}} \cap C_i| \geq 1] \geq 1 - 00.1\epsilon/\varphi^2$，当 α 充分大时成立。因此，由马尔可夫不等式，以至少 0.999 的概率，有 $|\{i \in [k] : \ C_i \cap S_{\text{test}} = \emptyset\}| \leq O(\epsilon/\varphi^2)$。

最后，根据引理 3.6 以及设定 $\delta = \Omega((\epsilon/\varphi^2)^{2/3})$（依据定义 4.3），我们有 $|B_\delta| = O(n \cdot (\epsilon/\varphi)^{1/3})$，因此由马尔可夫不等式，以至少 0.999 的概率，有 $|S_{\text{test}} \cap B_\delta| \leq O((\epsilon/\varphi)^{1/3} \log(\varphi^2/\epsilon))$。综合以上结果，由并集界可知，以至少 $1 - n^{-99} - n^{-100} - 0.0002 - 0.001 - 0.001 > 0.997$ 的概率，有

$$\begin{aligned}
&|\{i \in [k] : C_i \text{ is } \textit{not} \text{ well-represented by } S\} \cup \{i \in [k] : S_{\text{test}} \cap C_i \setminus B_\delta = \emptyset\}| \\
&\leq |\{i \in [k] : C_i \text{ is } \textit{not} \text{ well-represented by } S\}| + |\{i \in [k] : S_{\text{test}} \cap C_i = \emptyset\}| + |S_{\text{test}} \cap B_\delta| \\
&\leq O(k \cdot (\epsilon/\varphi)^{1/3} \log(\varphi^2/\epsilon)),
\end{aligned}$$

这就得到了所需的公式 (30)。

证明 (2): 首先注意到，以高概率，只有 S_{test} 中可能导致计数 c_y 增加的元素，对于 $\{y \in R_{\text{cand}} : \ C_{i(y)}$，是未被 $S\}$ 良好表示的 $x \in B_\delta$ 或 x，使得 $C_{i(x)}$ 未被 S 良好表示。事实上，假设 $x \in S_{\text{test}} \setminus B_\delta$ 且 $C_{i(x)}$ 被 S 良好表示。那么根据引理 4.26，以至少 $1 - n^{-100}$ 的概率，FindNeighbors($x, \mathcal{T}_{\text{cand}}$.根节点) 恰好返回集合$C_{i(x)} \cap R_{\text{cand}}$，特别是，它会增加元素 y 的计数器 c_y，而该元素的簇被 S 良好表示。因此，由并集界可知，以至少 $1 - |S_{\text{test}}| \cdot n^{-100} \geq 1 - n^{-98}$ 的概率，有

$$\begin{aligned}
&\left|\left\{y \in R_{\text{cand}} : C_{i(y)} \text{ is } \textit{not} \text{ well-represented by } S \text{ and } c_y > 0\right\}\right| \\
&\leq |S_{\text{test}} \cap B_\delta| + |\{x \in S_{\text{test}} : \ C_{i(x)} \text{ is } \textit{not} \text{ well-represented by } S\}|
\end{aligned} \tag{31}$$

现在我们来界定公式（31）右侧的大小。

根据引理 3.6 以及 $\delta = c \cdot (\epsilon/\varphi^2)^{2/3}$ 的设定（根据定义 4.3），我们有 $|B_\delta| = O(n \cdot (\epsilon/\varphi)^{1/3})$，因此根据马尔可夫不等式，以至少 0.999 的概率，我们有 $|S_{\text{test}} \cap B_\delta| \leq O((\epsilon/\varphi)^{1/3} \log(\varphi^2/\epsilon))$。这就界定了公式（31）中的第一项。现在我们来界定第二项。定义

$$\mathcal{F} \coloneqq \{i \in [k] : C_i \text{ is } \textit{not} \text{ well-represented by } S\}.$$

根据引理 4.28，以至少 0.9998 的概率，有

$$|\mathcal{F}| \leq O\left(k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log(\varphi^2/\epsilon)\right). \tag{32}$$

以方程（32）成立这一事件为条件，则来自属于 $\mathcal{F}$ 的簇、位于 S_{test} 中的样本的期望数量为

$$\begin{aligned}\mathbb{E}[|\{x \in S_{\text{test}} : i(x) \in \mathcal{F}\}|] &= \frac{|S_{\text{test}}|}{n}|\{x \in V : i(x) \in \mathcal{F}\}| \\ &\leq \frac{|S_{\text{test}}|}{n}\max_i |C_i| \cdot |\mathcal{F}| \\ &\leq \frac{O(k\log(\varphi^2/\epsilon))}{n} \cdot \eta \cdot \frac{n}{k} \cdot O\left(k \cdot \left(\epsilon/\varphi^2\right)^{1/3} \cdot \log(\varphi^2/\epsilon)\right) \\ &= O\left(k \cdot \left(\epsilon/\varphi^2\right)^{1/3} \log^2(\varphi^2/\epsilon)\right).\end{aligned}$$

因此，根据马尔可夫不等式，以至少 0.999 的概率，有 $|\{x \in S_{\text{test}} : i(x) \in \mathcal{F}\}| \leq O\left(k \cdot \left(\epsilon/\varphi^2\right)^{1/3} \log^2(\varphi^2/\epsilon)\right)$. 成立。综合以上结果，根据并集界，以至少$1 - n^{-98} - 0.001 - 0.0002 - 0.001 \geq 0.997$的概率，有

$$\begin{aligned}\left|\left\{y \in R_{\text{cand}} : C_{i(y)} \text{ 未被很好地表示为 } S \text{ 和 } c_y > 0\right\}\right| &\leq |S_{\text{test}} \cap B_\delta| + |\{x \in S_{\text{test}} : i(x) \in \mathcal{F}|\} \\ &\leq O\left(k \cdot \left(\epsilon/\varphi^2\right)^{1/3} \log^2(\varphi^2/\epsilon)\right).\end{aligned}$$

□

最后，我们已经准备好证明定理 4.7。

定理 *4.7* 的证明。 我们先定义集合 B。令 S 为算法 5 第 1 行中采样的集合。首先，定义

$$\mathcal{F} \coloneqq \{i \in [k] : C_i \text{ is not well-represented by } S\} \cup \{i \in [k] : C_i \cap R = \emptyset\}$$

为这样的簇的集合：它们未能满足条件，原因要么是 R_{cand} 未能为其计算出良好的代表，要么是它们未被 S_{test} 发现。现在，我们将引理陈述中的集合 B 定义为

$$B \coloneqq \{x \in V : i(x) \in \mathcal{F}\} \cup \{x \in V : x \text{ is } \textit{not} \text{ typical with respect to } S\} \cup B_\delta.$$

首先，我们界定 B的大小。然后，我们将证明引理中的保证对所有 $x \in V \setminus B$都成立 .

界定 B的大小： 根据 B的定义，我们有

$$\begin{aligned}|B| \leq |\{x : C_{i(x)} \text{ in } \textit{not} \text{ well-represented by } S\}| + |\{x : C_{i(x)} \cap R = \emptyset\}| \\ + |\{x : x \text{ is } \textit{not} \text{ typical with respect to } S\}| + |B_\delta|.\end{aligned}$$

现在我们分别界定这四项。根据引理 4.28，以至少 0.999 的概率，有$|\{i \in [k] : C_{i(x)} \text{ 在 未被 } S\}| \leq O(k \cdot (\epsilon/\varphi^2)^{1/3}\log(\varphi^2/\epsilon))$良好表示，因此

$$\begin{aligned}|\{x : C_{i(x)} \text{ in } \textit{not} \text{ well-represented by } S\}| &\leq \max_i |C_i| \cdot O(k \cdot (\epsilon/\varphi^2)^{1/3} \log(\varphi^2/\epsilon)) \\ &\leq \eta \cdot \frac{n}{k} \cdot O(k \cdot (\epsilon/\varphi^2)^{1/3} \\ &= O(n \cdot (\epsilon/\varphi^2)^{1/3}).\end{aligned}$$

根据引理 4.30，以至少 0.994 的概率，我们有 $\{y \in R : C_{i(y)} \text{ 由 } S\} \geq k - O(k \cdot (\epsilon/\varphi^2)^{1/3}\log(\varphi^2/\epsilon))$ 良好表示；根据引理 4.29，以至少 $1 - k \cdot n^{-50}$ 的概率，每个良好表示的簇在 R 中至多包含一个顶点。因此，根据并集界，以至少 $1 - k \cdot n^{-50} - 0.006$ 的概率，有 $|\{i \in [k] : |R \cap C_i| \geq 1\}| \geq k - O(k \cdot (\epsilon/\varphi^2)^{1/3}\log(\varphi^2/\epsilon)$，或者等价地，$|\{i \in [k] : R \cap C_i = \emptyset\}| \leq O(k \cdot (\epsilon/\varphi^2)^{1/3}\log(\varphi^2/\epsilon)$。由此，我们得到

$$\begin{aligned}|\{x \in V : R \cap C_{i(x)} = \emptyset\}| &\leq \max_i |C_i| \cdot |\{i \in [k] : R \cap C_i = \emptyset\}| \\ &\leq O(n \cdot (\epsilon/\varphi^2)^{1/3} \log(\varphi^2/\epsilon)).\end{aligned}$$

由于 S 是一组独立均匀随机采样的顶点，根据引理 4.12，以至少 0.9999 的概率， S 是良好分散集（依照定义 4.9）。如果 S 是良好分散集，那么根据引理 4.17，我们有

$$|\{x \in V : x \text{ is } \textit{not} \text{ typical with respect to } S\}| \le O(n \cdot (\epsilon/\varphi^2)^{1/3} \log(\varphi^2/\epsilon)).$$

这里，我们利用这样一个事实：相对于 S 而言，强典型的每个顶点（根据定义 4.15）也是典型的（根据定义 4.13）。

最后，根据引理 3.6，我们有

$$|B_\delta| \le O(n \cdot (\epsilon/\varphi^2)^{1/3}).$$

综上所述，根据并集界，以至少 $1-0.001-k\cdot n^{-50}-0.006-0.0001 \ge 0.992$ 的概率，有

$$\begin{aligned}|B| &\le |\{x : C_{i(x)} \text{ in } \textit{not} \text{ well-represented by } S\}| + |\{x : C_{i(x)} \cap R = \emptyset\}| \\ &\quad + |\{x : x \text{ is } \textit{not} \text{ typical with respect to } S\}| + |B_\delta| \\ &\le O(n \cdot (\epsilon/\varphi^2)^{1/3} \log(\varphi^2\epsilon)).\end{aligned}$$

证明对所有 $x \in V \setminus B$**：** 假设 $x \in V \setminus B$。那么，根据 B 的定义，簇 $C_{i(x)}$ 由 S 良好表示（根据定义 4.27），并且 $R \cap C_{i(x)} \neq \emptyset$。根据引理 4.29，由于 $C_{i(x)}$ 由 S 良好表示，我们有 $|R_{\text{cand}} \cap C_{i(x)}| = 1$。因此，由于 $R \cap C_{i(x)} \neq \emptyset$ 和 $R \subseteq R_{\text{cand}}$，必有 $|R \cap C_{i(x)}| = 1$，也就是说， $C_{i(x)}$ 在 R 中有一个唯一的代表顶点。

根据 B 的定义，我们还可得 $x \notin B_\delta$，并且相对于 S， x 是典型的。特别地，由于 $R \subseteq S$，根据引理 4.11，相对于 R， x 也是典型的。令 $\mathcal{T}$ 表示 R 的草图树（根据定义 4.5）。根据引理 4.26，由于 $x \notin B_\delta$，且相对于 R， x 是典型的，以至少 $1 - n^{-100}$ 的概率，FindNeighbors（x， $\mathcal{T}$.根节点）返回 $R \cap C_{i(x)}$ 的唯一元素，结论得证。

运行时间： 首先， FindNeighbors 在第 2 行计算草图树 $\mathcal{T}_S$。这需要对大小至多为 $|S|$ 的 $O(\log|S|)$ 个集合运行 Sketch。根据引理 2.11，所需时间为 $O(\log|S|) \cdot O^*\left(|S|\sqrt[k]{n} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right) = O^*\left(\sqrt[k]{n}k \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$.

给定草图树 $\mathcal{T}_S$，根据引理 4.26，第 6 行中每次调用 FindNeighbors$(x, \mathcal{T}_S.\text{根节点})$ 的时间为 $O^*\left(\sqrt[k]{n} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$。由于 FindNeighbors 返回的集合大小至多为 $O^*(1)$，因此第 7 行更新集合 A 的时间为 $O^*(1)$。所以，第 5 至 8 行的总运行时间为 $|S| \cdot O^*\left(\sqrt[k]{n} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right) = O^*\left(\sqrt{nk} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$。

在第 11–12 行中，我们定义图 H 并计算其连通分量。 H 的度至多为 $O^*(1)$，顶点数至多为 $|S| = O^*(k)$ 个，边数至多为 $|S| \cdot O^*(1) = O^*(k)$ 条，因此，找出所有连通分量需要 $O^*(k) \le O^*(\sqrt{nk})$ 的时间。在第 17–20 行中，我们只需从每个连通分量中选取一个顶点，总耗时至多为 $O^*(k)$。

与计算 $\mathcal{T}_S$ 类似，在第 21 行计算 R_{cand} 的草图树需要 $O^*\left(\sqrt{nk} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$. 的时间。给定一棵草图树 $\mathcal{T}_{\text{cand}}$，根据引理 4.26，在第 27 行中，每次调用 FindNeighbors$(x, \mathcal{T}_{\text{cand}}.\text{根节点})$ 需要 $O^*\left(\sqrt[k]{n} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$ 的时间。因此，第 26–32 行的总运行时间为 $|S_{\text{test}}| \cdot O^*\left(\sqrt[k]{n} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right) = O^*\left(\sqrt{nk} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$。最后，在第 35 行中，我们计算 R 的草图树。与计算 $\mathcal{T}_S$ 和 $\mathcal{T}_{\text{cand}}$ 类似，这需要 $O^*\left(\sqrt{nk} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$. 的时间。

综合以上各项，整体运行时间为 $O^*\left(\sqrt{nk} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$，如所述。

□

注记 4.32。 由 *5* 生成的集合 R 隐式定义了簇标签。固定一个双射 $l : R \to [|R|]$，并令 $\mathcal{T}$ 为 R 的草图树（根据定义 *4.5*）。查询 x 时，运行 FindNeighbors$(x, \mathcal{T}.\text{root})$，它会返回 R 的一个子集。如果输出只包含一个顶点 y，则用 $l(y)$ 标记 x。

4.7 定理 4.1 的证明

最后 y，我们已准备好证明定理 4.1，现将其重述如下，以方便读者阅读 .

定理 4.1 (定理 1.2 的正式版本)。 对于每个 ϵ、 $\varphi > 0$，以及每个 $k \geq 1$，都存在一个(k, φ, ϵ)-聚类预言机，其误分类率为 $O\left((\epsilon/\varphi^2)^{1/3} \text{对数}^2(\varphi^2/\epsilon)\right)$，并且它具有

- 预处理时间和空间复杂度 $O^*\left(\sqrt{nk} \cdot n^{O(\epsilon/\varphi^2 \text{对数}(\overline{1/\epsilon}))} \cdot (1/\epsilon)^{O(\text{对数}(\varphi^2/\epsilon))}\right)$
- 查询时间 $O^*\left(\sqrt{\frac{n}{k}} \cdot n^{O(\epsilon/\varphi^2 \text{ 对数}(1/\epsilon))} \cdot (1/\epsilon)^{O(\text{对数}(\varphi^2/\epsilon))}\right)$。

证明。 根据定理 4.7，以至少 0.999 的概率， PREPROCESSING(算法 5) 会计算集合 $R \subseteq V$ 的草图树 $\mathcal{T}$，使得以下结论成立：存在一个大小为 $O\left(n \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \text{对数}(\varphi^2/\epsilon)\right)$的集合 B，使得对于每个顶点 $x \in V \setminus B$，x 的簇 $C_{i(x)}$ 都有一个代表顶点 $y_{i(x)} \in R$ （见定义 4.6），并且

$$\Pr\left[\textsc{FindNeighbors}(x, \mathcal{T}.\text{root}) = y_{i(x)}\right] \geq 1 - n^{-50},$$

其中，概率是针对 FINDNEIGHBORS 的内部随机性而言的。

由此可知，过程 FINDNEIGHBORS(·， $\mathcal{T}$.根节点) 定义了一个聚类预言机，其误分类的顶点至多为

$$O\left(n \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log\left(\frac{\varphi^2}{\epsilon}\right)\right)$$

个。根据定理 4.7，预处理时间为 $O^*\left(\sqrt{nk} \cdot n^{O(\epsilon/\varphi^2 \overline{\log(1/\epsilon)})} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$.

我们存储集合 R 的草图树，因此需要为大小至多为$O(k)$ 的集合存储 $O(\log k)$ 个草图。根据引理 2.11，其空间复杂度为 $O^*\left(\sqrt{nk} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$。最后，根据引理 4.26， FINDNEIGHBORS (算法 7) 的运行时间为 $O^*\left(\sqrt{\frac{n}{k}} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$。

随机比特。 请注意，当前版本的算法在每次查询时都使用新的随机性来采样随机游走。尽管该算法以高概率成功，但结果并不一致。通过使用标准技术，我们可以仅使用少量在预处理阶段采样的随机比特，而不会降低数据结构成功的概率。更新随机游走的采样方式后，预处理阶段完成后，我们的聚类预言机将完全确定。

事实上，注意到 FINDNEIGHBORS (算法 7) 的随机性来自使用过程 SKETCH (算法 3) 来近似成对谱点积。在分析 SKETCH 的性能时，我们利用了随机游走是四重独立的这一点（见引理 4.21 的证明），但不需要更高阶的独立性。因此，对于每次查询，我们需要 $W = O^*\left(\sqrt{\frac{n}{k}} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$ 条长度为 $T = O(\log(n)/\varphi^2)$ 的随机游走，我们可以按以下方式生成这些随机游走：

- 生成一个大小为 $s = T \times W$ 的表 H，其中每个条目都独立选取为一个四重独立哈希函数 $h\colon V \to [d]$;
- 将第 w 个（共 W 个）游走的第 t 步（共 T 步）选作第 $t-1$ 个游走的第 w 步的第 $H_{t,\,w}$ 个邻居。

构建上述表所需的随机比特数为

$$O(\log d \cdot T \cdot W) = O^*\left(\sqrt{\frac{n}{k}} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right),$$

因此，所有这些值都可以在预处理阶段进行采样，并存储在次线性数据结构中。因此，查询阶段完全是确定性的。

如前所述，由于所有函数 h 都被选为四重独立， FINDNEIGHBORS 的性能分析仍然成立，因此，对于集合 B 之外的每个顶点，我们的确定性分类器都以概率 $1-n^{-50}$ 正确工作。根据并集界，算法以至少 $1-n^{-49}$ 的概率对 B 之外的所有顶点都正确工作。 □

参考文献

[AA22] Amol Aggarwal 和 Josh Alman。指数函数及高斯核密度估计的最优次数多项式近似。第 37 届计算复杂性会议论文集，2022 年。[Abb17] Emmanuel Abbe。社群检测与随机块模型：近期进展。《机器学习研究杂志》，18:177:1–177:86，2017 年。[ACL08] Reid Andersen、Fan R. K. Chung 和 Kevin J. Lang。使用 PageRank 对有向图进行局部划分。《Internet Mathematics》，5(1):3–22，2008 年。[AGPT16] Reid Andersen、Shayan Oveis Gharan、Yuval Peres 和 Luca Trevisan。使用演化集合实现近乎最优的局部图聚类。《ACM 汇刊》，63(2):15:1–15:31，2016 年。[ALM13] Zeyuan Allen Zhu、Silvio Lattanzi 和 Vahab S. Mirrokni。一种寻找良好连通簇的局部算法。载于第 30 届国际机器学习会议论文集（ICML 2013），美国佐治亚州亚特兰大，2013 年 6 月 16—21 日，第 396–404 页，2013 年。[CKCLL+13] Tsz Chiu Kwok、Lap Chi Lau、Yin Tat Lee、Shayan Oveis Gharan 和 Luca Trevisan。改进的切格不等式：通过高阶谱隙分析谱划分算法。ACM计算理论年度研讨会论文集，2013 年 1 月。[CKK+18] Ashish Chiplunkar、Michael Kapralov、Sanjeev Khanna、Aida Mousavifar 和 Yuval Peres。图的可聚类性测试：算法与下界。载于 2018 年 IEEE 第 59 届计算机科学基础年会（FOCS）论文集，第 497–508 页。IEEE，2018 年。[CPS15] Artur Czumaj、Pan Peng 和 Christian Sohler。图的簇结构测试。载于第 47 届 ACM计算理论年度研讨会论文集，第 723–732 页，2015 年。[CS10] Artur Czumaj 和 Christian Sohler。有界度图的扩张性测试。《组合学、概率与计算》，19(5-6):693–709，2010 年。[GKL+21] Grzegorz Gluch、Michael Kapralov、Silvio Lattanzi、Aida Mousavifar 和 Christian Sohler。次线性时间的谱聚类预言机。载于 2021 年 ACM-SIAM 离散算法研讨会论文集（SODA 2021），线上会议，2021 年 1 月 10—13 日，第 1598–1617 页。SIAM，2021 年。[GR11] Oded Goldreich 和 Dana Ron。有界度图的扩张性测试。载于 Oded Goldreich 编，《复杂性与密码学研究：随机性与计算交互的杂论——与 Lidor Avigad、Mihir Bellare、Zvika Brakerski、Shafi Goldwasser、Shai Halevi、Tali Kaufman、Leonid Levin、Noam Nisan、Dana Ron、Madhu Sudan、Luca Trevisan、Salil Vadhan、Avi Wigderson 和 David Zuckerman 合作》，《计算机科学讲义》第 6650 卷，第 68–75 页。Springer，2011 年。[IM98] Piotr Indyk 和 Rajeev Motwani。近似最近邻：迈向消除维数灾难。载于 Jeffrey Scott Vitter 编，《第 30 届 ACM计算理论研讨会论文集》，美国得克萨斯州达拉斯，1998 年 5 月 23—26 日，第 604–613 页。ACM，1998 年。[JK24] Agastya Vibhuti Jha 和 Akash Kumar。可聚类图上最大割的亚线性时间测试器。载于 Karl Bringmann、Martin Grohe、Gabriele Puppis 和 Ola Svensson 编，《第 51 届自动机、语言与编程国际研讨会（ICALP 2024）》，《莱布尼茨国际信息学论文集》（LIPIcs）第 297 卷，第 91:1–91:17 页，德国达格施图尔，2024 年。Schloss Dagstuhl – 莱布尼茨信息学中心。

[KPS08] Satyen Kale、Yuval Peres 和 C. Seshadhri。扩张图的噪声容忍性与次线性时间扩张图重构。载于第49届IEEE计算机科学基础研讨会（FOCS 2008），2008年10月25日至28日，美国宾夕法尼亚州费城，第719–728页。IEEE计算机学会，2008年。[KPS13] Satyen Kale 和 C. Seshadhri。扩张图的噪声容忍性与次线性时间扩张重构。《SIAM计算杂志》，42(1):305–323，2013年。[KS08] Satyen Kale 和 C. Seshadhri。有界度图的扩张性测试器。载于 Luca Aceto、Ivan Damg˚ard、Leslie Ann Goldberg、Magn´us M. Halld´orsson、Anna Ing´olfsd´ottir 和 Igor Walukiewicz 编，《自动机、语言与编程：第35届国际研讨会，ICALP 2008》，冰岛雷克雅未克，2008年7月7日至11日，论文集，第一部分：专题A：算法、自动机、复杂性与博弈，《计算机科学讲义》第5125卷，第527–538页。Springer，2008年。[LGT14] James R Lee、Shayan Oveis Gharan 和 Luca Trevisan。多路谱划分与高阶Cheeger不等式。《美国计算机协会会刊》（JACM），61(6):37，2014年。[MH02] John C Mason 和 David C Handscomb。切比雪夫多项式。Chapman and Hall/CRC，2002年。[NS10] Asaf Nachmias 和 Asaf Shapira。图的扩张性测试。《信息与计算》，208(4):309–314，2010年。[OA14] Lorenzo Orecchia 和 Zeyuan Allen Zhu。用于局部图聚类的基于流的算法。载于第二十五届ACM-SIAM离散算法研讨会论文集，SODA 2014，美国俄勒冈州波特兰，2014年1月5日至7日，第1267–1286页，2014年。[Pen20] Pan Peng。图簇结构的鲁棒聚类预言机与局部重构器。载于Shuchi Chawla编，《2020年ACM-SIAM离散算法研讨会论文集》，SODA 2020，美国犹他州盐湖城，2020年1月5日至8日，第2953–2972页。SIAM，2020年。[PY23] Pan Peng 和 Yuichi Yoshida。扩张图上最大割、最大2-LIN(q)和唯一标签覆盖的次线性时间算法。载于Nikhil Bansal 和 Viswanath Nagarajan编，《2023年ACM-SIAM离散算法研讨会论文集》，SODA 2023，意大利佛罗伦萨，2023年1月22日至25日，第4936–4965页。SIAM，2023年。[Rob55] Herbert E. Robbins。关于斯特林公式的一则评注。《美国数学月刊》，62:402–405，1955年。[Sin16] Ali Kemal Sinop。如何对子空间进行舍入：一种新的谱聚类算法。载于第二十七届ACM-SIAM离散算法研讨会论文集，SODA 2016，美国弗吉尼亚州阿灵顿，2016年1月10日至12日，第1832–1847页，2016年。[SP23] Ranran Shen 和 Pan Peng。具有改进预处理时间的次线性时间谱聚类预言机。载于Alice Oh、Tristan Naumann、Amir Globerson、Kate Saenko、Moritz Hardt 和 Sergey Levine编，《神经信息处理系统进展》第36卷：神经信息处理系统年度会议2023，NeurIPS 2023，美国路易斯安那州新奥尔良，2023年12月10日至16日，2023年。[ST14] Daniel A. Spielman 和 Shang-Hua Teng。用于预处理和求解对称对角占优线性系统的近线性时间算法。《SIAM矩阵分析与应用杂志》，35(3):835–885，2014年。

定理 2.10 的证明

I本节中，我们证明定理 2.10。为方便读者，现重述如下 :

定理 2.10. 对于满足 $\epsilon/\varphi^2\log(1/\epsilon) \le c_1$ 的每个 $\epsilon, \varphi \ge 0$，若$\log(1/\epsilon)\log(\varphi^2/\epsilon) \le c_2 \log n$，其中绝对常数 c_1, c_2 充分小，并且对于每个 $t \ge 20^{\frac{\log n}{\varphi^2}}$，都存在一个形式如下的多项式 p

$$p(x) = x^t \cdot q(x),$$

其中 q 是一个度为$\deg q = O(\epsilon t + \log(\varphi^2/\epsilon))$ 的多项式，使得 p 的系数的绝对值有界于$(1/\epsilon)^{O(\epsilon t + \log(\varphi^2/\epsilon))}$，并且

- $|p(x) - 1| \le \epsilon/\varphi^2$ *for* $x \in [1-\epsilon, 1]$
- $|p(x)| \le n^{-4}$ *for* $x \in [0, 1-\varphi^2/4]$.

主要思路是将 x^t 乘以一个多项式 q，使其在区间 $[1-\epsilon, 1]$ 上几乎完全抵消。我们通过截断函数 $(1-\epsilon x)^{-t}$ 的切比雪夫展开来构造 q。具体而言，我们证明了定理 A.2，该定理为这种近似提供了必要的保证。

首先，我们正式定义切比雪夫多项式 $Q_d(x)$。

定义 A.1. 给定整数 $d \ge 0$，根据以下关系定义度 d 次首一切比雪夫多项式 $Q_d(x)$

$$Q_d(x) = 2^{1-d}\cos(d\theta) \qquad when\ x = \cos\theta.$$

定理 A.2. 设 $t \ge 2$、 $\epsilon > 0$、 $d \ge 10\epsilon t$，并令

$$P(x) = \sum_{v=0}^{d} 2^{v-1} Q_v(x) c_{v,\epsilon,t}$$

为取切比雪夫展开的前 $d+1$ 项所得到的多项式，其中展开对象为$(1-\epsilon x)^{-t}$，定义在 $[-1,1]$. 上。则

- $\max_{v \le d} |c_{v,\epsilon,t}| \le 2^{O(\epsilon t)}$, *and*
- $\sup_{x\in[-1,1]} |P(x) - (1-\epsilon x)^{-t}| = O(de^{-d})$.

我们先假设定理 A.2 成立，证明定理 2.10，然后在本节余下部分证明定理 A.2。

定理 *2.10* 的证明。设

$$d = 20\epsilon t + 2\log(\varphi^2/\epsilon),$$

并令

$$P(x) \coloneqq \sum_{v=0}^{d} 2^{v-1} Q_v(x) c_{v,\epsilon,t}$$

为取 $(1-\epsilon x)^{-t}$ 在$[-1,1]$. 上的切比雪夫展开的前 $d+1$ 项所得的多项式。令

$$p(x) \coloneqq x^t \cdot P\left(\frac{1-x}{\epsilon}\right).$$

现在证明 p 满足所有所需的性质。

显然，对于， p 的形式为 $x^t q(x)$

$$q(x) \coloneqq P\left(\frac{1-x}{\epsilon}\right),$$

且 $\deg(q) = d = 20\epsilon t + 2\log(\varphi^2/\epsilon)$。要给 p 的系数绝对值设定上界，我们只需给 $q(x)$ 的系数绝对值设定上界。以下命题即可证明这一点:

命题 A.3。 $q(x)$ 的系数绝对值的上界为 $\left(\frac{1}{\epsilon}\right)^{O(d)}$.

为证明命题 A.3，我们将 $c_{v,\epsilon,t}$ 的切比雪夫系数与相对于标准基 $\{1, x, \dots x^d\}$的系数联系起来，然后进行变量替换 $y := \frac{1-x}{\epsilon}$。证明见 A.2 节。由命题 A.3， p 上的最大系数界现可由 $p(x) = x^t q(x)$ 得出。

接下来，我们证明 p的两个既定性质，即 $|p(x)-1| \le \epsilon/\varphi^2$ 对 $x \in [1-\epsilon, 1]$ 成立，以及 $|p(x)| \le n^{-4}$对 $x \in [0, 1-\varphi^2/4]$成立。

界定 $|p(x) - 1|$ 在 $[1-\epsilon, 1]$上的界： 我们将使用定理 A.2 中的界 $\sum_{x\in[-1,1]} |P(x) - (1-\epsilon x)^{-t}| = O(de^{-d})$。更具体地说，我们有

$$\begin{aligned}
\sup_{x\in[1-\epsilon,1]} |p(x)-1| &\le \sup_{x\in[1-\epsilon,1]} \left|x^t\right| \left|P\left(\frac{1-x}{\epsilon}\right) - x^{-t}\right| \\
&\le \sup_{x\in[1-\epsilon,1]} \left|P\left(\frac{1-x}{\epsilon}\right) - x^{-t}\right| \\
&= \sup_{y\in[0,1]} \left|P(y) - (1-\epsilon y)^{-t}\right| \\
&\le \sup_{y\in[-1,1]} \left|P(y) - (1-\epsilon y)^{-t}\right| \\
&\le O(de^{-d}) && \text{由定理 A.2} \\
&\le O\left(\left(\epsilon t + \log(\varphi^2/\epsilon)\right) e^{-(\epsilon t + 2\log(\varphi^2/\epsilon))}\right) \\
&\le \epsilon t \cdot e^{-\epsilon t} \cdot \frac{\epsilon}{\varphi^2} + \log(\varphi^2/\epsilon) \cdot \left(\frac{\epsilon}{\varphi^2}\right)^2 \\
&\le \frac{\epsilon}{\varphi^2}.
\end{aligned}$$

这里最后一步使用了不等式 $xe^{-x} \le 1/2$ 当 $x \in \mathbb{R}$ 且 $\log(x)/x \le 1/2$ 当 $x > 0$。

对 $|p(x)|$ 在 $[0, 1-\varphi^2/4]$ 上进行界定： 我们有

$$\begin{aligned}
\max_{x\in[0,1-\varphi^2/4]} |p(x)| &\le \max_{x\in[0,1-\varphi^2/4]} |x^t| \cdot \max_{x\in[0,1-\varphi^2/4]} |q(x)| \\
&\le \left(1 - \frac{\varphi^2}{4}\right)^t \cdot \deg(q) \cdot \max |\operatorname{coeff}(q)| \\
&\le \left(1 - \frac{\varphi^2}{4}\right)^t \cdot O(\epsilon t + \log(\varphi^2/\epsilon)) \cdot \left(\frac{1}{\epsilon}\right)^{O(\epsilon t + \log(\varphi^2/\epsilon))}.
\end{aligned} \tag{33}$$

接下来，我们注意到上面的最后一个表达式关于 t 递减。为此，我们使用假设$\epsilon/\varphi^2 \log(1/\epsilon) < c_1$，其中 c_1 是一个充分小的常数，从而得到

$$\left(1 - \frac{\varphi^2}{4}\right)\left(\frac{1}{\epsilon}\right)^{O(\epsilon)} = \left(1 - \frac{\varphi^2}{4}\right) e^{O(\epsilon \log(1/\epsilon))} \le \left(1 - \frac{\varphi^2}{4}\right) e^{\varphi^2/4} < 1.$$

因此，表达式 $\left(1 - \frac{\varphi^2}{4}\right)^t \cdot O(\epsilon t + \log(\varphi^2/\epsilon)) \cdot \left(\frac{1}{\epsilon}\right)^{O(\epsilon t + \log(\varphi^2/\epsilon))}$ 关于 t 递减。

因此，通过令 $t = 20 \log n/\varphi^2$，我们可以对公式 (33) 中的表达式给出上界。对于 $x \in [0, 1-\varphi^2/4]$，有

$$|x^t| \le \left(1 - \frac{\varphi^2}{4}\right)^{20\log(n)/\varphi^2} < n^{-5}.$$

令 $t = 20\log\, n/\varphi^2$，并结合公式 (33)，可得

$$\begin{aligned}\max_{x\in[0,1-\varphi^2/4]} |p(x)| &\le \left(1-\frac{\varphi^2}{4}\right)^t \cdot O(\epsilon t + \log(\varphi^2/\epsilon)) \cdot \left(\frac{1}{\epsilon}\right)^{O(\epsilon t+\log(\varphi^2/\epsilon))} \\ &\le n^{-5} \cdot O(\epsilon/\varphi^2 \log n + \log(\varphi^2/\epsilon)) \cdot \left(\frac{1}{\epsilon}\right)^{O(\epsilon/\varphi^2 \log n+\log(\varphi^2/\epsilon))} \\ &\le n^{-5} \cdot \left(2^{O(\epsilon/\varphi^2 \log(1/\epsilon)\log n} + 2^{O(\log(\varphi^2/\epsilon)\cdot\log(1/\epsilon))}\right) \\ &\le n^{-4},\end{aligned}$$

其中，最后一个不等式由假设 $\epsilon/\varphi^2 \log(1/\epsilon) \le c_1$ and $\log(1/\epsilon)\log(\varphi^2/\epsilon) \le c_2 \log\, n$得出，前提是常数 c_1, c_2充分小。因此， p 具备所有所需性质，证明完毕。

□

A.1 在 $[-1,1]$ 上对 $(1-\epsilon x)^{-t}$ 的切比雪夫逼近（定理 A.2 的证明）

本节证明定理 A.2，为方便读者，现将其重述如下：

定理 A.2. 设 $t \ge 2$、 $\epsilon > 0$、 $d \ge 10\epsilon t$，并设

$$P(x) = \sum_{v=0}^{d} 2^{v-1} Q_v(x) c_{v,\epsilon,t}$$

为将 $1-\epsilon x$ 在 $[-1,1]$. 上的切比雪夫展开取前 $d+1$ 项后得到的多项式。那么

- $\max_{v\le d} |c_{v,\epsilon,t}| \le 2^{O(\epsilon t)}$, *and*
- $\sup_{x\in[-1,1]} |P(x) - (1-\epsilon x)^{-t}| = O(de^{-d})$.

众所周知，光滑函数可以通过截断其切比雪夫扩张性而得到良好的近似。下面的引理给出了误差的标准界。该结果已为人所知，可直接由定义 A.1 得出。证明可参见 [AA22] 中的命题 2.2。

引理 A.4。 设 $d \ge 1$ 为整数，且 a_0、 a_1、 $\cdots \in \mathbb{R}$ 满足 $\sum_{j=0}^{\infty} |a_j| < \infty$。设 $f\colon\ [-1,1] \to \mathbb{R}$ 由绝对收敛级数 $f(x) = \sum_{j=0}^{\infty} 2^{j-1} a_j Q_j(x)$定义，并设 $\widehat{f}_d := \sum_{j=0}^{d} 2^{j-1} a_j Q_j(x)$ 为取 f的切比雪夫展开的前 $d+1$ 项所得的多项式。那么

$$\sup_{x\in[-1,1]} |\widehat{f}_d(x) - f(x)| \le \sum_{j=d+1}^{\infty} |a_j|.$$

利用上述事实，我们将证明，在 $[-1,1]$ 上，函数 $(1-\epsilon x)^{-t}$ 可以通过截断切比雪夫展开得到良好的近似。具体来说，我们将证明，$(1-\epsilon x)^{-t}$ 的切比雪夫展开中的系数 $c_{v,\ \epsilon,\ t}$ 按等比数列衰减（当 $v \to \infty$时），从而可以给误差 $\sum_{v=1}^{\infty} |c_{v,\ \epsilon,\ t}|$求得上界。首先，我们需要计算系数 $c_{v,\ \epsilon,\ t}$。下面的引理将帮助我们建立泰勒展开与切比雪夫展开之间的联系。

引理 A.5 ((2.14) in [MH02])。

$$x^n = \sum_{k=0}^{\lfloor n/2\rfloor} 2^{-2k} \binom{n}{k} Q_{n-2k}(x),$$

现在我们明确计算每个系数$c_{v,\ \epsilon,\ t}$。

引理 A.6. 对于每个 $\epsilon > 0$、$t > 0$ 和 $x \in [-1, 1]$，我们有

$$(1 - \epsilon x)^{-t} = \sum_{v=0}^{\infty} 2^{v-1} Q_v(x) c_{v,\epsilon,t},$$

其中

$$c_{v,\epsilon,t} = 2 \sum_{n-v \in 2\mathbb{Z}_{\geq 0}} \epsilon^n 2^{-n} \binom{n-1+t}{n} \binom{n}{\frac{n-v}{2}}.$$

证明。$(1-x)^{-t}$ 的泰勒展开为

$$(1-x)^{-t} = \sum_{n=0}^{\infty} \binom{-t}{n} (-1)^n x^n = \sum_{n=0}^{\infty} \binom{n-1+t}{n} x^n,$$

其中最后一个等式成立，是因为

$$(-1)^n \binom{-t}{n} = (-1)^n \frac{(-t)(-t-1)\cdots(-t-(n-1))}{n!} = \frac{t(t+1)\cdots(t+n-1)}{n!} = \binom{n-1+t}{n}.$$

由此得到

$$(1-\epsilon x)^{-t} = \sum_{n=0}^{\infty} \binom{n-1+t}{n} \epsilon^n x^n.$$

根据引理 A.5，我们有

$$x^n = \sum_{k=0}^{\lfloor n/2 \rfloor} 2^{-2k} \binom{n}{k} Q_{n-2k}(x).$$

代入后，得到

$$\begin{aligned}
(1-\epsilon x)^{-t} &= \sum_{n=0}^{\infty} \binom{n-1+t}{n} \epsilon^n \sum_{k=0}^{\lfloor n/2 \rfloor} 2^{-2k} \binom{n}{k} Q_{n-2k}(x) \\
&= \sum_{v=0}^{\infty} 2^{v-1} Q_v(x) \sum_{n-v \in 2\mathbb{Z}_{\geq 0}} \binom{n-1+t}{n} \epsilon^n 2^{1-n} \binom{n}{\frac{n-v}{2}} \\
&= \sum_{v=0}^{\infty} 2^{v-1} Q_v(x) c_{v,\epsilon,t}.
\end{aligned}$$

□

引理 A.6 给出了系数 $c_{v,\ \epsilon,\ t}$ 的无穷和表达式。为了了解这些系数的行为，我们可以对它们作如下界定。

引理 A.7。我们有

$$c_{v,\epsilon,t} = \Theta\left(\sum_{n-v \in 2\mathbb{Z}_{\geq 0}} \sqrt{\frac{(n+t)}{t((n+1)^2 - v^2)}} \exp(F_{v,\epsilon,t}(n)) \right),$$

其中

$$\begin{aligned}
F_{v,\epsilon,t}(n) \coloneqq & -n \log \frac{1}{\epsilon} - n \log 2 + (n+t-1)\log(n+t-1) \\
& - (t-1)\log(t-1) - \left(\frac{n-v}{2}\right) \log\left(\frac{n-v}{2}\right) - \left(\frac{n+v}{2}\right) \log\left(\frac{n+v}{2}\right).
\end{aligned}$$

证明。我们将使用以下形式的斯特林公式，它由 [Rob5 推导得出 5]:

存在常数 c, $C > 0$, 使得对所有 $n \geq 0$,

$$c(n+1)^{1/2} n^n e^{-n} \leq n! \leq C(n+1)^{1/2} n^n e^{-n}.$$

由引理 A.6 可立即得出该引理，因为

$$c_{v,\epsilon,t} = 2 \sum_{n-v \in 2\mathbb{Z}_{\geq 0}} \epsilon^n 2^{-n} \frac{(n+t-1)!}{(t-1)!(\frac{n-v}{2})!(\frac{n+v}{2})!}.$$

□

我们希望用一个单一表达式而非无穷和来界定系数 $c_{v,\ \epsilon,\ t}$。为此，我们将证明函数 $F_{v,\ \epsilon,\ t}(n)$ 至少以线性速度随 n 衰减，其中 n 足够大。这样，我们就可以用几何级数来给出 $c_{v,\ \epsilon,\ t}$ 的上界。

为简化记号，我们进行变量替换 $t \mapsto t+1$。更具体地说，给定 v, t, ϵ，我们在 $[v, \infty)$ 上定义函数 $F(n)$，使得

$$\begin{aligned} F(n) \coloneqq F_{v,\epsilon,t+1}(n) = &-n \log \frac{1}{\epsilon} - n \log 2 + (n+t) \log(n+t) \\ &- t \log t - \left(\frac{n-v}{2}\right) \log \left(\frac{n-v}{2}\right) - \left(\frac{n+v}{2}\right) \log \left(\frac{n+v}{2}\right) \end{aligned}$$

此外，定义

$$\begin{aligned} n_0 &\coloneqq \frac{t\epsilon^2 + \sqrt{\epsilon^2 t^2 + v^2(1-\epsilon^2)}}{1-\epsilon^2} \\ \Phi_{v,\epsilon,t} &\coloneqq F(n_0) \end{aligned}$$

现在我们分析函数 $F(n)$ 的行为，并证明当 n 足够大时，它至少以线性速度衰减。

引理 A.8. 固定 v、t、ϵ。我们有

(1) F 在 n_0 和 $F(n_0) = \Phi_{v,\ \epsilon,\ t}$ 处取得最大值；***(2)*** 对于 $z \geq 2n_0$，我们有 $F(n_0 + z) \leq F(n_0) - \underline{\text{对数}_4}^{\,2} \cdot z$。

证明。 首先，我们证明 **(1)**，然后证明 **(2)**。

证明 (1)： 函数 F 二阶可微，且

$$F'(n) = -\log(1/\epsilon) - \log(2) + \log(n+t) - \frac{1}{2}\log\left(\frac{n-v}{2}\right) - \frac{1}{2}\log\left(\frac{n+v}{2}\right)$$

and

$$F''(n) = -\frac{nt+v^2}{(n+t)(n^2-v^2)} < 0,$$

其中，最后一个不等式成立是因为 $n \geq v$。因此， F 在其整个定义域上是凹函数。方程 $F'(n_0) = 0$ 有两个解

$$n = \frac{t\epsilon^2 \pm \sqrt{\epsilon^2 t^2 + v^2(1-\epsilon^2)}}{1-\epsilon^2},$$

其中只有正解 $n = n_0 = {}^{t\epsilon^2+\sqrt{\epsilon^2t^2+v^2}}_{(1-\epsilon^2)\ 1-\epsilon^2}$ 位于 $\overline{F \text{ 的定义}}$域内。因此， F 在 n_0 处取得最大值。根据 $\Phi_{v,\epsilon,t}$ 的定义，我们有 $F(n_0) = \Phi_{v,\epsilon,t}$。

证明 (2)： 首先，我们需要

断言 A.9. $F'(2n_0) \le -\frac{\log 2}{2}$——

证明. 我们有

$$\begin{aligned} F'(n) &= -\log(2/\epsilon) + \log(n+t) - \frac{1}{2}\log\left(\frac{n-v}{2}\right) - \frac{1}{2}\log\left(\frac{n+v}{2}\right) \\ &= \log\left(\frac{\epsilon(n+t)}{\sqrt{n^2-v^2}}\right) = \frac{1}{2}\log\left(\frac{\epsilon^2(n+t)^2}{n^2-v^2}\right), \end{aligned}$$

so

$$F'(2n_0) = \frac{1}{2}\log\left(\frac{\epsilon^2(2n_0^2+t)^2}{4n_0^2-v^2}\right).$$

因此，只需证明

$$\frac{\epsilon^2(2n_0^2+t)^2}{4n_0^2-v^2} \le \frac{1}{2}.$$

注意， $n_0 \ge v, t\epsilon_\circ$ 因此，对于 $\epsilon \le 1/10$，我们有

$$4n_0^2 - v^2 \ge 3n_0^2 \ge 2\epsilon^2 t^2 + n_0^2 \ge 2\epsilon^2 t^2 + \epsilon n_0 \cdot 10 n_0 \ge 2\epsilon^2 t^2 + \epsilon n_0(8\epsilon n_0 + 8\epsilon t) = 2\epsilon^2(2n_0^2+t)^2,$$

如所要求。 □

利用上述论断，对于 $z \ge n_0$，我们有

$$\begin{aligned} F(n_0+z) - F(n_0) &\le F(n_0+z) - F(2n_0) && \text{by maximality of } F(n_0) \\ &\le (z-n_0)F'(2n_0) && \text{since } F''(n) > 0 \text{ for all } n \ge v \\ &\le -\frac{\log 2}{2}(z-n_0) && \text{since } F(2n_0) \le -\frac{\log 2}{2} \\ &\le -\frac{\log 2}{4} z && \text{using } n_0 \le z/2, \end{aligned}$$

这正是我们需要证明的结论。 □

现在我们可以用 $\Phi_{v,\ \epsilon,\ t}$ 给出系数 $c_{v,\ \epsilon,\ t}$ 的上界。

引理 A.10. 对于所有 v，都有

$$c_{v,\epsilon,t} = O\left((\epsilon t + v)\exp(\Phi_{v,\epsilon,t})\right)$$

证明。 回顾引理 A.7，可得

$$c_{v,\epsilon,t} = \Theta\left(\sum_{n-v \in 2\mathbb{Z}_{\ge 0}} \sqrt{\frac{(n+t)}{t((n+1)^2-v^2)}}\exp(F_{v,\epsilon,t}(n))\right),$$

注意到，对于所有 $n \ge v$，我们都有

$$\sqrt{\frac{(n+t)}{t((n+1)^2-v^2)}} \le \sqrt{\frac{(n/t+1)}{(n^2+2n-1)}} \le \sqrt{\frac{n}{2n}} < 1. \tag{34}$$

因此我们有

$$\begin{aligned}
c_{v,\epsilon,t} &\le C\sum_{n\ge v}^{\infty}\exp F(n) && \text{by Equation (34) and Lemma A.7}\\
&= C\sum_{n=v}^{3n_0-1}\exp F(n_0) + C\sum_{n\ge 3n_0}\exp F(n)\\
&\le C\cdot 3n_0F(n_0) + C\sum_{z\ge 2n_0}\exp(F(n_0)-cz) && \text{by Lemma A.8}\\
&= C\cdot 3n_0F(n_0) + CF(n_0)\sum_{z\ge 2n_0}\exp(-\log 2z/4)\\
&\le C\cdot 3n_0F(n_0) + C\cdot 10F(n_0)\\
&= O((\epsilon t+v)\exp(\Phi_{v,\epsilon,t})) && \text{since } n_0\le 2(\epsilon t+v) \text{ and since } F(n_0)=\Phi_{v,\epsilon,t}
\end{aligned}$$

□

接下来，为了给系数 $c_{v,\epsilon,t}$确定上界，我们需要了解 $\Phi_{v,\epsilon,t}$ 作为 v的函数时的行为。以下引理表明，对于足够大的 v值，$\Phi_{v,\epsilon,t}$ 可以用一个至少随 v 线性衰减的函数来给出上界。这反过来有助于用几何级数给出切比雪夫展开误差的上界。

引理 A.11. 若 $v\ge 10\epsilon t$，则 $\Phi_{v,\epsilon,t}\le -v$

证明。给定任意固定的 v,t,ϵ，我们可以将函数 $F(n)$ 重写为

$$\begin{aligned}
F(n) &= -n\log\frac{2}{\epsilon} + (n+t)\log(n+t) - t\log t - \left(\frac{n+v}{2}\right)\log\left(\frac{n+v}{2}\right) - \left(\frac{n-v}{2}\right)\log\left(\frac{n-v}{2}\right)\\
&= -n\log\frac{2}{\epsilon} + n\log(n+t) + t\log(1+n/t) - n\log\left(\frac{n+v}{2}\right) + \left(\frac{n-v}{2}\right)\log\left(\frac{n+v}{2}\right) - \left(\frac{n-v}{2}\right)\log\left(\frac{n-v}{2}\right)\\
&= -n\log\frac{2}{\epsilon} + n\log\left(\frac{2(n+t)}{n+v}\right) + t\log(1+n/t) + \left(\frac{n-v}{2}\right)\log\left(\frac{n+v}{n-v}\right)
\end{aligned}$$

我们将证明，对于 $n=n_0=\frac{t\epsilon^2+\sqrt{\epsilon^2t^2+v^2-(1-\epsilon^2)1-\epsilon^2}}{}$，这三项 $n\log\left(\frac{2(n+t)}{n+v}\right)$, $t\log(1+n/t)$,$\left(\frac{n-v}{2}\right)\log\left(\frac{n+v}{n-v}\right)$ 都可以被 $-n\log\frac{2}{\epsilon}$项抵消。(35)

首先，注意到我们有

$$v < n_0 \le 2v$$

（下界显然成立，而上界成立是因为 $v\ge 10\epsilon t$ 和 ϵ 足够小）。现在我们有

$$\begin{aligned}
n_0\log\left(\frac{2(n_0+t)}{n_0+v}\right) &\le n_0\log\left(\frac{2(n_0+t)}{n_0+n_0/2}\right)\\
&= n_0\left(\log 4/3 + \log(1+t/n_0)\right) \qquad (36)\\
&\le n_0\left(\log 4/3 + \log\left(1+\frac{1}{10\epsilon}\right)\right) \quad \text{since } n_0\ge v\ge 10\epsilon t
\end{aligned}$$

接下来，应用不等式 $\frac{1}{x}\log(1+x)\le 1$，其中 $x=n_0/t$，得到

$$t\log(1+n_0/t)\le n_0. \qquad (37)$$

最后，我们对 $\left(\frac{n-v}{2}\right)\log\left(\frac{n+v}{n-v}\right)$项进行界定。给定 $y>0$，函数 $f(x)=x\log(y/x)$ 在 $x=y/e$ 处取最大值，因此对于所有 $x,\ y>0$，有

$$x\log(y/x)\le\frac{y}{e}\log(e)=\frac{y}{e}.$$

将此结果应用于 $x = n - v$ 和 $y = n + v$，得到

$$\frac{n_0 - v}{2} \log\left(\frac{n_0 + v}{n_0 - v}\right) \leq \frac{(n_0 + v)}{2e} \leq \frac{n_0}{e}. \tag{38}$$

在将所有内容整合起来之前，注意到

$$\log 4/3 + \log\left(1 + \frac{1}{10\epsilon}\right) + \frac{1}{e} + 2 \leq \log\left(\frac{2}{\epsilon}\right). \tag{39}$$

为说明这一点，注意到

$$\frac{4}{3}\frac{e^{1/e+2}}{10} \approx 1.423 < 2,$$

因此，当 ϵ足够小时，我们得到

$$\frac{4}{3} \cdot \left(1 + \frac{1}{10\epsilon}\right) e^{1/e+2} \leq \frac{2}{\epsilon}$$

对两边取对数，得到方程（39）。

现在，我们可以结合方程（35）、（36）、（37）、（38）和（39），得到

$$\Phi_{v,\epsilon,t} = F(n_0) \leq -n_0 \log\frac{2}{\epsilon} + n_0 \log 4/3 + n_0 \log\left(1 + \frac{1}{10}\right) + \frac{n_0}{e} + n_0 \leq -n_0 \leq -v,$$

这正是所需的。 □

接下来需要控制较小 v 值时的 $\Phi_{v,\epsilon,t}$ 。这由下面的引理得出。

引理 A.12. 若 $v \leq 10\epsilon t$，则 $\Phi_{v,\epsilon,t} \leq 100\epsilon t$

证明。 回顾

$$F(n) = -n \log\frac{2}{\epsilon} + n \log\left(\frac{2(n+t)}{n+v}\right) + t \log(1 + n/t) + \left(\frac{n-v}{2}\right) \log\left(\frac{n+v}{n-v}\right)$$

and

$$n_0 = \frac{t\epsilon^2 + \sqrt{\epsilon^2 t^2 + v^2(1-\epsilon^2)}}{1 - \epsilon^2}.$$

注意

$$v, t\epsilon < n_0 \leq 20\epsilon t$$

对于 $0 \leq v \leq 10\epsilon t$。现在我们将分别界定 $F(n)$ 中的每一项。要界定 $t\log(1 + n/t)$ 项：如公式（37）所示，我们有

$$t \log(1 + n/t) \leq n_0$$

要界定 $\binom{n-v}{2} \log \binom{n+v}{n-v}$ 项：如公式（38）所示，我们有

$$\frac{n_0 - v}{2} \log\left(\frac{n_0 + v}{n_0 - v}\right) \leq \frac{(n_0 + v)}{2e} \leq \frac{n_0}{e} < n_0.$$

为了界定 n_0 对数 $\binom{2(n_0+t)}{n_0+v}$ 项：我们有

$$n_0 \log\left(\frac{2(n_0+t)}{n_0+v}\right) \leq n_0 \log\left(\frac{2(n_0+t)}{n_0}\right) = n_0 \log(2(1+t/n_0)) \leq n_0 \log(2(1+1/\epsilon)) \leq n_0 \log(4) + n_0 \log(1/\epsilon).$$

综合起来，我们得到

$$F(n_0) \leq -n_0 \log(2/\epsilon) + n_0 \log(4) + n_0 \log(1/\epsilon) + n_0 + n_0 \leq 3n_0 \leq 60\epsilon t.$$

□

现在我们可以证明定理 A.2。

定理 *A.2* 的证明。第一个要点可直接由引理 A.10、引理 A.11 和引理 A.12 得出，同时注意到所有系数均为非负数。

为证明第二个要点，根据引理 A.4，我们有

$$\sup_{x\in[-1,1]} |P(x) - (1-\epsilon x)^{t+1}| \le \sum_{v=d}^{\infty} |c_{v,\epsilon,t}|.$$

利用这一点，我们得到

$$\begin{aligned}
\sup_{x\in[-1,1]} |P(x) - (1-\epsilon x)^{t+1}| &\le \sum_{v=d}^{\infty} |c_{v,\epsilon,t}| \\
&\le C_1 \sum_{v\ge d} (\epsilon t + v)\exp(\Phi_{v,\epsilon,t}) && \text{by Lemma A.10} \\
&\le C_2 \sum_{v\ge d} v\exp(-v) && \text{by Lemma A.11, since } d \ge 10\epsilon t \\
&\le C_3 d e^{-d},
\end{aligned}$$

其中 C_1、C_2 和 C_3 是某些普适常数。 □

A.2 多项式基与定义域的变换（命题 A.3 的证明）

本节证明命题 A.3，为方便读者，现将其重述如下。

命题 A.3。 $q(x)$ 的系数绝对值有上界 $\left(\frac{1}{\epsilon}\right)^{O(d)}$ – .

首先，我们需要建立切比雪夫基下的系数与标准基 $\{1, x, \ldots, x^d\}$ 下的系数之间的关系。下面的引理给出了这种关系。

引理 A.13。 对于 $l \ge 0$，令 $C(l)$ 表示 $2^{l-1}Q_l(x)$ 在标准基 $\{1, x, \ldots x^n\}$ 下的系数的最大绝对值。那么

$$C(l) \le (1+\sqrt{2})^l.$$

证明。 注意到，多项式 $T_l(x) = 2^{l-1}Q_l(x)$ 满足以下递推关系

$$T_{l+1}(x) = 2xT_l(x) + T_{l-1}(x),$$

这意味着

$$C(l+1) \le 2C(l) + C(l-1).$$

因此，要给 $C(l)$ 求上界，只需解递推关系

$$\widetilde{C}(l+1) = 2\widetilde{C}(l) + \widetilde{C}(l-1), \widetilde{C}(0) = 1, \widetilde{C}(1) = 1.$$

这里的初 始条件由以下事实得出： $T_0(x) = 1$ 且 $T_1(x) = x$。很容易看出 t

$$C(l) = 0.5\left((1-\sqrt{2})^l + (1+\sqrt{2})^l\right) \le (1+\sqrt{2})^l.$$

□

作为推论，我们给出定理 A.2 中定义的多项式 P 相对于标准基的系数上界。

推论 A.14. 设 $d \geq 10\epsilon t$ 且令 $P(x) = \sum_{v=0}^{d} 2^{v-1} Q_v(x) c_{v,\ \epsilon,\ t}$ 为 次数为 d 的切比雪夫逼近$(1-\epsilon x)^{-t}$。则 P 相对于标准基的系数绝对值以上界 $2^{O(d)}$为界

证明。 记 $[a_m]Q(x)$ 为多项式 $Q(x)$ 在单项式 x^m 处的系数。 n,

$$\begin{aligned}
|[a_m]P(x)| &= \sum_{v=0}^{d} [a_m] 2^{v-1} Q_v(x) c_{v,\epsilon,t} \\
&\leq d \cdot \max_{v \leq d} \left| [a_m] 2^{v-1} Q_v(x) c_{v,\epsilon,t} \right| \\
&\leq d \cdot \max_{i \leq d} \left| [a_m] 2^{i-1} Q_i(x) \right| \cdot \max_{j \leq d} c_{j,\epsilon,t}
\end{aligned}$$

由引理 A.13，对于所有 i $\max_{m \leq d} \left| [a_m] 2^{i-1} Q_i(x) \right| \leq C(i) \leq 2^{O(i)}$，因此

$$\max_{i \leq d} \left| [a_m] 2^{i-1} Q_i(x) \right| \leq \max_{m, i \leq d} \left| [a_m] 2^{i-1} Q_i(x) \right| \leq \max_{i} 2^{O(i)} = 2^{O(d)}.$$

由引理 A.10， $C_{j,\ \epsilon,\ t} \leq O\left((\epsilon t + j) \exp(\Phi_{j,\ t,\ \epsilon})\right)$，并且由引理 A.11 和 A.12 $\Phi_{j,\ t,\ \epsilon} \leq \max\{-j, 100\epsilon t\}$，因此

$$C_{j,\epsilon,t} \leq O\left((\epsilon t + j) \exp(100\epsilon t)\right) \leq O\left(d \cdot 2^{O(\epsilon t)}\right).$$

因此，

$$[a_m]P(x) \leq d \cdot \max_{m, i \leq d} \left| [a_m] Q_i(x) 2^{i-1} \right| \cdot \max_{j \leq d} C_{j,\epsilon,t} \leq O\left(d^2 \cdot 2^{O(d)} \cdot 2^{O(\epsilon t)}\right) = 2^{O(d)}.$$

□

最后，我们可以证明命题 A.3 了。

命题 *A.3* 的证明。 回顾一下

$$q(y) := P\left(\frac{y-1}{\epsilon}\right).$$

令 a_0、.....、 a_d 表示 P相对于标准基的系数，并令 b_0、.....、 b_d 表示 q 相对于标准基的系数。换言之，

$$P(x) = a_d x^d + \ldots + a_0,$$

and

$$q(x) = b_d x^d + \ldots + b_0.$$

我们有

$$b_i = \sum_{j=i}^{d} a_j \left(-\frac{1}{\epsilon}\right)^{j-i} \binom{j}{i}.$$

根据推论 A.14，对于所有 i，我们都有 $|a_i| \leq 2^{O(d)}$ ，因此 q 的 i^{th}系数的绝对值的上界为

$$|b_i| \leq d \max_{j=i,\ldots,d} |a_j| \left(\frac{1}{\epsilon}\right)^{d-i} \binom{d}{i} \leq d \cdot 2^{O(d)} \left(\frac{1}{\epsilon}\right)^{d-i} \binom{d}{i}.$$

因此， q 的系数的绝对值有上界

$$\max_{i} |b_i| \leq \max_{i} d \cdot 2^{O(d)} \left(\frac{1}{\epsilon}\right)^{d-i} \binom{d}{i} \leq d \cdot 2^{O(d)} \left(\frac{1}{\epsilon}\right)^{d} \binom{d}{d/2} \leq 2^{O(\log(1/\epsilon)d)} = \left(\frac{1}{\epsilon}\right)^{O(d)},$$

正如所要求的。 □

B 对所有 $x \in V$ 的 ℓ_2-范数进行界定

我们在 D 节中的若干证明需要一个关于 $\|p_x^t\|_2 = \|M^t 1_x\|_2$ 的良好上界，该上界对所有 $x \in V$ 都成立，只要 t 足够大。注意，关于好顶点 $x \in V \setminus B_\delta$ 的上界可由以下事实推出：对所有 $x \in V \setminus B_\delta$，$\|f_x\|_2^2 = O(k/n)$（见注记 3.8）。直观地说，这是因为$\|f_x\|_2^2 = \sum_{i=1}^k \langle \mathbb{1}_x, v_i\rangle^2$，并且当 t 足够大时，$M^t \mathbb{1}_x$ 在 v_i、$i > k$ 上的投影都可以忽略不计。然而，对于坏顶点 $x \in B_\delta$，我们也需要类似的上界。在引理 4.17 和引理 4.28 的证明中会用到这一点；在证明中，我们对违反定义 4.13 中条件 $\left\langle p_x^t, \left(p_S^l\right)^2 \right\rangle \le O^*\left(\frac{1}{n^{3/2}} \cdot \frac{k^{3/2}}{|S|^2}\right)$ 或 $\left\langle (p_x^t)^2, p_S^l \right\rangle \le O^*\left(\frac{1}{n^{3/2}} \cdot \frac{k^{3/2}}{|S|}\right)$ 的 $x \in V$ 和 $x \in S$ 的数量给出上界。

在文献中，同时适用于 V 中所有顶点的界并不常见。我们所知的唯一此类界是 [GKL+21] 中的引理 22，它是通过仔细分析拉普拉斯矩阵底部 k 个特征向量满足的平衡条件得出的。他们证明了 $\|M^t \mathbb{1}_x\|_2 = O\left(k \cdot n^{-1/2+O(\epsilon/\varphi^2)}\right)$，比我们所需的界大 $\sqrt{k}$ 倍。遗憾的是，我们无法承受这一额外损失。下面我们将用不同的方法证明这个紧界。

引理 B.1. 设 M 为惰性随机游走矩阵 $M = \frac{1}{2}I + \frac{1}{2d}A$。那么，对于 $t \ge {}^{10}\log(n)\varphi^2$ 以及所有 $x \in V$，

$$\|M^t \mathbb{1}_x\|_2 = O\left(\sqrt{\frac{k}{n}}\right).$$

事实上，我们证明以下更强的结果：

引理 B.2. 设 $k \ge 2$ 为整数，$\varphi \in (0,1)$，且 $\epsilon \in (0,1)$。设 $G = (V, E)$ 为一个 d-正则图，且存在一个 (k, φ, ϵ)-聚类 $C_1, \ldots, C_k$。设 M 为惰性随机游走的转移矩阵，其以概率 $3/4$ 停留在当前顶点（以概率 $1/4d$沿任意一条边移动），即

$$\widetilde{M} = \frac{3}{4}I + \frac{1}{4d}A.$$

那么，对所有 $t \ge {}^{10}\log(n)\varphi^2$ 以及所有 $x \in V$，

$$\|\widetilde{M}^t \mathbb{1}_x\|_2 = O\left(\sqrt{\frac{k}{n}}\right).$$

引理 B.1 是引理 B.2 的一个推论。

引理 *B.1* 的证明。注意到 $M = I - \frac{1}{2}\mathcal{L}$和 $\widetilde{M} = I - \frac{1}{4}\mathcal{L}$。矩阵 M、$\widetilde{M}$、$\mathcal{L}$具有相同的特征向量。回忆一下，我们用 $\lambda_1 \le \ldots \le \lambda_n$表示 $\mathcal{L}$的特征值，并按升序排列。我们可以用 $\lambda_1, \ldots, \lambda_n$表示 M和 $\widetilde{M}$的特征值，分别为 $1 - \frac{\lambda_1}{2}, \ldots, 1 - \frac{\lambda_n}{2}$和 $1 - \frac{\lambda_1}{4}, \ldots, 1 - \frac{\lambda_n}{4}$。

定义 Σ为以下矩阵特征值构成的对角矩阵：M 为 $\widetilde{M}$的特征值构成的对角矩阵，并按非递增顺序排列。回忆一下，我们用 Σ 表示 M 的特征值构成的对角矩阵，并按非递增顺序排列。注意，由于 $0 \le \lambda_1 \le \ldots \le \lambda_n \le 2$，我们有 $0 \preceq \Sigma \preceq \Sigma$，因此对每个 t，都有 $\Sigma^t \preceq \Sigma^t$。所以，$M^t = U\Sigma^t U^\top \preceq U\widetilde{\Sigma}^t U^\top = \widetilde{M}^t$；对于每个 x和每个 $t \ge {}^{10}\log_{\varphi^2} n$，根据引理 B.2，

$$\|M^t \mathbb{1}_x\|_2^2 = \mathbb{1}_x^\top M^{2t} \mathbb{1}_x \le \mathbb{1}_x^\top \widetilde{M}^{2t} \mathbb{1}_x = \|\widetilde{M}^t \mathbb{1}_x\|_2^2 \le O(k/n),$$

如所需。 □

剩下要证明引理 B.2。

引理 *B.2* 的证明。令 G_{cross} 为 G 的子图，它由端点位于不同簇 C_1、.....、C_k 中的所有边组成，并添加自环，使每个顶点的度为 d。令 G_{in} 为 G 的子图，它由各簇内部的所有边组成，并添加自环，使每个顶点的度为 d。

令 M_{in} 和 M_{cross} 分别为 G_{in} 和 G_{cross} 上的惰性随机游走矩阵，其中停留在顶点 1/2 的概率为。于是，

$$\widetilde{M} = 1/2\widetilde{M}_{in} + 1/2\widetilde{M}_{cross}.$$

对于顶点上的任意分布 π，根据三角不等式，

$$\|\widetilde{M}\pi\|_2 \le 1/2\|\widetilde{M}_{in}\pi\|_2 + 1/2\|\widetilde{M}_{cross}\pi\|_2,$$

此外，

$$\|\widetilde{M}_{cross}\pi\|_2 \le \|\pi\|_2 \tag{40}$$

由于 M_{cross} 的特征值取值于 $[0,1]$,，因为 M_{cross} 是 d-正则图上随机游走的矩阵。

对于分布 π 和任意 $i \in [n]$，令 π_i 表示 π 在 C_i 上的限制，即 $(\pi_i)_x$ 在 $x \in C_i$ 时等于 π_x，否则为零。令 $M_{\mathrm{in},\ i}$ 表示 M_{in} 在 $C_i \times C_i$ 上的限制；也就是说，$M_{\mathrm{in},\ i} \in \mathbb{R}^{n\times n}$ 在 $C_i \times C_i$ 中的指标上与 M_{in} 一致，其他位置为零。注意，当 x 和 y 属于不同的簇时，$(M_{in})_{x,\ y} = 0$，因此

$$\|\widetilde{M}_{in}\pi\|_2^2 = \left\|\sum_{i=1}^{k}\widetilde{M}_{in,i}\pi_i\right\|_2^2 = \sum_{i=1}^{k}\|\widetilde{M}_{in,i}\pi_i\|_2^2. \tag{41}$$

固定 $M_{in,\ i}$ 的一组标准正交特征基 $v_{1,\ i}$，…，$v_{n,\ i}$。我们将 $\lambda_{l,\ i}$ 记为 $M_{in,\ i}$ 的第 l 大特征值，将 $v_{l,\ i}$ 记为其对应的单位范数特征向量。

注意，$M_{in,i}$ 的最小 $n - |C_i|$ 个特征值为 0。由于每个 C_i 都是一个 φ-扩展图，根据切格不等式，我们有 $\lambda_{2,\ C_i} \le 1 - \frac{\varphi^2}{4}$。由于对于每个 $i \in [k]$，$G_{in} \cap C_i$ 都是一个 d-正则图，$\lambda_{1,C_i} = 1$ 且$v_{1,i}$是每个 $i \in [k]$对应的簇 C_i 的归一化指示向量。

$$\begin{aligned}\|\widetilde{M}_{in,i}\pi_i\|_2^2 = \|\pi_i^\top v_{1,i} + \sum_{l=2}^{|C_i|}\lambda_{l,i}\pi_i^\top v_{l,i}\|_2^2 &\le |\pi_i^\top v_{1,i}|^2 + \left(1-\frac{\varphi^2}{4}\right)^2\left\|\sum_{l=2}^{|C_i|}\pi_i^\top v_{l,i}\right\|_2^2 \\ &\le \left(1-\frac{\varphi^2}{4}\right)^2\|\pi_i\|_2^2 + \left(1-\left(1-\frac{\varphi^2}{4}\right)^2\right)|\pi_i^\top v_{1,i}|^2 \\ &\le \left(1-\frac{\varphi^2}{4}\right)^2\|\pi_i\|_2^2 + O(\varphi^2)\cdot|\pi_i^\top v_{1,i}|^2.\end{aligned} \tag{42}$$

注意，$|\pi_i^\top v_{1,i}| = \frac{1}{\sqrt{|C_i|}}\|\pi_i\|_1$，因为$v_{1,i}$ 就是簇 C_i上的归一化指示向量。因此，

$$\sum_{i=1}^{k}|\pi_i^\top v_{1,i}|^2 = \sum_{i=1}^{k}\frac{1}{|C_i|}\|\pi_i\|_1^2 \le \frac{\eta k}{n}\sum_{i=1}^{k}\|\pi_i\|_1^2 \le \frac{\eta k}{n}\left\|\sum_{i=1}^{k}\pi_i\right\|_1^2 \le O(k/n), \tag{43}$$

其中，第二个不等式成立是因为 $\sum_{i=1}^{k}\pi_i = \pi$ 和 π 是一个概率分布。

Su对所有 $i \in [k]$ 求和方程 (42)，并结合方程 (41) 和 (43)，可得

$$\|\widetilde{M}_{in}\pi\|_2^2 \le \left(1-\frac{\varphi^2}{4}\right)^2\|\pi\|_2^2 + O(\varphi^2 k/n). \tag{44}$$

设 C 为一个足够大的普适常数，使得 $\|\widetilde{M}_{in}\pi\|_2^2 \le \left(1-\frac{\varphi^2}{4}\right)^2\|\pi\|_2^2 + C\varphi^2 k/n$. 我们可以将 M^t 表示为以下矩阵之和： $M^t = (1/2M_{in} + 1/2M_{cross})^t$. 回顾公式 (40)，对于任意分布 π'，都有 $\|M_{cross}\pi'\|_2^2 \le \|\pi'\|_2^2$. 因此，反复使用公式 (44) 和公式 (40)，可知，对于任意矩阵 M_l，由 l 个矩阵 M_{in} 和 $t-l$ 个矩阵 M_{cross}相乘所得的乘积，均满足

$$\begin{aligned}\|\widetilde{M}_l\pi\|_2^2 &\le \left(1-\frac{\varphi^2}{4}\right)^{2l}\|\pi\|_2^2 + C\varphi^2 k/n\cdot\left(1+\left(1-\frac{\varphi^2}{4}\right)^2+\ldots+\left(1-\frac{\varphi^2}{4}\right)^{2(l-1)}\right)\\ &\le \left(1-\frac{\varphi^2}{4}\right)^{2l}\|\pi\|_2^2 + 4Ck/n,\end{aligned} \tag{45}$$

其中，最后一个不等式成立，是因为和 $\left(1+\left(1-\frac{\varphi^2}{4}\right)^2+\ldots+\left(1-\frac{\varphi^2}{4}\right)^{2(l-1)}\right)$ 是一个几何数列，对于任意 l，它都以 $\frac{1}{1-\left(1-\frac{\varphi^2}{4}\right)^2}\le\frac{4}{\varphi^2}$ 为界，这是因为 $\varphi\le 1$。

因此，

$$\begin{aligned}\|\widetilde{M}^t\pi\|_2 &\le 2^{-t}\sum_{l=1}^{t}\binom{t}{l}\sqrt{\left(1-\frac{\varphi^2}{4}\right)^{2l}\|\pi\|_2^2+4Ck/n} \le 2^{-t}\sum_{l=1}^{t}\binom{t}{l}\left(\left(1-\frac{\varphi^2}{4}\right)^{l}\|\pi\|_2+\sqrt{4Ck/n}\right)\\ &\le 2^{-t}\left(1+1-\frac{\varphi^2}{4}\right)^{t}\|\pi\|_2+\sqrt{4Ck/n} = \left(1-\frac{\varphi^2}{8}\right)^{t}\|\pi\|_2+\sqrt{4Ck/n},\end{aligned} \tag{46}$$

其中第三个不等式使用了恒等式 $(1+z)^t=\sum_{l=0}^{t}\binom{t}{l}z^t$，该恒等式对所有 $z\in\mathbb{R}$ 都成立，并应用于 $z=\left(1-\frac{\varphi^2}{4}\right)$ 和 $z=1$。由于 π 是一个分布，我们有 $\|\pi\|_2\le 1$，因此对于每个 $t\ge\frac{10}{\log(n)\,\varphi^2}$，

$$\|\widetilde{M}^t\pi\|_2 \le 1/n+\sqrt{4Ck/n}\le O(\sqrt{k/n}).$$

□

C 谱嵌入的性质（引理 3.7 和引理 4.24 的证明）

本节首先证明引理 3.7，重述如下。

引理 3.7。对于所有 x、$y\in V\setminus B_\delta$，如果 x 和 y 属于同一个簇 C_i，那么

$$\left|\langle f_x,f_y\rangle-\frac{1}{|C_i|}\right|\le\left(\frac{4\sqrt{\epsilon}}{\varphi}+4\sqrt{\delta}\right)\frac{1}{|C_i|}.$$

证明。假设 x，$y\in C_i\setminus B_\delta$。那么

$$\begin{aligned}\left|\langle f_x,f_y\rangle-\|\mu_i\|_2^2\right| &= \left|\langle(f_x-\mu_i)+\mu_i,(f_y-\mu_i)+\mu_i\rangle-\|\mu_i\|_2^2\right|\\ &\le |\langle f_x-\mu_i,f_y-\mu_i\rangle|+|\langle f_x-\mu_i,\mu_i\rangle|+|\langle\mu_i,f_y-\mu_i\rangle| \qquad \text{by triangle inequality}\\ &\le \|f_x-\mu_i\|_2\|f_y-\mu_i\|_2+\|f_x-\mu_i,\mu_i\|_2+\|\mu_i\|_2\|f_y-\mu_i\|_2\\ &\le \frac{\delta}{|C_i|}+\frac{\sqrt{2\delta}}{|C_i|}+\frac{\sqrt{2\delta}}{|C_i|} \quad \text{by Definition 3.4 and Lemma 2.7}\\ &\le \frac{4\sqrt{\delta}}{|C_i|}.\end{aligned} \tag{47}$$

因此，

$$\left|\langle f_x,f_y\rangle-\frac{1}{|C_i|}\right|\le\left|\langle f_x,f_y\rangle-\|\mu_i\|_2^2\right|+\left|\|\mu_i\|_2^2-\frac{1}{|C_i|}\right|\le\frac{4\sqrt{\delta}}{|C_i|}+\frac{4\sqrt{\epsilon}}{\varphi}\frac{1}{|C_i|} \qquad \text{by (47) and Lemma 2.7.}$$

□

接下来，我们证明如下重述的引理 4.24。

引理 4.24. 对于每个顶点 $x \in V \setminus B_\delta$，都有

$$\Pr\left[\left|\langle \textsc{Sketch}(x), \textsc{Sketch}(x)\rangle - \|f_x\|_2^2\right| \le 0.01\|f_x\|_2^2\right] \le 10^{-6}.$$

上述概率是关于 Sketch 的内部随机性的，并且 Sketch 的两次调用相互独立。

证明。令 $\sigma_x \sum_t c_t \cdot \widehat{p}_x^t$ 和 $\sigma'_x \sum_t c_t \cdot \widehat{p}_x'^t$ 为对 Sketch(x) 的两次独立调用的输出，其中我们将两次调用内部随机符号的两个独立实例分别记为 σ_x 和 σ'_x。给定 $t_1, t_2 \in [t_{\min}, t_{\min} + t_\Delta]$，根据引理 B.1，我们有 $p_x^{t_1}(x) = \|M^{t_1}\mathbb{1}_x\|_2^2 \le O(\frac{k}{n})$，且 $\left\langle p_x^{t_1}, (p_x^{t_2})^2 \right\rangle \le |p_x^{t_1}\|_2 \|(p_x^{t_2})^2\|_2 \le \|p_x^{t_1}\|_2 \|p_x^{t_2}\|_2^2 \le O\left(\frac{k^{3/2}}{n^{3/2}}\right)$。这里的不等式 $\|(p_x^{t_2})^2\|_2 \le \|p_x^{t_2}\|_2^2$ 成立，这是因为注意到 $\|(p_x^{t_2})^2\|_2^2 = \sum_{v\in V} p_x^{t_2}(v)^4 \le \left(\sum_{v\in V} p_x^{t_2}(v)^2\right)^2 = \|p_x^{t_2}\|_2^4$。因此，我们可以应用引理 4.21，其中 $\beta = O(1), \gamma = O\left(\sqrt{\frac{n}{k}}\right), \rho = 10^{-6}$ ^{1} $(t_\Delta+1)_2 = O^*(1)$，且 $\xi = \frac{5\cdot 10^{-4}}{\eta\cdot|c_{t_1}||c_{t_2}|}(t_\Delta+1)^2 = O^*\left(\frac{1}{1/\epsilon}\right)^{t_\Delta}$（最后一个等式由定理 2.10 中的系数大小界得出）。在引理 4.21 中令 $Q, R = 7\cdot\sqrt{\frac{n}{k}}\cdot\frac{1}{\rho\cdot\xi^2} = O^*\left(\sqrt{\frac{n}{k}}\cdot\left(\frac{1}{\epsilon}\right)^{O(t_\Delta)}\right)$，并利用对所有 $x \in V \setminus B_\delta$ 均有 $\|f_x\|_2^2 = \Theta(\frac{k}{n})$ 这一事实（见注记 3.8），即可得到 [4]

$$\Pr_{\text{random walks}}\left[\left|\left\langle \sigma_x \widehat{p}_x^{t_1}, \sigma'_x \widehat{p}_x^{t_2}\right\rangle - \left\langle \sigma_x M^{t_1}\mathbb{1}_x, \sigma'_x M^{t_2}\mathbb{1}_x\right\rangle\right| \le \frac{5\cdot 10^{-4}}{|c_{t_1}||c_{t_2}|(t_\Delta+1)^2}\cdot\|f_x\|_2^2\right] \ge 1 - 10^{-6}\frac{1}{(t_\Delta+1)^2}. \tag{48}$$

对 $t_1, t_2 \in [t_{\min}, t_{\min}+t_\Delta]$ 的 $(t_\Delta+1)^2$ 个可能配对应用并集界，可得

$$\Pr_{\text{random walks}}\left[\left|\langle \textsc{Sketch}(x), \textsc{Sketch}(x)\rangle - \left\langle \sigma_x \sum_t c_t \cdot M^t\mathbb{1}_x, \sigma'_x \sum_t c_t \cdot M^t\mathbb{1}_x\right\rangle\right| > 0.0005\|f_x\|_2^2\right] \le 10^{-6}. \tag{49}$$

最后，我们有

断言 C.1. 对所有 $x \in V \setminus B_\delta$,

$$\left|\left\|\sum_t c_t M^t \mathbb{1}_x\right\|_2^2 - \|f_x\|_2^2\right| \le 0.0005\|f_x\|_2^2.$$

证明。该结论基本上由定理 2.10 得出。将 $p(M)\mathbb{1}_x$ 和 f_x 按特征向量

[4] 公式 (48) 给出了 $\langle p_x^{t_1}, p_x^{t_2}\rangle$ 与其经验版本之间的差值界。然而，由于我们的最终目标是近似多项式 $p(M)$，因此用 $M^{t_1}\mathbb{1}_x$，$M^{t_1}\mathbb{1}_x$ 而不是 $p_x^{t_1}, p_x^{t_2}$ 来表示该表达式是合理的。

tors $u_1, \dots u_n$ 拉普拉斯矩阵 $\mathcal{L}$表示，我们得到

$$\begin{aligned}
\left| \left\| \sum_t c_t M^t \mathbb{1}_x \right\|_2^2 - \|f_x\|_2^2 \right| &= \left| \|p(M)\mathbb{1}_x\|_2^2 - \|f_x\|_2^2 \right| \\
&= \left| \sum_{i=1}^{n} \langle u_i, \mathbb{1}_x \rangle^2 p((1-\lambda_i/2))^2 - \sum_{i=1}^{k} \langle u_i, \mathbb{1}_x \rangle^2 \right| \\
&\le \left| \sum_{i=1}^{k} \langle u_i, \mathbb{1}_x \rangle^2 p((1-\lambda_i/2))^2 - \sum_{i=1}^{k} \langle u_i, \mathbb{1}_x \rangle^2 \right| + \left| \sum_{i=k+1}^{n} \langle u_i, \mathbb{1}_x \rangle^2 p((1-\lambda_i/2))^2 \right| \\
&\le \left| \sum_{i=1}^{k} \langle u_i, \mathbb{1}_x \rangle^2 (1-\epsilon/\varphi^2)^2 - \sum_{i=1}^{k} \langle u_i, \mathbb{1}_x \rangle^2 \right| + n^{-8} \left| \sum_{i=k+1}^{n} \langle u_i, \mathbb{1}_x \rangle^2 \right| \\
&\le \frac{2\epsilon}{\varphi^2} \left| \sum_{i=1}^{k} \langle u_i, \mathbb{1}_x \rangle^2 \right| + n^{-8} \|\mathbb{1}_x\|_2^2 \\
&= \frac{2\epsilon}{\varphi^2} \|f_x\|_2^2 + n^{-8} \\
&\le 0.0005 \|f_x\|_2^2.
\end{aligned}$$

这里，第四行由备注 3.2 和定理 2.10 得出；最后一行则由 ϵ/φ^2 小于某个常数这一假设（见注 4.2），以及 $\|f_x\|_2^2 = \Theta(\binom{n}{k})$ 对于 $x \in V \setminus B_\delta$ 成立这一事实（见注记 3.8）得出。 □

结合公式（49）和断言 C.1，以至少 $1 - 10^{-6}$ 的概率，有$|\langle \textsc{Sketch}(x), \textsc{Sketch}(x) \rangle| \in \left[0.99\|f_x\|_2^2, 1.01|f_x\|_2^2\right]$，如所需。 □

最后，我们证明两个辅助引理，它们将在引理 4.28 和引理 4.12 的证明中用到。

引理 C.2。

$$\sum_{y \in V} f_y f_y^\top = I_k.$$

特别地，对于所有 $\alpha \in \mathbb{R}^k$，都有 $\sum_{y \in V} \langle \alpha, f_y \rangle^2 = \|\alpha\|^2$。

证明。我们有

$$\begin{aligned}
\sum_{y \in V} f_y f_y^\top &= \sum_{y \in V} U_{[k]}^\top \mathbb{1}_y \mathbb{1}_y^\top U_{[k]} && \text{根据 } f_y \text{的定义} \\
&= U_{[k]}^\top \left(\sum_{y \in V} \mathbb{1}_y \mathbb{1}_y^\top \right) U_{[k]} \\
&= U_{[k]}^\top I_n U_{[k]} \\
&= U_{[k]}^\top U_{[k]} \\
&= I_k,
\end{aligned}$$

最后一个等式由特征向量的正交归一性得出，因为 $U_{[k]}$ 是一个矩阵，其列为 $\mathcal{L}$ 的底部 k 个特征向量。

引理的第二部分由此得出，因为

$$\sum_{y \in V} \langle \alpha, f_y \rangle^2 = \alpha^\top \left(\sum_{y \in V} f_y f_y^\top \right) \alpha = \|\alpha\|_2^2.$$

□

引理 C.3。对于每个 $i \in [k]$

$$\mu_i \mu_i^\top \bullet \sum_{y \in V \setminus C_i} f_y f_y^\top \leq 4 \cdot \frac{\sqrt{\epsilon}}{\varphi} \cdot \|\mu_i\|_2^2$$

证明。我们有

$$\begin{aligned} \mu_i \mu_i^\top \bullet \sum_{y \in V \setminus C_i} f_y f_y^\top = \sum_{y \in V \setminus C_i} \langle \mu_i, f_y \rangle^2 &= \sum_{y \in V} \langle \mu_i, f_y \rangle^2 - \sum_{y \in C_i} \langle \mu_i, f_y \rangle^2 \\ &= \|\mu_i\|^2 - \sum_{y \in C_i} \langle \mu_i, f_y \rangle^2 \qquad \text{by Lemma C.2.} \end{aligned} \tag{50}$$

现在我们将对 $\sum_{y \in C_i} \langle \mu_i, f_y \rangle^2$ 项给出下界。对于每个 $y \in C_i$，我们有

$$\begin{aligned} \langle \mu_i, f_y \rangle^2 &= \left(\langle \mu_i, f_y - \mu_i \rangle + \langle \mu_i, \mu_i \rangle \right)^2 \\ &= \langle \mu_i, f_y - \mu_i \rangle^2 + 2 \langle \mu_i, f_y - \mu_i \rangle \|\mu_i\|_2^2 + \|\mu_i\|_2^4 \end{aligned} \tag{51}$$

此外，注意到

$$\begin{aligned} \sum_{y \in C_i} \langle \mu_i, f_y - \mu_i \rangle &= |C_i| \cdot \frac{1}{|C_i|} \sum_{y \in C_i} \langle \mu_i, f_y - \mu_i \rangle \\ &= |C_i| \left\langle \mu_i, \frac{1}{|C_i|} \sum_{y \in C_i} (f_y - \mu_i) \right\rangle \\ &= 0, \qquad \text{根据 } \mu_i \text{ (4) 的定义。} \end{aligned} \tag{52}$$

对 $y \in C_i$ 求和方程（51），并使用方程（52），得到

$$\begin{aligned} \sum_{y \in C_i} \langle \mu_i, f_y \rangle^2 &= \sum_{y \in C_i} \left(\langle \mu_i, f_y - \mu_i \rangle^2 + 2 \langle \mu_i, f_y - \mu_i \rangle \|\mu_i\|_2^2 + \|\mu_i\|_2^4 \right) \\ &\geq 2\|\mu_i\|_2^2 \sum_{y \in C_i} \langle \mu_i, f_y - \mu_i \rangle + |C_i| \cdot \|\mu_i\|_2^4 \qquad \text{by dropping the } \langle \mu_i, f_y - \mu_i \rangle^2 \text{ terms} \\ &= |C_i| \cdot \|\mu_i\|_2^4 \qquad \text{by Equation (52)} \\ &\geq \|\mu_i\|_2^2 \cdot \left(1 - \frac{4\sqrt{\epsilon}}{\varphi} \right) \qquad \text{by Lemma 2.7} \end{aligned} \tag{53}$$

结合方程（50）和方程（53）可得

$$\mu_i \mu_i^\top \bullet \sum_{y \in V \setminus C_i} f_y f_y^\top \leq \|\mu_i\|_2^2 - \left(1 - 4 \cdot \frac{\sqrt{\epsilon}}{\varphi} \right) \|\mu_i\|_2^2 = 4 \cdot \frac{\sqrt{\epsilon}}{\varphi} \|\mu_i\|_2^2.$$

□

D 良好分散集和典型顶点的性质

在本节中，我们证 明引理 4.11、引理 4.12、引理 4.17、引理 4.18、命题 4.20 和引理 4.28 .

引理 4.11。 如果 S 是一个良好分散集，那么其每个子集 $S' \subseteq S$ 也都是良好分散集。

证明。 假设 S 分布良好（依据定义 4.9），并假设 $S' \subseteq S$。现在我们来验证定义 4.9 中的每个条件对 S' 都成立。

条件（1）。 这可由假设 $S' \subseteq S$ 立即得出。

条件（2）。集合 S 和 S' 的低自碰撞概率也可以等价地表述为

$$\left\| \sum_{y \in S} M^t \mathbb{1}_y \right\|_2^2 = O^*\left(\frac{k^2}{n}\right)$$

and

$$\left\| \sum_{y \in S'} M^t \mathbb{1}_y \right\|_2^2 = O^*\left(\frac{k^2}{n}\right)$$

分别成立。由于所有向量 $M^t \mathbb{1}_y$ 的所有分量均为非负数，因此我们得到

$$\left\| \sum_{y \in S'} M^t \mathbb{1}_y \right\|_2^2 \leq \left\| \sum_{y \in S} M^t \mathbb{1}_y \right\|_2^2 = O^*\left(\frac{k^2}{n}\right).$$

条件 (3)。第一个条件可直接由假设 $S' \subseteq S$ 得出。为证明其余两个条件，注意对于每个 i 和每个 y，都有

$$\mu_i \mu_i^\top \bullet f_y f_y^\top = \langle \mu_i, f_y \rangle^2 \geq 0,$$

so

$$\mu_i \mu_i^\top \bullet \sum_{y \in S' \setminus C_i} f_y f_y^\top \leq \mu_i \mu_i^\top \bullet \sum_{y \in S \setminus C_i} f_y f_y^\top \leq 10^{-10} \cdot \|\mu_i\|_2^4$$

and

$$\mu_i \mu_i^\top \bullet \sum_{y \in S'} f_y f_y^\top \leq \mu_i \mu_i^\top \bullet \sum_{y \in S} f_y f_y^\top \leq O\left(\left(\varphi^2/\epsilon\right)^{1/3}\right) \|\mu_i\|_2^4.$$

□

引理 4.12。若 S 是一个至多含有 $O(k\log(\varphi^2/\epsilon))$ 个顶点的（多重）集合，且这些顶点是独立均匀随机采样得到的，那么该集合 S 以至少 0.9999 的概率是良好分散的。

证明。我们证明每个条件都以高常数概率成立。

条件（1）：与 B_δ 的重叠较小。我们证明条件（1）以至少 0.99998 的概率成立。由引理 3.6 可得 $|B_\delta| = O\left(\frac{1}{\delta} \cdot \epsilon/\varphi^2 \cdot n\right)$。因此我们有

$$\mathbb{E}[|S \cap B_\delta|] \leq O\left(\frac{1}{\delta} \cdot \frac{\epsilon}{\varphi^2} k \log(\varphi^2/\epsilon)\right).$$

根据马尔可夫不等式，我们得到，对于某个足够大的常数 C，有

$$\Pr\left[|S \cap B_\delta| \geq C \cdot \frac{\epsilon/\varphi^2}{\delta} \cdot k \log(\varphi^2/\epsilon)\right] \leq 0.00001.$$

条件（2）：低自碰撞概率。 我们证明条件（2）以至少 0.99998 的概率成立。回顾公式（15），$p_S^t = \frac{1}{|S|} \sum_{y \in S} M^t \mathbb{1}_y$，并注意到条件（2）等价于

$$\left\| \sum_{y \in S} M^t \mathbb{1}_y \right\|^2 \leq O^*\left(\frac{k^2}{n}\right) \quad \text{for all } t \in [t_{\min}, t_{\min} + t_\Delta].$$

更准确地说，我们将证明其以至少 0.99994 的概率成立

$$\left\|\sum_{y\in S} M^l \mathbb{1}_y\right\|^2 \le O\left(\frac{k^2}{n}\cdot \log^2(\varphi^2/\epsilon)\cdot(t_\Delta+1)\right) \quad \text{for all } t\in[t_{\min}, t_{\min}+t_\Delta], \tag{54}$$

其中 $t_\Delta = 20\cdot \frac{\epsilon}{\varphi^2}\log\ n + 2\log(\varphi^2/\epsilon) = O^*(1)$。将 $S=\{y_1,\dots,y_L\}$ 写作大小固定为$L = O(k\cdot\log(\varphi^2/\epsilon))$ 的随机（多重）集合，我们有 $y_1,\dots,y_L$ 是在 V 上均匀分布的独立随机变量。由 y_ls 的独立性可得

$$\begin{aligned}
\mathop{\mathbb{E}}_{S}\left[\sum_{l,l'\in[L]:l\neq l'}\left\langle M^t\mathbb{1}_{y_l}, M^t\mathbb{1}_{y_{l'}}\right\rangle\right] &= \mathop{\mathbb{E}}_{S}\left[\sum_{l,l'\in[L]:l\neq l'}\sum_{v\in V} M^t\mathbb{1}_{y_l}(v) M^t\mathbb{1}_{y_{l'}}(v)\right]\\
&= L\cdot(L-1)\cdot \mathop{\mathbb{E}}_{y,y'}\left[\sum_{v\in V} M^t\mathbb{1}_y(v) M^t\mathbb{1}_{y'}(v)\right]\\
&= L\cdot(L-1)\cdot\frac{1}{n^2}\sum_{v\in V}\left(\sum_{y\in V} M^t\mathbb{1}_y(v)\right)^2\\
&= \sum_{v\in V} L\cdot(L-1)\cdot\frac{1}{n^2} \qquad \text{since } \sum_{y\in V} M^t\mathbb{1}_y = \mathbb{1}\\
&\le \frac{L^2}{n}.
\end{aligned} \tag{55}$$

接下来，对于每个 $y\in V$ 和 $t\ge t_{\min}$，根据引理 B.1，有 $\|M^t\mathbb{1}_y\|_2^2 \le O(k/n)$。结合公式 (55)，得到

$$\mathop{\mathbb{E}}_{S}\left[\left\|\sum_{y\in S} M^t\mathbb{1}_y\right\|_2^2\right] = \mathop{\mathbb{E}}_{S}\left[\sum_{l\in[L]}\|M^t\mathbb{1}_{y_l}\|_2^2\right] + \mathop{\mathbb{E}}_{S}\left[\sum_{l,l'\in[L]:l\neq l'\in S}\left\langle M^t\mathbb{1}_{y_l}, M^t\mathbb{1}_{y_{l'}}\right\rangle\right] \le O\left(\frac{k^2\log^2(\varphi^2/\epsilon)}{n}\right).$$

根据马尔可夫不等式，以至少 $\frac{0.99998}{t_\Delta+1}$ 的概率，有

$$\sum_{v\in V}\left(\sum_{y\in S} M^l\mathbb{1}_y(v)\right)^2 \le O\left(\frac{k^2}{n}\cdot\log^2(\varphi^2/\epsilon)\cdot(t_\Delta+1)\right).$$

对所有 $l\in[t_{\min}, t_{\min}+t_\Delta]$. 应用并集界，即可得到结论

条件 (3)：嵌入空间中近似各向同性。我们证明条件 (3) 以至少 0.99994 的概率成立。

注意到

$$\mathop{\mathbb{E}}_{S}\left[\sum_{y\in S}\|f_y\|_2^2\right] = |S|\cdot\mathop{\mathbb{E}}_{y\sim V}\left[\|f_y\|_2^2\right] = \frac{|S|}{n}\sum_{v\in V}\|f_v\|_2^2 = \frac{|S|\cdot k}{n} = O\left(\frac{k^2\log(\varphi^2/\epsilon)}{n}\right),$$

其中第三个不等式由 $\sum_{v\in V}\|f_v\|_2^2 = \mathrm{Tr}\left(\sum_{v\in V} f_v f_v^\top\right) = \mathrm{Tr}(I_k) = k$ 得出，这是引理 C.2 的结论。因此，根据马尔可夫不等式，该结论以至少 0.99998 的概率成立:

$$\sum_{y\in S}\|f_y\|_2^2 \le O\left(\frac{k^2\log(\varphi^2/\epsilon)}{n}\right),$$

正如所需。接下来，注意到

$$\begin{aligned}\mathop{\mathbb{E}}_{S}\left[\sum_{i\in[k]}\left(\mu_i\mu_i^\top \bullet \sum_{y\in S\setminus C_i} f_y f_y^\top\right)\right] &= \sum_{i\in[k]}\left(\mu_i\mu_i^\top \bullet \mathop{\mathbb{E}}_{S}\left[\sum_{y\in S} f_y f_y^\top \mathbb{1}\{y\notin C_i\}\right]\right)\\ &= \frac{|S|}{n}\sum_{i\in[k]}\left(\mu_i\mu_i^\top \bullet \sum_{y\in V} f_y f_y^\top \mathbb{1}\{y\notin C_i\}\right) \qquad (56)\\ &= \frac{|S|}{n}\left(\sum_{i\in[k]}\mu_i\mu_i^\top \bullet \left(\sum_{y\in V\setminus C_i} f_y f_y^\top\right)\right).\end{aligned}$$

为了给 (56) 中的最终表达式一个上界，注意到根据引理 C.3，我们有

$$\frac{|S|}{n}\sum_{i\in[k]}\left(\mu_i\mu_i^\top \bullet \sum_{y\in V\setminus C_i} f_y f_y^\top\right) = \frac{|S|}{n}\sum_{i\in[k]}\sum_{y\in V\setminus C_i}\langle \mu_i, f_y\rangle^2 \le 4\cdot\frac{|S|}{n}\cdot\frac{\sqrt{\epsilon}}{\varphi}\sum_{i\in[k]}\|\mu_i\|_2^2 \le O\left(\left(\frac{\epsilon}{\varphi^2}\right)^{1/2}\cdot\frac{|S|\cdot k^2}{n^2}\right), \qquad (57)$$

其中，最后一个不等式利用了以下事实：根据注 2.8，对于每个 i， $\|\mu_i\|_2^2 = \Theta(k/n)$。结合 (56) 和 (57)，并应用马尔可夫不等式，以至少 0.99998 的概率，

$$\sum_{i\in[k]}\mu_i\mu_i^\top \bullet \sum_{y\in S\setminus C_i} f_y f_y^\top \le \alpha\left(\frac{\epsilon}{\varphi^2}\right)^{1/2}\cdot\frac{|S|\cdot k^2}{n^2} \qquad (58)$$

对于某个充分大的常数 α。对于所有 i，除了至多 $\frac{\alpha}{c}\cdot k\cdot(\epsilon/\varphi^2)^{1/3}\log(\varphi^2/\epsilon) = O\left(k\cdot(\epsilon/\varphi^2)^{1/3}\log(\varphi^2/\epsilon)\right.$，其中 c 是充分小的常数

$$\mu_i\mu_i^\top \bullet \sum_{y\in S\setminus C_i} f_y f_y^\top \le c\cdot\frac{|S|\cdot k^2\cdot(\epsilon/\varphi^2)^{1/2}}{k\cdot(\epsilon/\varphi^2)^{1/3}\log(\varphi^2/\epsilon)\cdot n^2} \le c\cdot\left(\frac{\epsilon}{\varphi^2}\right)^{1/6}\cdot\|\mu_i\|_2^4 < 10^{-10}\|\mu_i\|_2^4. \qquad (59)$$

如果违反条件的指标数量 i更高，就会与方程 (58) 矛盾。最后，注意到

$$\begin{aligned}\mathop{\mathbb{E}}_{S}\left[\sum_{i\in[k]}\left(\mu_i\mu_i^\top \bullet \sum_{y\in S} f_y f_y^\top\right)\right] &= \sum_{i\in[k]}\mu_i\mu_i^\top \bullet \mathbb{E}\left[\sum_{y\in S} f_y f_y^\top\right]\\ &= \frac{|S|}{n}\sum_{i\in[k]}\mu_i\mu_i^\top \bullet \sum_{y\in V} f_y f_y^\top\\ &= \frac{|S|}{n}\sum_{i\in[k]}\|\mu_i\|_2^2 && \text{因为 } \sum_{y\in V} f_y f_y^\top = I_k\\ &= O\left(\frac{|S|\cdot k^2}{n^2}\right) && \text{根据注 2.8。}\end{aligned}$$

因此，根据马尔可夫不等式，以至少 0.99998 的概率，有

$$\sum_{i\in[k]}\mu_i\mu_i^\top \bullet \sum_{y\in S} f_y f_y^\top \le \kappa\frac{|S|\cdot k^2}{n^2}.$$

对于足够大的常数 κ。对于至多 $\frac{\kappa}{c'}\cdot k\cdot(\epsilon/\varphi^2)^{1/3}\log(\varphi^2/\epsilon) = O(k\cdot(\epsilon/\varphi^2)^{1/3}\log(\varphi^2/\epsilon))$ 个 $i\in[k]$ 的取值（其中 c' 足够小），

$$\mu_i\mu_i^\top \bullet \sum_{y\in S} f_y f_y^\top \le c'\cdot\frac{|S|\cdot k^2}{k\cdot(\epsilon/\varphi^2)^{1/3}\log(\varphi^2/\epsilon)\cdot n^2} \le c'\cdot\frac{k^2\cdot(\varphi^2/\epsilon)^{1/3}}{n^2} \le (\varphi^2/\epsilon)^{1/3}\|\mu_i\|_2^4,$$

运用我们推导方程（59）时所用的相同论证。

最后，由于两个条件各自对除至多 $O(k\cdot(\epsilon/\varphi^2)^{1/3}\log(\varphi^2/\epsilon))$ 个 $i\in[k]$ 值以外的所有值，都以至少 0.99998 的概率成立，因此，两个条件同时成立的概率至少为 0.99996，且除至多 $O(k\cdot(\epsilon/\varphi^2)^{1/3}\log(\varphi^2/\epsilon))$ 个 $i\in[k]$ 值以外，对所有值均成立。

根据并集界，三个条件（**1**）、（**2**）和（**3**）同时成立的概率至少为 0.9999。

□

接下来，我们证明引理 4.17。为方便读者，现将其重述如下。

引理 4.17. 对于每个良好分散集 S，均有

$$|\{x\in V: x \text{ is not } \textit{strongly typical with respect to } S\}|\le O\left(n\cdot\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\cdot\log(\varphi^2/\epsilon)\right).$$

证明. 设 S 是一个良好分散集（依据定义 4.9）。回顾定义 4.13 和定义 4.15，关于 S 的强典型顶点由 5 个条件定义。对于每个条件，我们定义 V 中违反该条件的所有点所构成的集合，并分别界定这些集合的大小。令

$$F_1 \coloneqq \{x\in V: S\cap B_\delta\cap C_{i(x)}\neq\emptyset\}$$

$$F_2 \coloneqq \left\{x\in V: |S\cap C_{i(x)}|\ge 10^{-5}\delta^{-1/2}\right\}$$

$$F_3 \coloneqq \left\{x\in V: p_S^t(x)\ge 10^5\cdot\left(\frac{\varphi^2}{\epsilon}\right)^{1/3}\cdot\frac{t_\Delta+1}{|S|}\cdot\frac{k}{n} \text{ for some } t\in[2t_{\min},2t_{\min}+2t_\Delta]\right\}$$

$$F_4 \coloneqq \left\{\begin{array}{l} x\in V: \left\langle p_x^t,\left(p_S^l\right)^2\right\rangle\ge 10^5\cdot\left(\frac{\varphi^2}{\epsilon}\right)^{1/3}\cdot\frac{(t_\Delta+1)^3}{|S|^2}\cdot\left(\frac{k}{n}\right)^2\log(\varphi^2/\epsilon) \\ \qquad \text{or } \left\langle\left(p_x^t\right)^2,p_S^l\right\rangle\ge 10^5\cdot\left(\frac{\varphi^2}{\epsilon}\right)^{1/3}\cdot\frac{(t_\Delta+1)^3}{|S|}\cdot\left(\frac{k}{n}\right)^2\log(\varphi^2/\epsilon) \text{ for some } t,l\in[t_{\min},t_{\min}+t_\Delta]\end{array}\right\}$$

$$F_5 \coloneqq \left\{x\in V: (f_x-\mu_{i(x)})(f_x-\mu_{i(x)})^\top\bullet\sum_{y\in S}f_yf_y^\top\ge 10^{-10}\|f_x\|_2^4\right\}$$

分别表示定义 4.13 中不满足条件（**1**）、（**2**）、（**3**）、（**4**）和（**5**）的顶点集合。于是，

$|\{x\in V:$ x 相对于 S$\}|\le|F_1|+|F_2|+|F_3|+|F_4|+|F_5|$，不是强典型的，

因此，只需证明 $|F_1|$、$|F_2|$、$|F_3|$、$|F_4|$、$|F_5|\le O\left(n\cdot\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\cdot\log(\varphi^2/\epsilon)\right)$。现在我们分别对这五个集合进行界定。

界定 $|F_1|$： 我们有

$$\begin{aligned}|F_1| &= \sum_{i=1}^k\mathbb{1}\{S\cap B_\delta\cap C_i\neq\emptyset\}|C_i| \\ &\le \max_i|C_i|\cdot|S\cap B_\delta| \\ &\le \eta\cdot\frac{n}{k}\cdot O\left((\epsilon/\varphi^2)^{1/3}\cdot k\cdot\log(\varphi^2/\epsilon)\right) \qquad \text{根据定义 4.9 中的条件 (1)} \\ &= O\left(n\cdot(\epsilon/\varphi^2)^{1/3}\cdot\log(\varphi^2/\epsilon)\right).\end{aligned}$$

界定 $|F_2|$： 我们有

$$\begin{aligned}
|F_2| &= \sum_{i=1}^{k} \mathbb{1}\{|S \cap C_i| \geq 10^{-5} \cdot \delta^{-1/2}\}|C_i| \\
&\leq \max_i |C_i| \cdot |\{i \in [k] : |S \cap C_i| \geq 10^{-5}\delta^{-1/2}\}| \\
&\leq \eta \cdot \frac{n}{k} \frac{|S|}{10^{-5}\delta^{-1/2}} \\
&= O\left(\delta^{1/2} \log(\varphi^2/\epsilon) \cdot n\right) && \text{since } |S| \leq O(k \log(\varphi^2/\epsilon)) \text{ by Definition 4.9} \\
&= O\left(n \cdot (\epsilon/\varphi^2)^{1/3} \log(\varphi^2/\epsilon)\right) && \text{by choice of } \delta = O\left((\epsilon/\varphi^2)^{2/3}\right).
\end{aligned}$$

上界 $|F_3|$： </style>对于，令，$t := \left\{x \in V\colon\ p_S^t(x) \geq 10^5 \cdot \left(\frac{\varphi^2}{\epsilon}\right)^{1/3} \cdot \frac{t_\Delta+1}{|S|} \cdot \frac{k}{n}\right\}$。回顾式（15），$p_S^t(x) = \frac{1}{|S|}\sum_{y\in S} M^t \mathbb{1}_y(x) = \frac{1}{|S|}\langle \mathbb{1}_x,\ \sum_{y\in S} M^t \mathbb{1}_y\rangle$。因此，我们有

$$\begin{aligned}
\sum_{x\in V} p_S^t(x) &= \frac{1}{|S|}\sum_{x\in V}\left\langle \mathbb{1}_x, \sum_{y\in S} M^t \mathbb{1}_y \right\rangle \\
&= \frac{1}{|S|}\sum_{y\in S}\left\langle \mathbb{1}, M^{t_2}\mathbb{1}_y\right\rangle \\
&= 1 && \text{since } \sum_{v\in V} M^{t_2}\mathbb{1}_y(v) = 1, \text{ for all } y \in V
\end{aligned}$$

现在，对于所有 x，我们都有 $0 \leq p_S^t(x)$，且 $\sum_{x\in V} p_S^t(x) = 1$，由此可得

$$|F_{3,t}| \leq \frac{1}{10^5(\varphi^2/\epsilon)^{1/3}(t_\Delta+1)/|S| \cdot k/n} = O\left(\frac{1}{t_\Delta+1}\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\log(\varphi^2/\epsilon)\cdot n\right).$$

回顾一下，根据 F_3 的定义，我们有 $F_3 = \cup_{t\in[2t_{\min}, 2t_{\min}+2t_\Delta]} F_{3,\ t}$。对 $[2t_{\min}, 2t_{\min}+2t_\Delta]$ 中 t 的 $2(t_\Delta+1)$ 个取值取并集，可得

$$|F_3| \leq \sum_{t\in[2t_{\min}, 2t_{\min}+2t_\Delta]} |F_{3,t}| \leq O\left(\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\log(\varphi^2/\epsilon)\cdot n\right).$$

界定 $|F_4|$： 由（15）可知，$p_S^l(v) = \frac{1}{|S|}\sum_{y\in S} M^l\mathbb{1}_y(v)$，因此 $(p_S^l(v))^2 = \frac{1}{|S|^2}\sum_{y,\ y'\in S} M^l\mathbb{1}_y(v) M^l\mathbb{1}_{y'}(v)$。对于每个 $t,\ l \in [t_{\min}, t_{\min}+t_\Delta]$，定义

$$\begin{aligned}
F_{4,t,l} &= \left\{x \in V : \left\langle p_x^t, \left(p_S^l\right)^2\right\rangle + \left\langle \left(p_x^t\right)^2, p_S^l\right\rangle \geq 10^5 \cdot \left(\frac{\varphi^2}{\epsilon}\right)^{1/3} \cdot \frac{(t_\Delta+1)^3}{|S|} \cdot \left(\frac{k}{n}\right)^2 \log(\varphi^2/\epsilon)\right\} \\
&= \left\{\begin{array}{l} x \in V : \sum_{v\in V}(M^t\mathbb{1}_x(v))^2 \sum_{y\in S} M^l\mathbb{1}_y(v) + \sum_v M^t\mathbb{1}_x(v) \sum_{y,y'\in S} M^l\mathbb{1}_y(v) M^l\mathbb{1}_{y'}(v) \\ \qquad \geq 10^5 \cdot \left(\frac{\varphi^2}{\epsilon}\right)^{1/3} (t_\Delta+1)^3 \left(\frac{k}{n}\right)^2 \log(\varphi^2/\epsilon) \end{array}\right\}.
\end{aligned}$$

于是 $F_4 \subseteq \bigcup_{t,\ l\in[t_{\min}, t_{\min}+t_\Delta]} F_{4,\ t,\ l}$。定义 4.9 中的条件**(2)** 可以等价地写成 as

$$\sum_{v\in V}\left(\sum_{y\in S} M^l\mathbb{1}_y(v)\right)^2 \leq O\left(\frac{k^2}{n}\cdot\log^2(\varphi^2/\epsilon)\cdot(t_\Delta+1)\right) \quad \text{for all } l \in [t_{\min}, t_{\min}+t_\Delta].$$

因此，对所有 $t,\ l \in [t_{\min}, t_{\min}+t_\Delta]$，我们有

$$\sum_{v\in V}\sum_{x\in V} M^t\mathbb{1}_x(v)\sum_{y,y'\in S} M^l\mathbb{1}_y(v)M^l\mathbb{1}_{y'}(v) = \sum_{v\in V}\sum_{y,y'\in S} M^l\mathbb{1}_y(v)M^l\mathbb{1}_{y'}(v) \qquad \text{since } \sum_{x\in V} M^t\mathbb{1}_x = \mathbb{1}$$
$$\leq O\left(\frac{k^2}{n}\cdot\log^2(\varphi^2/\epsilon)\cdot(t_\Delta+1)\right) \quad \text{by } (\mathbf{2}) \text{ in Definition 4.9.}$$

接下来，注意到 $M^t\mathbb{1}_x(v) = \langle \mathbb{1}_v,\ M^t\mathbb{1}_x\rangle = \langle \mathbb{1}_x,\ M^t\mathbb{1}_v\rangle = M^t\mathbb{1}_v(x)$，因为 M是一个对称矩阵。所以

$$\begin{aligned}
\sum_{v\in V}\sum_{x\in V}(M^t\mathbb{1}_x(v))^2\sum_{y\in S}M^l\mathbb{1}_y(v) &= \sum_{v\in V}\sum_{x\in V}(M^t\mathbb{1}_v(x))^2\sum_{y\in S}M^l\mathbb{1}_y(v)\\
&= \sum_{v\in V}\|M^t\mathbb{1}_v\|_2^2\sum_{y\in S}M^l\mathbb{1}_y(v)\\
&\leq O(k/n)\cdot\sum_{v\in V}\sum_{y\in S}M^l\mathbb{1}_y(v) && \text{因为根据引理 B.1，}\ \|M^t\mathbb{1}_v\|_2^2 = O(k/n)\\
&= O\left(\frac{k}{n}\cdot|S|\right) && \text{因为对所有 } y\in V,\ \|M^l\mathbb{1}_y\|_1 = 1\\
&= O\left(\frac{k^2\log(\varphi^2/\epsilon)}{n}\right).
\end{aligned}$$

因此，

$$\sum_{x\in V}\left(\sum_{v\in V}(M^t\mathbb{1}_x(v))^2\sum_{y\in S}M^l\mathbb{1}_y(v) + \sum_v M^t\mathbb{1}_x(v)\sum_{y,y'\in S}M^l\mathbb{1}_y(v)M^l\mathbb{1}_{y'}(v)\right) \leq \alpha\cdot\frac{k^2}{n}\cdot\log^2(\varphi^2/\epsilon)\cdot(t_\Delta+1) \tag{60}$$

对于一个足够大的常数 α。根据 $F_{4,\ t,\ l}$ 的定义，对于任意 $x\in F_{4,\ t,\ l}$，左端中对应的求和项必须至少为 $10^5\cdot\binom{\varphi^2}{\epsilon}^{1/3}(t_\Delta+1)^3\binom{k}{n}^2\log(\varphi^2/\epsilon)$。与此同时，这类求和项至多有 $\frac{\alpha}{10^5}\cdot\frac{1}{(t_\Delta+1)^2}\binom{\epsilon}{\varphi^2}^{1/3}\log(\varphi^2/\epsilon)\cdot n$ 项，否则等式 (60) 不成立。因此，

$$|F_{4,t,l}| \leq O\left(\frac{1}{(t_\Delta+1)^2}\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\log(\varphi^2/\epsilon)\cdot n\right). \tag{61}$$

对所有 $(t_\Delta+1)^2$ 对 $t,\ l\in[t_{\min}, t_{\min}+t_\Delta]$ 求和可得

$$|F_4| \leq \sum_{t,l\in[t_{\min},t_{\min}+t_\Delta]}|F_{4,t,l}| \leq O\left(\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\log(\varphi^2/\epsilon)\cdot n\right).$$

界定 $|F_5|$： 回顾引理 3.5，对于每个向量 $\alpha\in\mathbb{R}^k$，都有

$$\alpha^\top\left(\sum_{x\in V}(f_x-\mu_{i(x)})(f_x-\mu_{i(x)})^\top\right)\alpha \leq \frac{4\epsilon}{\varphi^2}\|\alpha\|_2^2. \tag{62}$$

因此，我们有

$$\begin{aligned}\sum_{x\in V}(f_x-\mu_{i(x)})(f_x-\mu_{i(x)})^\top \bullet \sum_{y\in S} f_y f_y^\top &= \mathrm{Tr}\left(\sum_{x\in V}(f_x-\mu_{i(x)})(f_x-\mu_{i(x)})^\top \sum_{y\in S} f_y f_y^\top\right)\\ &= \sum_{y\in S}\mathrm{Tr}\left(\sum_{x\in V}(f_x-\mu_{i(x)})(f_x-\mu_{i(x)})^\top f_y f_y^\top\right)\\ &= \sum_{y\in S} f_y^\top\left(\sum_{x\in V}(f_x-\mu_{i(x)})(f_x-\mu_{i(x)})^\top\right) f_y\\ &\le \frac{4\epsilon}{\varphi^2}\sum_{y\in S}\|f_y\|_2^2 && \text{by (62)}\\ &\le \alpha\frac{\epsilon}{\varphi^2}\cdot\frac{k^2}{n}\log(\varphi^2/\epsilon) && \text{根据注记 3.8}\end{aligned} \tag{63}$$

其中 α 是一个充分大的常数。这里，最后一个不等式也用到了假设 $|S| = O(k\log(\varphi/\epsilon)$，参见定义 4.9。

与推导方程 (61) 时所用的平均论证类似，我们对左端的和应用马尔可夫不等式，利用下界 $\|f_x\|_2^2 \ge c\cdot\frac{k}{n}$；当常数 c 充分小时，该下界对所有 $x\in V\setminus B_\delta$ 都成立（见注记 3.8）。这意味着，对于至多 $\frac{\alpha\cdot 10^{10}}{c^2}\frac{\epsilon}{\varphi^2}\log(\varphi^2/\epsilon)$ 个顶点 $\cdot n$，其对应的求和项才会超过 $10^{-10}\cdot\|f_x\|_2^4$。否则，仅对这些项求和就会大于 $\frac{\alpha}{c^2}\cdot\|f_x\|_2^4\frac{\epsilon}{\varphi^2}\log(\varphi^2/\epsilon)\cdot n \ge \alpha\frac{\epsilon}{\varphi^2}\cdot\frac{k^2}{n}\log(\varphi^2/\epsilon)$，这将与方程 (63) 矛盾。严格来说，我们得到

$$\left|\left\{x\in V\setminus B_\delta : (f_x-\mu_{i(x)})(f_x-\mu_{i(x)})^\top \bullet \sum_{y\in S} f_y f_y^\top \ge 10^{-10}\|f_x\|_2^4\right\}\right| \le O\left(\frac{\epsilon}{\varphi^2}\log(\varphi^2/\epsilon)\cdot n\right).$$

因此

$$\begin{aligned}\left|\left\{x\in V : (f_x-\mu_{i(x)})(f_x-\mu_{i(x)})^\top \bullet \sum_{y\in S} f_y f_y^\top \ge 10^{-10}\|f_x\|_2^4\right\}\right| &\le |B_\delta| + O\left(\frac{\epsilon}{\varphi^2}\log(\varphi^2/\epsilon)\cdot n\right)\\ &\le O\left(\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\log(\varphi^2/\epsilon)\cdot n\right).\end{aligned}$$

□

接下来，我们证明引理 4.18，并在下文重新陈述，以便读者查阅。

引理 4.18. 设 $S\subseteq V$ 为一个集合，并假设相对于 S 而言，x 是典型的（依据定义 *4.13*）。那么

$$f_x f_x^\top \bullet \sum_{y\in S\setminus C_{i(x)}} f_y f_y^\top \le 10^{-6}\cdot\|f_x\|_2^4,$$

and

$$f_x f_x^\top \bullet \sum_{y\in S} f_y f_y^\top \le (\varphi^2/\epsilon)^{1/3}\cdot\|f_x\|_2^4.$$

换言之，f_x 与其他簇的谱嵌入几乎没有相关性，而与整个集合 S 的相关性有界。

证明。 第一条性质可由下文的命题 D.1 以及假设 x 相对于 S 是典型的（根据定义 4.13）立即得出。

命题 D.1。 如果 $x \in V \setminus B_\delta$ 满足 $\mu_{i(x)}^\top \mu_{i(x)} \bullet \sum_{y \in S \setminus C_{i(x)}} f_y^\top f_y \leq 10^{-10} \cdot \|\mu_{i(x)}\|_2^4$ 且 $(f_x - \mu_{i(x)})(f_x - \mu_{i(x)})^\top \bullet \sum_{y \in S} f_y f_y^\top \leq 10^{-10} \|f_x\|_2^4$，那么

$$f_x^\top f_x \bullet \sum_{y \in S} f_y f_y^\top \leq 10^{-9} \cdot \|f_x\|_2^4.$$

证明。我们有

$$\begin{aligned} f_x^\top f_x \bullet \sum_{y \in S} f_y f_y^\top &= f_x^\top \left(\sum_{y \in S \setminus C_{i(x)}} f_y f_y^\top \right) f_x \\ &= (f_x - \mu_{i(x)})^\top \left(\sum_{y \in S \setminus C_{i(x)}} f_y f_y^\top \right) (f_x - \mu_{i(x)})^\top + (f_x - \mu_{i(x)}) \left(\sum_{y \in S \setminus C_{i(x)}} f_y f_y^\top \right) \mu_{i(x)} \\ &\quad + \mu_{i(x)}^\top \left(\sum_{y \in S \setminus C_{i(x)}} f_y f_y^\top \right) (f_x - \mu_{i(x)}) + \mu_{i(x)}^\top \left(\sum_{y \in S \setminus C_{i(x)}} f_y f_y^\top \right) \mu_{i(x)}. \end{aligned}$$

为了界定第一项，我们利用 $\sum_{y \in C_{i(x)}} f_y f_y^\top$ 是半正定矩阵这一事实，以及这一假设 (64)

$$(f_x - \mu_{i(x)})(f_x - \mu_{i(x)})^\top \bullet \sum_{y \in S \setminus C_{i(x)}} f_y f_y^\top \leq (f_x - \mu_{i(x)})(f_x - \mu_{i(x)})^\top \bullet \sum_{y \in S} f_y f_y^\top \leq 10^{-10} \|f_x\|_2^4 \tag{65}$$

为了界定公式（64）中的最后一项，我们利用这一假设

$$\mu_{i(x)}^\top \mu_{i(x)} \bullet \sum_{y \in S \setminus C_{i(x)}} f_y^\top f_y \leq 10^{-10} \cdot \|\mu_{i(x)}\|_2^4 \leq 4 \cdot 10^{-10} \|f_x\|_2^4 \qquad \text{by Remark 3.8.} \tag{66}$$

为了界定公式（64）中的第二项和第三项，我们利用 $\sum_{y \in S \setminus C_{i(x)}} f_y f_y^\top$ 是半正定矩阵这一事实，因此

$$\begin{aligned} &(f_x - \mu_{i(x)})^\top \left(\sum_{y \in S \setminus C_{i(x)}} f_y f_y^\top \right) \mu_{i(x)} + \mu_{i(x)}^\top \left(\sum_{y \in S \setminus C_{i(x)}} f_y f_y^\top \right) (f_x - \mu_{i(x)}) \\ &\leq (f_x - \mu_{i(x)})^\top \left(\sum_{y \in S \setminus C_{i(x)}} f_y f_y^\top \right) (f_x - \mu_{i(x)}) + \mu_{i(x)}^\top \left(\sum_{y \in S \setminus C_{i(x)}} f_y f_y^\top \right) \mu_{i(x)} \\ &\leq 3 \cdot 10^{-10} \|f_x\|_2^4 \qquad \text{根据公式（65）和（66）。} \end{aligned} \tag{67}$$

结合公式（64）、（65）、（66）和（67）即可得到该结论。 □

第二个性质可由下文的命题 D.2 以及以下假设立即得出： x 相对于 S 是典型的（依据定义 4.13）。

命题 D.2。 如果 $x \in V \setminus B_\delta$ 满足 $\mu_{i(x)}^\top \mu_{i(x)} \bullet \sum_{y \in S} f_y^\top f_y \leq O(\varphi^2/\epsilon)^{1/3} \cdot \|\mu_{i(x)}\|_2^4$ 以及 $(f_x - \mu_{i(x)})(f_x - \mu_{i(x)})^\top \bullet \sum_{y \in S} f_y f_y^\top \leq 10^{-10} \|f_x\|_2^4$，那么

$$f_x^\top f_x \bullet \sum_{y \in S} f_y f_y^\top \leq O(\varphi^2/\epsilon)^{1/3} \cdot \|f_x\|_2^4.$$

证明。我们有

$$
\begin{aligned}
f_x^\top f_x \bullet \sum_{y\in S} f_y f_y^\top &= f_x^\top \left(\sum_{y\in S} f_y f_y^\top\right) f_x \\
&= (f_x-\mu_{i(x)})^\top \left(\sum_{y\in S} f_y f_y^\top\right)(f_x-\mu_{i(x)})^\top + (f_x-\mu_{i(x)})\left(\sum_{y\in S} f_y f_y^\top\right)\mu_{i(x)} \qquad (68)\\
&+ \mu_{i(x)}^\top \left(\sum_{y\in S} f_y f_y^\top\right)(f_x-\mu_{i(x)}) + \mu_{i(x)}^\top \left(\sum_{y\in S} f_y f_y^\top\right)\mu_{i(x)}.
\end{aligned}
$$

为了给公式（68）中的第一项作出界，我们使用假设

$$
(f_x-\mu_{i(x)})(f_x-\mu_{i(x)})^\top \bullet \sum_{y\in S} f_y f_y^\top \le 10^{-10}\|f_x\|_2^4. \qquad (69)
$$

为了给公式（68）中的最后一项作出界，我们使用假设

$$
\mu_{i(x)}^\top \mu_{i(x)} \bullet \sum_{y\in S} f_y^\top f_y \le O(\varphi^2/\epsilon)^{1/3}\cdot\|\mu_{i(x)}\|_2^4 \le O(\varphi^2/\epsilon)^{1/3}\cdot\|f_x\|_2^4, \qquad \text{by Remark 3.8.} \qquad (70)
$$

为了给公式（68）中的第二项和第三项作出界，我们利用 $\sum_{y\in S} f_y f_y^\top$ 是半正定矩阵这一事实，因此

$$
\begin{aligned}
&(f_x-\mu_{i(x)})^\top \left(\sum_{y\in S} f_y f_y^\top\right)\mu_{i(x)} + \mu_{i(x)}^\top \left(\sum_{y\in S} f_y f_y^\top\right)(f_x-\mu_{i(x)}) \\
&\le (f_x-\mu_{i(x)})^\top \left(\sum_{y\in S} f_y f_y^\top\right)(f_x-\mu_{i(x)}) + \mu_{i(x)}^\top \left(\sum_{y\in S} f_y f_y^\top\right)\mu_{i(x)} \qquad (71)\\
&\le O\left(\left(\varphi^2/\epsilon\right)^{1/3}\right)\|f_x\|_2^4 \qquad \text{根据公式（69）和（70）。}
\end{aligned}
$$

结合公式（68）、（69）、（70）和（71）即可得到所述结论。 □

□

D.1 命题 4.20 的证明

本节中，我们证明命题 4.20。为方便读者，下面重新陈述该命题。

命题 4.20. $\Pr_\sigma[\mathcal{E}_{\text{equal}}] \le 0.5$.

证明. 我们有

$$
\Pr_\sigma[\mathcal{E}_{\text{equal}}] = \mathbb{1}\{|S\cap C_i| \text{ is even}\}\cdot\binom{|S\cap C_{i(x)}|}{\frac{|S\cap C_{i(x)}|}{2}} 2^{-|S\cap C_{i(x)}|}.
$$

为了界定右侧，我们分别考虑 $|S\cap C_i|=2$ 和 $|S\cap C_i|>2$ 这两种情况。
情况 a： $|S\cap C_i|=2$。于是 $\binom{|S\cap C_{i(x)}|}{|S\cap C_{i(x)}|/2}\cdot 2^{-|S\cap C_{i(x)}|} = 2\cdot 2^{-4} = 0.5$。

情况 b： $|S\cap C_i|>2$。只有当 $|S\cap C_i|$ 为偶数时，概率才非零，因此我们可以假设$|S\cap C_i|\ge 4$。利用以下斯特林公式的表述（由 [Rob55] 给出）

$$
n! = \sqrt{2\pi}n^{n+1}e^{-n}e^{r_n},
$$

其中 $r_n \in \left[\frac{1}{12n+1}, \frac{1}{12n}\right]$，我们得到对于所有 $n\ge 4$，$\binom{n}{n/2} \le \frac{2^{n+1}\sqrt{2}}{\sqrt{\pi\cdot n}} < 0.5\cdot 2^n$。令 $n=|S\cap C_i|$，即可证明该结论。

□

D.2 Al几乎所有簇都能被 S 良好表示（引理 4.28 的证明)

本节我们证明引理 4.28。我们首先证明引理 D.3，对相对于 S 而言不典型的顶点数量进行界定。

引理 D.3。若 S 是从 V中独立均匀随机采样得到的 $O(k\log(\varphi^2/\epsilon))$ 个顶点组成的（多重）集合，那么以至少 0.999 的概率（关于 S的随机性），有

$$|\{x \in S : x \text{ is not } \textit{typical with respect to } S\}| \le O\left(k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log^2(\varphi^2/\epsilon)\right).$$

证明。回顾定义 4.13，相对于 S 的典型顶点由 5 个条件定义。对于每个条件，我们定义 S 中所有违反该条件的点所构成的集合，并分别对每个集合的大小进行界定。令

$$F_1 \coloneqq \{x \in S : S \cap B_\delta \cap C_{i(x)} \neq \emptyset\}$$

$$F_2 \coloneqq \left\{x \in S : |S \cap C_{i(x)}| \ge 10^{-5}\delta^{-1/2}\right\}$$

$$F_3 \coloneqq \left\{x \in S : p_S^t(x) \ge 10^5 \cdot \left(\frac{\varphi^2}{\epsilon}\right)^{1/3} \cdot \frac{(t_\Delta+1)}{|S|} \cdot \frac{k}{n} \text{ for some } t \in [2t_{\min}, 2t_{\min}+2t_\Delta]\right\}$$

$$F_4 \coloneqq \left\{x \in S : \left\langle p_x^t, \left(p_S^\ell\right)^2\right\rangle \ge 10^5 \cdot \left(\frac{\varphi^2}{\epsilon}\right)^{1/3} \cdot \frac{(t_\Delta+1)^4}{|S|^2} \cdot \left(\frac{k}{n}\right)^{3/2} \log(\varphi^2/\epsilon)\right.$$
$$\left. \text{or } \left\langle \left(p_x^t\right)^2, p_S^\ell\right\rangle \ge 10^5 \cdot \left(\frac{\varphi^2}{\epsilon}\right)^{1/3} \cdot \frac{(t_\Delta+1)^4}{|S|} \cdot \left(\frac{k}{n}\right)^{3/2} \log(\varphi^2/\epsilon) \text{ for some } t,\ell \in [t_{\min}, t_{\min}+t_\Delta]\right\}$$

$$F_5 \coloneqq \left\{x \in S : (f_x - \mu_{i(x)})(f_x - \mu_{i(x)})^\top \bullet \sum_{y \in S} f_y f_y^\top \ge 10^{-10}\|f_x\|_2^4\right\}$$

分别表示定义 4.13 中未满足条件 **(1)**、**(2)**、**(3)**、**(4)** 和**(5)** 的顶点集合。那么

$$|\{x \in V : \ x \text{ 相对于 S}\}| \le |F_1| + |F_2| + |F_3| + |F_4| + |F_5|, \text{ 并非典型顶点}$$

因此，只需证明 $|F_1|,\ |F_2|,\ |F_3|,\ |F_4|,\ |F_5| \le O\left(k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log^2(\varphi^2/\epsilon)\right)$。

现在我们将证明，对每个 $i \in [5]$, $|F_i| \le O\left(k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log^2(\varphi^2/\epsilon)\right)$，其成立的概率至少为 0.9992。

因此，利用并集界即可得到该引理。

在下文的所有证明中，我们将 $S = \{x_1, \ldots, x_L\}$ 视为一个大小固定为$L = O(k \cdot \log(\varphi^2/\epsilon))$ 的随机（多重）集合；我们将 $x_1, \ldots, x_L$ 视为在 V上均匀分布的随机变量。

界定 $|F_1|$： 稍微滥用一下记号， F_1 是一个随机变量，用于统计违反约束 $S \cap B_\delta \cap C_{i(x_l)} \neq \emptyset$的元素 $x_l \in$ 的数量。我们的目标是证明

$$\Pr_S\left[|F_1| \le O\left(k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log^2(\varphi^2/\epsilon)\right)\right] \ge 0.9992.$$

我们通过证明，对于高常数概率事件 $\mathcal{E}$，$\Pr_S\left[|F_1| \le O\left(k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log^2(\varphi^2/\epsilon)\right) \middle| \mathcal{E}\right]$ 有一个常数下界来实现这一点。

高概率事件 $\mathcal{E}$。设 α 为一个充分小的常数。设

$$\mathcal{E} \coloneqq \left\{ |S \cap B_\delta| \le \alpha \cdot k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log(\varphi^2/\epsilon) \right\}.$$

由于如引理 3.6 $|B_\delta| \le O(\frac{1}{\delta} \cdot \epsilon/\varphi^2 \cdot n)$ 所示，我们有

$$\mathop{\mathbb{E}}_{S}[|S \cap B_\delta|] \le O\left(k \cdot \log(\varphi^2/\epsilon)\frac{\epsilon}{\delta\varphi^2}\right).$$

根据设定 $\delta = \Omega((\epsilon/\varphi^2)^{2/3})$（依照定义 4.3），由马尔可夫不等式

$$1 - \Pr_S[\mathcal{E}] = \Pr_S\left[S \cap B_\delta \ge O\left(k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log(\varphi^2/\epsilon)\right)\right] \le 0.0001.$$

为什么对 $\mathcal{E}$ 进行条件化会有所帮助。我们可以尝试用以下朴素论证来界定 $|F_1|$。若 $\mathcal{E}$ 成立，则 $|S \cap B_\delta| \le \alpha \cdot k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log(\varphi^2/\epsilon)$，因此满足 $S \cap B_\delta \cap C_i \ne \emptyset$ 的簇最多有 $\alpha \cdot k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log(\varphi^2/\epsilon)$ 个 C_i。由于对每个 i，都有 $|C_i| \le O(n/k)$，我们得到

$$|F_1| \le O\left(\left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log(\varphi^2/\epsilon) \cdot n\right).$$

虽然这个论证是有效的，但由此得到的上界太大。我们将利用该论证中提出的思路，对每个 $x_l \in S$ 界定 $\Pr_S\left[S \cap B_\delta \cap C_{i(x_l)} \ne \emptyset \middle| \mathcal{E}\right]$ ——然后通过马尔可夫不等式得到 $|F_1|$ 的所需上界。

界定 $\Pr_S\left[S \cap B_\delta \cap C_{i(x_l)} \ne \emptyset \middle| \mathcal{E}\right]$ 对所有 l。固定 $l \in [L]$ 并定义 $S \backslash x_l \coloneqq \{x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_L\}$。注意到

$$\begin{aligned} \Pr_S[\mathcal{E}] \cdot \Pr_S\left[S \cap B_\delta \cap C_{i(x_l)} \ne \emptyset \middle| \mathcal{E}\right] = \Pr_S\left[S \cap B_\delta \cap C_{i(x_l)} \ne \emptyset \text{ and } \mathcal{E}\right] \le \\ \Pr_S\left[S \cap B_\delta \cap C_{i(x_l)} \ne \emptyset \text{ and } (S \setminus x_l) \cap B_\delta \le \alpha \cdot k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log(\varphi^2/\epsilon)\right] \le \\ \Pr_{x_l}[x_l \in B_\delta] + \Pr_S\left[(S \setminus x_l) \cap B_\delta \cap C_{i(x_l)} \ne \emptyset \text{ and } (S \setminus x_l) \cap B_\delta \le \alpha \cdot k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log(\varphi^2/\epsilon)\right]. \end{aligned} \tag{72}$$

这里，第一个不等式成立是因为 $S \backslash x_l \subset S$。最后一个不等式成立是因为 $S \cap B_\delta \cap C_{i(x_l)} \ne \emptyset$ 当且仅当 x_l 是从 B_δ 中采样得到，或者 x_l 是从与 $(S \setminus x_l) \cap B_\delta$ 有非空交集的某个簇中采样得到。第一个概率可以界定为

$$\Pr_{x_l}[x_l \in B_\delta] = \frac{|B_\delta|}{n} \le O\left(\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\right).$$

注意，如果 $(S \setminus x_l) \cap B_\delta \le \alpha \cdot k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log(\varphi^2/\epsilon)$，那么与 $(S \setminus x_l) \cap B_\delta$ 有非空交集的簇数量不会超过 $\alpha \cdot k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log(\varphi^2/\epsilon) \ll k$。因此，在这种情况下，$(S \setminus x_l) \cap B_\delta \cap C_{i(x_l)} \ne \emptyset$ 当且仅当 x_l 是从这个簇的小子集中采样时：

$$\begin{aligned} \Pr_S\left[(S \setminus x_l) \cap B_\delta \le \alpha \cdot k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log(\varphi^2/\epsilon) \text{ and } (S \setminus x_l) \cap B_\delta \cap C_{i(x_l)} \ne \emptyset\right] \le \\ \alpha \cdot k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log(\varphi^2/\epsilon) \cdot \max_i \frac{|C_i|}{n} \le O\left(\left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log(\varphi^2/\epsilon)\right). \end{aligned}$$

最后，将这些界代入公式（72），我们得到

$$\Pr_S\left[S\cap B_\delta\cap C_{i(x_l)}\neq\emptyset\middle|\mathcal{E}\right]\le\frac{O\left(\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\cdot\log(\varphi^2/\epsilon)\right)}{\Pr_S[\mathcal{E}]}.$$

由于 $\Pr_S[\mathcal{E}]\ge 0.9999$,

$$\Pr_S\left[S\cap B_\delta\cap C_{i(x_l)}\neq\emptyset\middle|\mathcal{E}\right]\le O\left(\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\cdot\log(\varphi^2/\epsilon)\right).$$

推导 $|F_1|$**的界**。由于上述界对每个 $x_l\in S$ 都成立，而 F_1 恰好统计了满足 $S\cap B_\delta\cap C_{i(x_l)}\neq\emptyset$ 的 x_l 数量,

$$\mathop{\mathbb{E}}_S[|F_1||\mathcal{E}]\le O\left(\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\cdot\log(\varphi^2/\epsilon)\cdot|S|\right)=O\left(k\cdot\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\cdot\log^2(\varphi^2/\epsilon)\right).$$

根据马尔可夫不等式，对于足够大的常数 C

$$\Pr_S\left[|F_1|\le C\cdot k\cdot\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\cdot\log^2(\varphi^2/\epsilon)\middle|\mathcal{E}\right]\ge 0.9999$$

最后,

$$\Pr_S\left[|F_1|\le O\left(k\cdot\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\cdot\log^2(\varphi^2/\epsilon)\right)\right]\ge\Pr_S\left[|F_1|\le O\left(k\cdot\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\cdot\log^2(\varphi^2/\epsilon)\right)\middle|\mathcal{E}\right]\cdot\Pr_S[\mathcal{E}]\ge 0.9992$$

如所需。

界定 $|F_2|$：我们将证明，以至少 0.9992 的概率，有 $|F_2|\le O\left(k\cdot(\epsilon/\varphi^2)^{1/3}\cdot\log^2(\varphi^2/\epsilon)\right)$.

固定 $l\in[L]$。我们用记号 $S\setminus x_l$ 表示集合 $\{x_1,\dots,x_{l-1},x_{l+1},\dots,x_L\}$。由于 $|S|=O(k\cdot\log(\varphi^2/\epsilon))$（根据定义 4.9 中的条件（1）），因此至多存在 $\frac{|S|}{10^{-6}\delta^{-1/2}}=O(k\cdot\delta^{1/2}\cdot\log(\varphi^2/\epsilon))$ 个不同的簇 C_i 满足 $|(S\setminus x_l)\cap C_i|\ge 10^{-6}\delta^{-1/2}$。因此,

$$\Pr_{x_l}\left[(S\setminus x_l)\cap C_{i(x_l)}\ge 10^{-6}\delta^{-1/2}\right]\le O\left(k\cdot\delta^{1/2}\cdot\log(\varphi^2/\epsilon)\right)\cdot\max_i\frac{|C_i|}{n}$$

对于随机变量 $S\setminus x_l$ 的每一种取值，结论都成立。因此,

$$\begin{aligned}\Pr_S\left[|S\cap C_{i(x_l)}|\ge 10^{-5}\delta^{-1/2}\right]&\le\Pr_{S\setminus x_l}\left[\Pr_{x_l}\left[(S\setminus x_l)\cap C_{i(x_l)}\ge 10^{-6}\delta^{-1/2}\right]\mid S\setminus x_l\right]\\&=O\left(k\cdot\delta^{1/2}\cdot\log(\varphi^2/\epsilon)\right)\cdot\max_i\frac{|C_i|}{n}\\&=O\left(\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\cdot\log(\varphi^2/\epsilon)\right)\qquad\text{since }\delta=O((\epsilon/\varphi^2)^{2/3}),\end{aligned}$$

其中，第一个不等式成立是因为 $|S\cap C_{i(x_l)}|\ge 10^{-5}\delta^{-1/2}$ 蕴含 $(S\setminus x_l)\cap C_{i(x_l)}\ge 10^{-6}\delta^{-1/2}$。这是因为 $\delta=O((\epsilon/\varphi^2)^{2/3})$（根据定义 4.3），尤其是，$\delta$ 小于一个充分小的常数。

由于上述不等式对每个 $x_l\in S$ 都成立，且 $|F_2|$ 统计了满足 $|S\cap C_{i(x_l)}|\ge 10^{-5}\delta^{-1/2}$ 的 x_l 的数量，因此我们得到

$$\mathop{\mathbb{E}}_{S}[|F_2|] \le O\left(k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \log^2(\varphi^2/\epsilon)\right).$$

根据马尔可夫不等式，以至少 0.9992 的概率，我们得到

$$|F_2| \le O\left(k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log^2(\varphi^2/\epsilon)\right)$$

正如所需。

界定 $|F_3|$： 给定 t，令

$$F_{3,t} \coloneqq \left\{x \in S : p_S^t(x) \ge 10^5 \cdot \left(\frac{\varphi^2}{\epsilon}\right)^{1/3} \cdot \frac{t_\Delta + 1}{|S|} \cdot \frac{k}{n}\right\}.$$

根据 F_3 的定义，$F_3 = \cup_{t\in[2t_{\min},2t_{\min}+2t_\Delta]} F_{3,\,t}$，而且分别界定所有 $F_{3,\,t}$ 会更容易。回顾式（15），有

$$p_S^t(x) = \frac{1}{|S|}\sum_{y\in S} M^t \mathbb{1}_y(x) = \frac{1}{|S|}\left\langle \mathbb{1}_x, \sum_{y\in S} M^t \mathbb{1}_y\right\rangle = \frac{1}{|S|}\left\langle M^{t_1}\mathbb{1}_x, \sum_{y\in S} M^{t_2}\mathbb{1}_y\right\rangle$$

其中 $t_1 = \lceil t/2\rceil$, $t_2 = \lfloor t/2\rfloor$。

定义一个随机变量

$$Z \coloneqq \sum_{l=1}^{L}\left\langle M^{t_1}\mathbb{1}_{x_l}, \sum_{j=1}^{L} M^{t_2}\mathbb{1}_{x_j}\right\rangle.$$

注意，如果将 S 视为随机变量，则 $Z = |S|\cdot\sum_{x\in S} p_S^t(x)$。对 Z 的一个良好上界将使我们能够论证，不可能有太多 $x \in S$ 对应的 $p_S^t(x)$ 很大。由于集合 $F_{3,\,t}$ 由所有满足 $p_S^t(x)$ 超过某个阈值的 $x \in S$ 组成，这将推出 $|F_{3,\,t}|$ 的一个上界。

注意到

$$\mathop{\mathbb{E}}_{S}[Z] = \sum_{l=1}^{L}\mathop{\mathbb{E}}_{x_l}\left[\left\langle M^{t_1}\mathbb{1}_{x_l}, M^{t_2}\mathbb{1}_{x_l}\right\rangle\right] + \sum_{j,l\in[L]:l\ne j}\mathop{\mathbb{E}}_{x_l,x_j}\left[\left\langle M^{t_1}\mathbb{1}_{x_l}, M^{t_2}\mathbb{1}_{x_j}\right\rangle\right], \tag{73}$$

并且每个求和中的所有期望都相等。因此，

$$\mathop{\mathbb{E}}_{S}[Z] = |S|\cdot\mathop{\mathbb{E}}_{x_1}\left[\left\langle M^{t_1}\mathbb{1}_{x_1}, M^{t_2}\mathbb{1}_{x_1}\right\rangle\right] + |S|\cdot(|S|-1)\cdot\mathop{\mathbb{E}}_{x_1,x_2}\left[\left\langle M^{t_1}\mathbb{1}_{x_1}, M^{t_2}\mathbb{1}_{x_2}\right\rangle\right].$$

对于所有 x 和所有 $t \ge t_{\min}$，根据引理 B.1，有 $\|M^t\mathbb{1}_x\|_2^2 \le O(k/n)$，因此

$$\mathop{\mathbb{E}}_{x_1}\left[\left\langle M^{t_1}\mathbb{1}_{x_1}, M^{t_2}\mathbb{1}_{x_1}\right\rangle\right] \le \mathop{\mathbb{E}}_{x_1}\left[\|M^{t_1}\mathbb{1}_{x_1}\|_2^2 + \|M^{t_2}\mathbb{1}_{x_1}\|_2^2\right] \le O\left(\frac{k}{n}\right).$$

为界定方程 (73) 中的第二项，注意对于 x_1 的每一种取值，我们都有 $\mathbb{E}_{x_2}[\langle M^{t_1}\mathbb{1}_{x_1}, M^{t_2}\mathbb{1}_{x_2}\rangle] = \frac{1}{n}\sum_{v\in V}\langle M^{t_1}\mathbb{1}_{x_1}, M^{t_2}\mathbb{1}_v\rangle = \frac{1}{n}\langle M^{t_1}\mathbb{1}_{x_1}, \mathbb{1}\rangle = \frac{1}{n}$，因此 $\mathbb{E}_{x_1,\,x_2}[\langle M^{t_1}\mathbb{1}_{x_1}, M^{t_2}\mathbb{1}_{x_2}\rangle] = \frac{1}{n}$。综合可得

$$\mathop{\mathbb{E}}_{S}[Z] \le O\left(|S|\cdot\frac{k}{n}\right) + \frac{|S|^2}{n} \le O\left(\frac{k^2\log^2(\varphi^2/\epsilon)}{n}\right).$$

因此，对 S 的采样而言，以至少 $1 - \frac{0.0004}{t_\Delta+1}$ 的概率，根据马尔可夫不等式，我们得到

$$Z = O\left(\frac{k^2\log^2(\varphi^2/\epsilon)\cdot(t_\Delta+1)}{n}\right).$$

对于每个 $l \in [L]$，令 $Z_l \coloneqq \left\langle M^t \mathbb{1}_{x_l},\ \sum_{j=1}^{L} M^t \mathbb{1}_{x_j} \right\rangle$，并注意到 $Z = \sum_{l=1}^{L} Z_l$。在已采样得到满足 $Z \le O\left(k^2 \cdot (t_\Delta+1) \cdot \frac{}{n}{}^{2}\, \varphi^2/\epsilon\right)^{\log()}$ 的集合 S 这一事件的条件下，除至多 $O\left(k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \log^2(\varphi^2/\epsilon)\,(t_\Delta+1)\right)$ 个随机变量 $x_l \in S$ 外，所有随机变量的实现都满足

$$\left\langle M^{t_1} \mathbb{1}_{x_l}, \sum_{j=1}^{L} M^{t_2} \mathbb{1}_{x_j} \right\rangle \le 10^5 \cdot \left(\frac{\varphi^2}{\epsilon}\right)^{1/3} \cdot (t_\Delta+1)^2 \cdot \frac{k}{n},$$

这等价于写作 $|F_{3,\ t}| \le O\left(k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{\frac{1}{3}\ \text{对数}^2\ \varphi^2/\epsilon}_{t_\Delta+1}\right)^{()}$。

最后，回顾一下 $F_3 = \cup_{t\in[2t_{\min},2t_{\min}+2t_\Delta]} F_{3,\ t}$。根据并集界，以至少 $1 - \frac{0.0004}{t_\Delta+1} \cdot 2(t_\Delta+1) = 0.9992$ 的概率，有 $|F_{3,\ t}| \le O\left(k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3\ \text{对数}^2\ \varphi^2/\epsilon}_{t_\Delta+1}\right)^{()}$ 对所有 $t \in [2t_{\min}, 2t_{\min}+2t_\Delta]$ 均成立，因此

$$|F_3| \le O\left(k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \log^2(\varphi^2/\epsilon)\right).$$

界定 $|F_4|$： 对于所有 $t,\ u \in [t_{\min}, t_{\min}+t_\Delta]$ 定义 $F_{4,\ t,\ \ell}$：

$$F_{4,t,\ell} = \Bigg\{ x \in S : \sum_{v\in V} (M^t \mathbb{1}_x(v))^2 \sum_{y\in S} M^\ell \mathbb{1}_y(v) + \sum_v M^t \mathbb{1}_x(v) \sum_{y,y'\in S} M^\ell \mathbb{1}_y(v) M^\ell \mathbb{1}_{y'}(v) \ge 10^5 \cdot \left(\frac{\varphi^2}{\epsilon}\right)^{1/3} (t_\Delta+1)^4 \left(\frac{k}{n}\right)^{3/2} \log(\varphi^2/\epsilon) \Bigg\}.$$

注意，$F_4 \subseteq \cup_{t,\ \ell\in[t_{\min},t_{\min}+t_\Delta]} F_{4,\ t,\ \ell}$。为验证这一陈述，先回顾公式（15）：对于每个 $v \in V$, $p_S^t(v) = \frac{1}{|S|} \sum_{y\in S} M^t \mathbb{1}_y(v)$，都有 $p_S^t(v)^2 = \frac{1}{|S|^2} \sum_{y,\ y'\in S} M^t \mathbb{1}_y(v) M^t \mathbb{1}_{y'}(v)$。若 $x \in F_4$，则存在某些 t, $\ell \in [t_{\min}, t_{\min}+t_\Delta]$，使得下列情况之一成立：

$$\frac{1}{|S|^2} \sum_{v\in V} M^t \mathbb{1}_x(v) \sum_{y,y'\in S} M^\ell \mathbb{1}_y(v) M^\ell \mathbb{1}_{y'}(v) = \left\langle p_x^t, \left(p_S^\ell\right)^2 \right\rangle \ge 10^5 \cdot \left(\frac{\varphi^2}{\epsilon}\right)^{1/3} \cdot \frac{(t_\Delta+1)^4}{|S|^2} \cdot \left(\frac{k}{n}\right)^{3/2} \log(\varphi^2/\epsilon)$$

or

$$\frac{1}{|S|} \sum_{v\in V} (M^t \mathbb{1}_x(v))^2 \sum_{y\in S} M^\ell \mathbb{1}_y(v) = \left\langle \left(p_x^t\right)^2, p_S^\ell \right\rangle \ge 10^5 \cdot \left(\frac{\varphi^2}{\epsilon}\right)^{1/3} \cdot \frac{(t_\Delta+1)^4}{|S|} \cdot \left(\frac{k}{n}\right)^{3/2} \log(\varphi^2/\epsilon).$$

因此，根据 $F_{4,\ t,\ \ell}$ 的定义，我们得到 $x \in F_{4,\ t,\ \ell}$。

定义随机变量

$$Z \coloneqq \sum_{l\in[L]} \sum_v \sum_{j\ne j'\in[L]} M^t \mathbb{1}_{x_l}(v) M^\ell \mathbb{1}_{x_j}(v) M^\ell \mathbb{1}_{x_{j'}}(v)$$

and

$$Z' \coloneqq \sum_{l,j\in[L]} \sum_{v\in V} \left[(M^t \mathbb{1}_{x_l}(v))^2 M^\ell \mathbb{1}_{x_j}(v) + M^t \mathbb{1}_{x_l}(v) (M^\ell \mathbb{1}_{x_j}(v))^2 \right].$$

注意，如果我们将 S 视为一个随机变量，那么 $Z + Z' = \sum_{x\in S} \left(|S|^2 \cdot \langle p_x^t, \left(p_S^\ell\right)^2 \rangle + |S| \cdot \langle \left(p_x^t\right)^2,\ p_S^\ell \rangle \right)$。对 Z 和 Z' 给出良好的上界，将使我们能够论证：不存在太多的 $x \in S$ 使得相应的内积很大。由于集合 $F_{4,\ t,\ \ell}$ 由所有相应内积超过某个阈值的 $x \in S$ 构成，这将意味着 $|F_{4,\ t,\ \ell}|$ 的一个上界。

我们首先给出 Z 的一个上界。首先，我们将构成 Z 的各项分成两组：一组中 l，j，j' 都是不同的指标；另一组中，允许 l 与 j 或 j' 中的任一个相同。

$$\begin{aligned}\mathop{\mathbb{E}}_{S}[Z] = 2 \sum_{l\neq j\in[L]} \mathop{\mathbb{E}}_{x_l,x_j}\left[\sum_v M^t\mathbb{1}_{x_l}(v)M^\ell\mathbb{1}_{x_l}(v)M^\ell\mathbb{1}_{x_j}(v)\right] \\ + \sum_{\substack{l,j,j'\in[L]:\\ l\neq j\neq j'\neq l}} \mathop{\mathbb{E}}_{x_l,x_j,x_{j'}}\left[\sum_v M^t\mathbb{1}_{x_l}(v)M^\ell\mathbb{1}_{x_j}(v)M^\ell\mathbb{1}_{x_{j'}}(v)\right]\end{aligned}$$

注意到每个求和式中的期望都相等。因此，

$$\mathop{\mathbb{E}}_{S}[Z] \le 2|S|^2\cdot \mathop{\mathbb{E}}_{x_1,x_2}\left[\sum_v M^t\mathbb{1}_{x_1}(v)M^\ell\mathbb{1}_{x_1}(v)M^\ell\mathbb{1}_{x_2}(v)\right] + |S|^3\cdot \mathop{\mathbb{E}}_{x_1,x_2,x_3}\left[\sum_v M^t\mathbb{1}_{x_1}(v)M^\ell\mathbb{1}_{x_2}(v)M^\ell\mathbb{1}_{x_3}(v)\right]$$

注意，对于每个 v

$$\mathop{\mathbb{E}}_{x_1}\left[\mathop{\mathbb{E}}_{x_2}\left[\mathop{\mathbb{E}}_{x_3}\left[M^t\mathbb{1}_{x_1}(v)M^\ell\mathbb{1}_{x_2}(v)M^\ell\mathbb{1}_{x_3}(v)\right]\right]\right] = \frac{1}{n}\mathop{\mathbb{E}}_{x_1}\left[\mathop{\mathbb{E}}_{x_2}\left[M^t\mathbb{1}_{x_1}(v)M^\ell\mathbb{1}_{x_2}(v)\right]\right] = \frac{1}{n^2}\mathop{\mathbb{E}}_{x_1}\left[M^t\mathbb{1}_{x_1}(v)\right] = \frac{1}{n^3},$$

因此第二项 等于 $\frac{|S|^3}{n^3}$。上述所有变换成立是因为 $\sum_{x\in V} M^t\mathbb{1}_x = M^t\mathbb{1} = \mathbb{1}$。Next,

$$\mathop{\mathbb{E}}_{x_2}\left[\sum_v M^t\mathbb{1}_{x_1}(v)M^\ell\mathbb{1}_{x_1}(v)M^\ell\mathbb{1}_{x_2}(v)\right] = \frac{1}{n}\sum_v M^t\mathbb{1}_{x_1}(v)M^\ell\mathbb{1}_{x_1}(v) = \frac{1}{n}\left\langle M^t\mathbb{1}_{x_1}, M^\ell\mathbb{1}_{x_1}\right\rangle,$$

并且 $\mathbb{E}_{x_1}[\langle M^t\mathbb{1}_{x_1}, M^\ell\mathbb{1}_{x_1}\rangle] \le \mathbb{E}_{x_1}[\|M^t\mathbb{1}_{x_1}\|_2\cdot\|M^\ell\mathbb{1}_{x_1}\|_2] \le O\left(\frac{k}{n}\right)$，这可由引理 B.1 得出。因此，

$$\mathop{\mathbb{E}}_{S}[Z] \le \frac{2|S|^2\cdot k}{n^2} + \frac{|S|^3}{n^3} \le O\left(\frac{k^3\log^3(\varphi^2/\epsilon)}{n^2}\right).$$

类似地，我们对 Z' 进行界定。首先，我们将构成 Z' 的项分成两组：一组满足 $l=j$，另一组满足 $l\neq j$。

$$\begin{aligned}\mathop{\mathbb{E}}_{S}[Z'] &= \sum_{l\in[L]}\mathop{\mathbb{E}}_{x_l}\left[\sum_{v\in V}(M^t\mathbb{1}_{x_l}(v))^2M^\ell\mathbb{1}_{x_l}(v) + \sum_{v\in V}M^t\mathbb{1}_{x_l}(v)(M^\ell\mathbb{1}_{x_l}(v))^2\right] \\ &+ \sum_{l\neq j\in[L]}\mathop{\mathbb{E}}_{x_l,x_j}\left[\sum_{v\in V}(M^t\mathbb{1}_{x_l}(v))^2M^\ell\mathbb{1}_{x_j}(v) + \sum_{v\in V}M^t\mathbb{1}_{x_l}(v)(M^\ell\mathbb{1}_{x_j}(v))^2\right] \\ &\le O\left(\frac{|S|\cdot\sqrt{k}}{\sqrt{n}}\right)\cdot\mathop{\mathbb{E}}_{x_l}\left[\|M^t\mathbb{1}_{x_l}\|_2^2 + \|M^\ell\mathbb{1}_{x_l}\|_2^2\right] + |S|^2\cdot\left(\frac{\mathbb{E}_{x_l}\left[\sum_v(M^t\mathbb{1}_{x_l}(v))^2\right]}{n} + \frac{\mathbb{E}_{x_l}\left[\sum_v(M^\ell\mathbb{1}_{x_l}(v))^2\right]}{n}\right) \\ &\le O\left(\frac{|S|\cdot k^{3/2}}{n^{3/2}}\right) + O\left(\frac{|S|^2\cdot k}{n^2}\right),\end{aligned}$$

其中第三行由 $M^t\mathbb{1}_x(v) \le \|M^t\mathbb{1}_x\|_2 \le O(\sqrt{k/n})$ 得出，最后一行则由引理 B.1 得出，即对所有 x 和所有 $t\ge t_{\min}$，$\mathbb{E}_{x_l}[\sum_v(M^\ell\mathbb{1}_{x_l}(v))^2] = \mathbb{E}_{x_l}[\|M^t\mathbb{1}_{x_l}\|_2^2] \le O(k/n)$ 均成立。由此，由于 $|S| = O(k\log(\varphi^2/\epsilon))$，

$$\mathop{\mathbb{E}}_{S}[Z+Z'] \le O\left(\frac{k^3\log^3(\varphi^2/\epsilon)}{n^2}\right) + O\left(\frac{|S|\cdot k^{3/2}}{n^{3/2}}\right) \le O\left(\frac{k^{5/2}\log^3(\varphi^2/\epsilon)}{n^{3/2}}\right)$$

因为 $k/n\le 1$. 以至少 $1-\frac{0.001}{(t_\Delta+1)^2}$ 的概率，在集合 S 上进行采样，根据马尔可夫不等式

$$Z+Z' \le O\left(\frac{k^{5/2}\cdot(t_\Delta+1)^2\cdot\log^3(\varphi^2/\epsilon)}{n^{3/2}}\right).$$

注意，$Z+Z'=\sum_{l=1}^{L}Z_{x_l}+Z'_{x_l}$其中 $Z_{x_l}:=\sum_v\sum_{j,j'\in[L]}M^t\mathbb{1}_{x_l}(v)M^\ell\mathbb{1}_{x_j}(v)M^\ell\mathbb{1}_{x_{j'}}(v)$ 且$Z'_{x_l}:=\sum_{v\in V}\sum_{l\in[L]}\left[(M^t\mathbb{1}_{x_l}(v))^2M^\ell\mathbb{1}_{x_2}(v)+M^t\mathbb{1}_x(v)(M^\ell\mathbb{1}_y(v))^2\right]$。假设我们采样得到一个集合 S，使得 $Z+Z'\le C\cdot k^{5/2}\cdot(t_\Delta+1)^2\cdot\log^3(\varphi^2/\epsilon)\,n^{3/2}$，其中某个常数 C 足够大。那么，除至多 $C\cdot10^{-5}\cdot k\cdot\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\log^2(\varphi^2/\epsilon)\,(t_\Delta+1)^2$ 个随机变量 $x_l\in S$（这里稍微滥用一下记号，也记作 x_l）之外，所有随机变量的取值都满足

$$\sum_v(M^t\mathbb{1}_{x_l}(v))^2\sum_{j\in[L]}M^\ell\mathbb{1}_{x_j}(v)+\sum_vM^t\mathbb{1}_{x_l}(v)\sum_{j,j'\in[L]}M^\ell\mathbb{1}_{x_j}(v)M^\ell\mathbb{1}_{x_{j'}}(v)$$
$$\le10^5\cdot\left(\frac{\varphi^2}{\epsilon}\right)^{1/3}(t_\Delta+1)^4\left(\frac{k}{n}\right)^{3/2}\log(\varphi^2/\epsilon).$$

等价地，$|F_{4,\ t,\ \ell}|\le O\left(k\cdot\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\text{对数}^2(\varphi^2/\epsilon)\,(t_\Delta+1)^2\right)$。最后，回顾一下，$F_4=\cup_{t,\ \ell\in[t_{\min},t_{\min}+t_\Delta]}F_{4,\ t,\ \ell}$。以至少 $\left(1-\frac{0.001}{t_\Delta+1}()^2\right)^{t_\Delta^2}\ge0.9992$ 的概率，有$|F_{4,\ t,\ \ell}|\le O\left(k\cdot\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\text{对数}^2(\varphi^2/\epsilon)\,(t_\Delta+1)^2\right)$对所有$t,\ \ell\in[t_{\min},t_{\min}+t_\Delta]$都成立，因此

$$|F_4|\le O\left(k\cdot\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\log^2(\varphi^2/\epsilon)\right).$$

界定 $|F_5|$： 定义一个随机变量 Z:

$$Z=\sum_{l\in[L]}(f_{x_l}-\mu_{i(x_l)})(f_{x_l}-\mu_{i(x_l)})^\top\bullet\sum_{j\in[L]\setminus l}f_{x_j}f_{x_j}^\top$$

于是,

$$\mathop{\mathbb{E}}_S[Z]=\sum_{l\neq j\in[L]}\mathop{\mathbb{E}}_{x_l,x_j}\left[(f_{x_l}-\mu_{i(x_l)})(f_{x_l}-\mu_{i(x_l)})^\top\bullet f_{x_j}f_{x_j}^\top\right]$$

注意到求和中的所有期望都相等。因此,

$$\begin{aligned}\mathop{\mathbb{E}}_S[Z]&\le|S|^2\cdot\mathop{\mathbb{E}}_{x_1,x_2}\left[(f_{x_1}-\mu_{i(x_1)})(f_{x_1}-\mu_{i(x_1)})^\top\bullet f_{x_2}f_{x_2}^\top\right]\\&=\frac{|S|^2}{n^2}\sum_{x\in V}(f_x-\mu_{i(x)})(f_x-\mu_{i(x)})^\top\bullet\sum_{y\in V}f_yf_y^\top\\&=\frac{|S|^2}{n^2}\sum_{x\in V}\|f_x-\mu_{i(x)}\|_2^2&\text{根据引理 C.2}\\&\le O\left(\frac{\epsilon k}{\varphi^2}\cdot\frac{k^2\log^2(\varphi^2/\epsilon)}{n^2}\right)&\text{by Equation (13) in Lemma 3.5.}\end{aligned}$$

根据马尔可夫不等式，对于集合 S 的抽样，其概率至少为 0.9999

$$Z\le C\cdot\frac{\epsilon k}{\varphi^2}\cdot\frac{k^2\log^2(\varphi^2/\epsilon)}{n^2}$$

对于某个足够大的常数 C。注意，$Z=\sum_{l=1}^{L}Z_{x_l}$ 其中 $Z_{x_l}:=(f_{x_l}-\mu_{i(x_l)})(f_{x_l}-\mu_{i(x_l)})^\top\bullet\sum_{j\in[L]\setminus l}f_{x_j}f_{x_j}^\top$。假设我们已采样了一个集合 S，使得 $Z\le C\cdot\frac{\epsilon k}{\varphi^2}\cdot k^2\log^2(\varphi^2/\epsilon)n^2$。那么，除至多 $\frac{C}{c}\cdot k\cdot\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\log(\varphi^2/\epsilon)$ 个以充分小的常数 c 为参数的随机变量 $x_l\in S$ 外，所有随机变量的实现——这里稍微滥用记号，也记为 x_l ——都满足

$$(f_{x_l}-\mu_{i(x_l)})(f_{x_l}-\mu_{i(x_l)})^\top \bullet \sum_{j\in[L]\setminus l} f_{x_j}f_{x_j}^\top \le \frac{c\cdot k^2\cdot(\epsilon/\varphi^2)^{2/3}\cdot\log(\varphi^2/\epsilon)}{n^2}. \tag{74}$$

根据引理 4.12，S 以至少 0.9999 的概率具有良好分布，因此根据定义 4 的条件（**1**）。 9

$$|S\cap B_\delta| \le O\left(k\cdot(\epsilon/\varphi^2)^{1/3}\cdot\log(\varphi^2/\epsilon)\right).$$

定义

$$D \coloneqq \left\{ x_l\in S : x_l\in B_\delta \text{ or } (f_{x_l}-\mu_{i(x_l)})(f_{x_l}-\mu_{i(x_l)})^\top \bullet \sum_{j\in[L]\setminus l} f_{x_j}f_{x_j}^\top > \frac{c\cdot k^2\cdot(\epsilon/\varphi^2)^{2/3}\cdot\log(\varphi^2/\epsilon)}{n^2} \right\}.$$

以至少 0.9998 的概率，$|D| \le O\left(k\cdot\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\log^2(\varphi^2/\epsilon)\right)$。接下来只需证明 $F_5\subseteq D$。事实上，对于每个 $x\notin D$

$$\begin{aligned}(f_x-\mu_{i(x)})(f_x-\mu_{i(x)})^\top \bullet \sum_{y\in S} f_yf_y^\top &\le \|f_x-\mu_{i(x)}\|_2^2\cdot\|f_x\|_2^2 + c\cdot\frac{k^2\cdot(\epsilon/\varphi^2)^{2/3}\cdot\log(\varphi^2/\epsilon)}{n^2}\\ &< \frac{\sqrt{\delta}k}{n}\cdot\|f_x\|_2^2 + c\cdot\frac{k^2\cdot(\epsilon/\varphi^2)^{2/3}\cdot\log(\varphi^2/\epsilon)}{n^2} \le 10^{-10}\|f_x\|_2^4,\end{aligned}$$

因此 $x\notin F_5$。这里我们用到：对于每个 $x\in V\setminus B_\delta$，都有 $\|f_x-\mu_{i(x)}\|_2^2 \le \frac{\delta k}{n}$，且根据定义 3.4 和注记 3.8，有 $\|f_x\|_2^2=\Theta(k/n)$。根据设定 $\delta=\Theta((\epsilon/\varphi^2)^{2/3})$（见定义 4.3），$\delta$ 小于任意充分小的常数。

□

利用引理 D.3，我们现在可以证明引理 4.28，下面为方便读者重新陈述该引理。

引理 4.28。如果从 V 中独立均匀随机采样得到 $O(k\log(\varphi^2/\epsilon))$ 个顶点组成的 S（多重）集合，那么以至少 0.9998 的概率，有

$$|\{i\in[k] : C_i \textit{ is not well-represented by } S\}| \le O\left(k\cdot\left(\frac{\epsilon}{\varphi^2}\right)^{1/3}\cdot\log(\varphi^2/\epsilon)\right).$$

证明。令 $\mathcal{F}\coloneqq\{i\in[k]:\ C_i \text{ is} \textit{not} \text{ well-represented by } S\}$ 表示未满足定义 4.27 的簇集合。现在我们将分别界定违反定义 4.27 中两个条件的簇数。令

$\mathcal{F}_1\coloneqq\{i\in[k]:\ S\cap C_i=\emptyset\}$ 和 $\mathcal{F}_2\coloneqq\{i\in[k]:\ \exists x\in S\cap C_i$ 表示 x 相对于 $S\}$ 不典型的情况

分别表示违反条件（**1**）和（**2**）的簇集合。于是 $\mathcal{F}=\mathcal{F}_1\cup\mathcal{F}_2$。

首先，我们界定 $|\mathcal{F}_1|$。下面的断言可以得到这一结果。

断言 D.4. 若 $|S|\ge 100k\log(\varphi^2/\epsilon)\cdot\eta$，则以至少 0.9999 的概率，有 $|\mathcal{F}_1|\le O(k\cdot\epsilon/\varphi^2$)

证明。令 S 为一个至少包含 $100k\log(\varphi^2/\epsilon)\cdot\eta$ 个顶点的（多重）集合，这些顶点从 V 中独立地均匀随机采样得到。固定一个簇 C_i，并定义随机变量 $X=|S\cap C_i|$，其随机性取决于 S。于是 X 是 $|S|$ 个独立伯努利随机变量之和，每个变量的均值为 $\frac{|C_i|}{n}$，因此有 $\mu\coloneqq\mathbb{E}[X]=|S|\cdot\frac{|C_i|}{n}\ge 100k\log(\varphi^2/\epsilon)\cdot\eta\cdot\frac{1/\eta}{k}=100\log(\varphi^2/\epsilon)$。令 δ 由方程 $(1-\delta)\mu=1$ 定义，即 $\delta=1-\frac{1}{\mu}$。注意，由于当 ϵ/φ^2 足够小时有 $\mu\ge 2$，所以 $\delta\ge 1/2$。应用切尔诺夫界可得

$$\begin{aligned}\Pr[|S\cap C_i|<1] &\le \Pr[X\le(1-\delta)\mu]\\ &\le e^{-\delta^2\mu/2}\\ &\le e^{-\mu/8} \qquad\qquad \text{since } \delta\ge\frac{1}{2}\\ &\le e^{-100\log(\varphi^2/\epsilon)/8}\\ &\le 0.0001\epsilon/\varphi^2.\end{aligned} \tag{75}$$

对于每个 $i \in [k]$，定义示性随机变量 $Y_i \coloneqq \mathbb{1}\{|C_i \cap S| < 1\}$，并令 $Y \coloneqq \sum_{i\in[k]} Y_i$。注意， $Y = |\mathcal{F}_1|$。根据（75），我们有

$$\mathbb{E}[Y] = \sum_{i=1}^{k} \Pr[Y_i = 1] \le 0.0001k \cdot \epsilon/\varphi^2.$$

因此，根据马尔可夫不等式，

$$\Pr[Y \ge k \cdot \epsilon/\varphi^2] \le 0.0001,$$

这给出了关于 $|\mathcal{F}_1|$ 的所需界。 □

接下来，我们需要给出 $|\mathcal{F}_2|$ 的界。根据引理 D.3，以 0.9999 的概率，有

$$|\{x \in S : \text{x is } \textit{not} \text{ typical with respect to S}\}| \le O\left(k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log(\varphi^2/\epsilon)\right).$$

在此事件条件下，我们有

$$|\mathcal{F}_2| \le |\{x \in S : \text{x is } \textit{not} \text{ typical with respect to S}\}| \le O\left(k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log(\varphi^2/\epsilon)\right).$$

为完成证明，对 $\mathcal{F}_1$ 和 $\mathcal{F}_2$ 对应的失败事件取并集界。由此可知，以至少 0.9998 的概率，有

$$|\mathcal{F}| \le |\mathcal{F}_1| + |\mathcal{F}_2| \le O(k \cdot \epsilon/\varphi^2) + O\left(k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log(\varphi^2/\epsilon)\right) = O\left(k \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/3} \cdot \log(\varphi^2/\epsilon)\right).$$

□

E 用随机游走近似 $\left\langle f_x, \sum_y \sigma_y f_y \right\rangle$（引理 4.21 的证明）

在本节中，我们证明碰撞计数引理（引理 4.21 将在下文重述），该引理表明，$\left\langle M^{t_1}\mathbb{1}_x, \sum_{y\in S}\sigma_u \mathbb{1}_y \right\rangle$ 可以通过运行少量随机游走来估计。该引理假设，在 p_S^t 下分配给 x 的概率质量与均匀分布所分配的概率质量相近（条件**(1)**），并且 x 的 t 步分布与 p_S^t 弱相关（条件**(2)**）。注意，我们考虑了两种不同长度的随机游走 t_1, t_2。这是因为我们将在第 4.4 节的引理 4.22 中使用该引理，来论证 $\left\langle p(M)\mathbb{1}_x, \sum_{y\in S} p(M)\sigma_y \mathbb{1}_y \right\rangle = \sum_{t_1,t_2} c_{t_1} c_{t_2} \langle M^{t_1}\mathbb{1}_x, M^{t_2}\mathbb{1}_y\rangle$ 的方差。

假设我们从 x 进行 Q 次长度为 t_1 的惰性随机游走，并从每个 $y \in S$ 进行 R 次长度为 t_2 的惰性随机游走。令 $\widehat{p}_x^{t_1}$ 表示从 x 出发的随机游走的经验分布；对于每个 $y \in S$，令 $\widehat{p}_y^{t_1}$ 表示从 y 出发的随机游走的经验分布。那么我们有如下结论：

引理 4.21（碰撞计数）。设 β、$\gamma \ge 1$ 为参数，设 $x \in V$、$S \subseteq V$ 和 t_1、$t_2 \ge 1$。假设满足以下条件：

(1) $p_S^{t_1+t_2}(x) \le \beta \cdot \frac{1}{n} \cdot \frac{k}{|S|}$，

(2) $\left\langle p_x^{t_1}, \left(p_S^{t_2}\right)^2 \right\rangle \le \gamma \cdot \frac{1}{n^2} \cdot \frac{k^2}{|S|^2}$ *and* $\left\langle \left(p_x^{t_1}\right)^2, p_S^{t_2} \right\rangle \le \gamma \cdot \frac{1}{n^2} \cdot \frac{k^2}{|S|}$.

对于每个 $\rho > 0$（期望失败概率）、$\xi > 0$（期望精度）和 $\sigma \in \{-1,1\}^S$（符号赋值），如果

$$Q \cdot R \ge \frac{7\beta}{\rho \cdot \xi^2} \cdot \frac{n}{k} \text{ and } Q, R \ge \frac{7(\gamma+\beta^2)}{\rho \cdot \xi^2}$$

那么，

$$\Pr_{\textit{random walks}}\left[\left|\left\langle \widehat{p}_x^{t_1}, \sum_{y\in S}\sigma_y \widehat{p}_y^{t_2}\right\rangle - \left\langle M^{t_1}\mathbb{1}_x, \sum_{y\in S}\sigma_y M^{t_2}\mathbb{1}_y\right\rangle\right| \le \xi \cdot \frac{k}{n}\right] \ge 1-\rho.$$

证明。令 $W_x^1,\dots,W_x^Q \sim M^{t_1}\mathbb{1}_x$ 表示从 x 出发的 Q 随机游走的端点；对每个 $y\in S$，令 $W_y^1,\dots,W_y^R \sim M^{t_2}\mathbb{1}_y$ 表示从 y 出发的 R 随机游走的端点。那么 $\widehat{p}_x^{t_1} = \frac{1}{R}\sum_{r=1}^R W_x^r$ 并且对每个 $y\in S$，$\widehat{p}_y^{t_2} = \frac{1}{Q}\sum_{q=1}^Q W_y^q$。

我们有

$$\mathbb{E}[W_x^q(v)] = M^{t_1}\mathbb{1}_x(v), \qquad \mathbb{E}[W_y^r(v)] = M^{t_2}\mathbb{1}_y(v).$$

Let

$$Z = \frac{1}{Q\cdot R}\sum_{q=1}^{Q}\sum_{r=1}^{R}\sum_{y\in S}\sum_{v\in V}\sigma_y W_x^q(v)W_y^r(v) = \left\langle \widehat{p}_x^{t_1}, \sum_{y\in S}\sigma_y \widehat{p}_y^{t_2}\right\rangle$$

表示有符号碰撞率的随机变量。我们有

$$\mathbb{E}[Z] = \frac{1}{Q\cdot R}\sum_{v,q,r,y}\sigma_y\,\mathbb{E}[W_x^q(v)]\,\mathbb{E}[W_y^r(v)] = \sum_{v,y}\sigma_y M^{t_1}\mathbb{1}_x(v)M^{t_2}\mathbb{1}_y(v) = \left\langle M^{t_1}\mathbb{1}_x, \sum_y \sigma_y M^{t_2}\mathbb{1}_y\right\rangle,$$

$$\begin{aligned}\mathbb{E}[Z]^2 &= \frac{1}{Q^2\cdot R^2}\sum_{q,q',r,r',y,y',v,u}\sigma_y\sigma_{y'}\,\mathbb{E}[W_x^q(v)]\,\mathbb{E}[W_y^r(v)]\,\mathbb{E}[W_x^{q'}(u)]\,\mathbb{E}[W_{y'}^{r'}(u)]\\ &= \left(\sum_y \sigma_y \left\langle M^{t_1}\mathbb{1}_x, M^{t_2}\mathbb{1}_y\right\rangle\right)^2,\end{aligned} \tag{76}$$

and

$$\begin{aligned}\mathbb{E}[Z^2] &= \frac{1}{Q^2\cdot R^2}\,\mathbb{E}\left[\left(\sum_{q,r,y,v}\sigma_y W_x^q(v)W_y^r(v)\right)\left(\sum_{q',r',y',u}\sigma_{y'}W_x^{q'}(u)W_{y'}^{r'}(u)\right)\right]\\ &= \frac{1}{Q^2\cdot R^2}\sum_{q,q',r,r',y,y',v,u}\sigma_y\sigma_{y'}\,\mathbb{E}\left[W_x^q(v)W_y^r(v)W_x^{q'}(u)W_{y'}^{r'}(u)\right].\end{aligned} \tag{77}$$

现在我们将对方差 $\mathsf{Var}[Z] = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2$ 给出上界。具体做法是仔细地对公式（77）右侧给出的 $\mathbb{E}[Z^2]$ 表达式中的每一项进行上界估计。根据 $\{q,q',(r,y)$ 、$(r',y')\}$ 的大小对各项分类会很有用。这是因为，对于满足 $|\{q,q',(r,y)$、$(r',y')\}| = 4$ 的项，四个变量 $W_x^q(v)$、 $W_y^r(v)$、 $W_x^{q'}(u)$ 和 $W_{y'}^{r'}(u)$ 相互独立，因此 $\mathbb{E}\left[W_x^q(v)W_y^r(v)W_x^{q'}(u)W_{y'}^{r'}(u)\right] = \mathbb{E}[W_x^q(v)]\,\mathbb{E}[W_y^r(v)]\,\mathbb{E}[W_x^{q'}(u)]\,\mathbb{E}[W_{y'}^{r'}(u)],$，我们可以将这些项与 $\mathbb{E}[Z]^2$ 抵消。另一方面，满足 $|\{q,q',(r,y)$、$(r',y')\}| < 4$ 的项无法与 $\mathbb{E}[Z]^2$ 抵消，因此我们将利用引理陈述中的条件对它们进行上界估计。

对以下项进行界定： $|\{q,q',(r,\ y),\ (r',\ y')\}| = 2$**：** 那么 $q = q'$、 $r = r'$ 和 $y = y'$，因此我们只需考虑 $u = v$ 和 $u \neq v$ 的情形。我们将利用引理陈述中的条件（**1**）来界定这些项。由于我们处理的是形如 $M^{t_1}\mathbb{1}_x$ 和 $M^{t_2}\mathbb{1}_y$ 的表达式，将该条件改写成以下等价形式会更方便：$\left\langle M^{t_1}\mathbb{1}_x,\ \sum_{y\in S}M^{t_2}\mathbb{1}_y\right\rangle \le \beta\cdot\frac{k}{n}$。现在，我们来界定公式（77）中满足 $|\{q,q',(r,\ y),\ (r',\ y')\}| = 2$ 的项所作出的贡献。

- 满足 $u = v,\ y = y'$、 $q = q',\ r = r'$ 的项所作出的贡献

$$\begin{aligned}\frac{1}{Q^2\cdot R^2}\sum_{v,y,q,r}\mathbb{E}\left[W_x^q(v)^2W_y^r(v)^2\right] &= \frac{1}{Q^2\cdot R^2}\cdot QR\sum_v\sum_y M^{t_1}\mathbb{1}_x(v)M^{t_2}\mathbb{1}_y(v)\\ &= \frac{1}{QR}\sum_y\left\langle M^{t_1}\mathbb{1}_x M^{t_2}\mathbb{1}_y\right\rangle\\ &= \frac{1}{QR}\cdot\beta\cdot\frac{k}{n} && \text{by Condition (}\mathbf{1}\text{)}\\ &\le \frac{\rho\cdot\xi^2}{7}\cdot\frac{k^2}{n^2} && \text{根据对 } Q,R \text{ 的假设。}\end{aligned}$$

• 包含 $u \neq v, y = y', q = q', r = r'$ 的项贡献为 0，因为同一个随机游走不可能在两个不同的端点结束。

对 $|\{q, q', (r, y), (r', y')\}| = 3$ 相关项进行界定：我们将利用引理陈述中的条件（**2**）来界定这些项。由于我们处理的是形如 $M^{t_1}\mathbb{1}_x$和 $M^{t_2}\mathbb{1}_y$的表达式，方便起见，我们将该条件改写为以下等价形式：

$$\sum_v \sum_{y,y' \in S} M^{t_1}\mathbb{1}_x(v) M^{t_2}\mathbb{1}_y(v) M^{t_2}\mathbb{1}_{y'}(v) \leq \gamma \cdot \frac{k^2}{n^2} \text{ and } \sum_v \sum_{y \in S} M^{t_1}\mathbb{1}_x(v)^2 M^{t_2}\mathbb{1}_y(v) \leq \gamma \cdot \frac{k^2}{n^2}.$$

现在我们 界定公式（77）中 $|\{q, q', (r, y), (r', y')\}| = 3$这些项的贡献 .

• 与 $u = v, y = y', q \neq q', r = r'$ 相关的项的贡献为

$$\begin{aligned}
\frac{1}{Q^2 \cdot R^2} \sum_{v,y,q \neq q',r} \mathbb{E}\left[W_x^q(v) W_x^{q'}(v) W_y^r(v)^2\right] &\leq \frac{1}{Q^2 \cdot R^2} \cdot Q^2 R \sum_v \sum_y M^{t_1}\mathbb{1}_x(v)^2 M^{t_2}\mathbb{1}_y(v) \\
&\leq \frac{1}{R} \cdot \gamma \cdot \frac{k^2}{n^2} && \text{by Condition (}\mathbf{2}\text{)} \\
&\leq \frac{\rho \cdot \xi^2}{7} \cdot \frac{k^2}{n^2} && \text{by the assumption on } R.
\end{aligned}$$

• 与 $u = v, y = y', q = q', r \neq r'$ 相关的项的贡献为

$$\begin{aligned}
\frac{1}{Q^2 \cdot R^2} \sum_{v,y,q,r \neq r'} \mathbb{E}\left[W_x^q(v)^2 W_y^r(v) W_y^{r'}(v)\right] &\leq \frac{1}{Q^2 \cdot R^2} \cdot Q R^2 \sum_v \sum_y M^{t_1}\mathbb{1}_x(v) M^{t_2}\mathbb{1}_y(v)^2 \\
&\leq \frac{1}{Q} \cdot \gamma \cdot \frac{k^2}{n^2} && \text{by Condition (}\mathbf{2}\text{)} \\
&\leq \frac{\rho \cdot \xi^2}{7} \cdot \frac{k^2}{n^2} && \text{by the assumption on } Q.
\end{aligned}$$

• 与 $u = v, y \neq y', q = q'$ 相关的项的贡献为

$$\begin{aligned}
\frac{1}{Q^2 \cdot R^2} \sum_{v,y \neq y',q,r,r'} \sigma_y \sigma_{y'} \mathbb{E}\left[W_x^q(v)^2 W_y^r(v) W_{y'}^{r'}(v)\right] &= \frac{1}{Q} \sum_{y,y':y \neq y'} \sigma_y \sigma_{y'} \sum_v M^{t_1}\mathbb{1}_x(v) M^{t_2}\mathbb{1}_y(v) M^{t_2}\mathbb{1}_{y'}(v) \\
&\leq \frac{1}{Q} \sum_{y,y':y \neq y'} \sum_v M^{t_1}\mathbb{1}_x(v) M^{t_2}\mathbb{1}_y(v) M^{t_2}\mathbb{1}_{y'}(v) \\
&\leq \frac{1}{Q} \cdot \gamma \cdot \frac{k^2}{n^2} && \text{by Condition (}\mathbf{2}\text{)} \\
&\leq \frac{\rho \cdot \xi^2}{7} \cdot \frac{k^2}{n^2} && \text{by the assumption on } R.
\end{aligned}$$

• 如果 $q = q'$，或 $(r, y) = (r', y')$，则带有 $u \neq v$ 的项贡献为 0，因为同一个随机游走不可能以两个不同的端点结束。

对带有 $|\{q, q', (r, y)、(r', y')\}| = 4$ 的项进行界定：</style}带有 (, y)、(r′, y′)}| = 4 的大多数项会与 $\mathbb{E}[Z^2]$抵消。不过，会剩下一些无法与 $\mathbb{E}[Z^2]$抵消的低阶小项。我们将利用引理陈述中的条件证明，这些剩余项的上界为 $O\left(\rho \cdot \xi^2 \cdot \frac{k^2}{n^2}\right)$。

• 含有 $u = v, y = y'、q \neq q', r \neq r'$ 的项会产生贡献

$$\frac{1}{Q^2 \cdot R^2} \sum_{v,y,q,r \neq r'} \mathbb{E}\left[W_x^q(v) W_x^{q'}(v) W_y^r(v) W_y^{r'}(v)\right] \leq \sum_y \sum_v M^{t_1}\mathbb{1}_x(v)^2 M^{t_2}\mathbb{1}_y(v)^2$$

- 含有 $u=v,\ y\neq y'、\ q\neq q'$ 的项会产生贡献

$$
\begin{aligned}
&\frac{1}{Q^2\cdot R^2}\sum_{v,y\neq y',q\neq q',r,r'}\sigma_y\sigma_{y'}\,\mathbb{E}\left[W_x^q(v)W_x^{q'}(v)W_y^r(v)W_{y'}^{r'}(v)\right]\\
&=\frac{Q-1}{Q}\sum_{y,y':y\neq y'}\sigma_y\sigma_{y'}\sum_v M^{t_1}\mathbb{1}_x(v)^2M^{t_2}\mathbb{1}_y(v)M^{t_2}\mathbb{1}_{y'}(v)\\
&\le\sum_{y,y':y\neq y'}\sigma_y\sigma_{y'}\sum_v M^{t_1}\mathbb{1}_x(v)^2M^{t_2}\mathbb{1}_y(v)M^{t_2}\mathbb{1}_{y'}(v)+\frac{1}{Q}\sum_{y,y':y\neq y'}\sum_v M^{t_1}\mathbb{1}_x(v)^2M^{t_2}\mathbb{1}_y(v)M^{t_2}\mathbb{1}_{y'}(v)\\
&\le\sum_{y,y':y\neq y'}\sigma_y\sigma_{y'}\sum_v M^{t_1}\mathbb{1}_x(v)^2M^{t_2}\mathbb{1}_y(v)M^{t_2}\mathbb{1}_{y'}(v)+\frac{1}{Q}\sum_{y,y'}\sum_v M^{t_1}\mathbb{1}_x(v)M^{t_2}\mathbb{1}_y(v)M^{t_2}\mathbb{1}_{y'}(v)\\
&\le\sum_{y,y':y\neq y'}\sigma_y\sigma_{y'}\sum_v M^{t_1}\mathbb{1}_x(v)^2M^{t_2}\mathbb{1}_y(v)M^{t_2}\mathbb{1}_{y'}(v)+\frac{\rho\cdot\xi^2}{7}\cdot\frac{k^2}{n^2}\quad\text{by Condition (}\mathbf{2}\text{) and assumption on }Q.
\end{aligned}
$$

- 包含 $u\neq v$和 $q\neq q'$ 以及 $(r,y)\neq(r',q')$ 的项贡献

$$
\begin{aligned}
&\frac{1}{Q^2\cdot R^2}\sum_{u\neq v,q\neq q',(r,y)\neq(r',y')}\sigma_y\sigma_{y'}\,\mathbb{E}\left[W_x^q(v)W_x^{q'}(u)W_y^r(v)W_{y'}^{r'}(u)\right]\\
&=\frac{Q-1}{Q}\cdot\frac{R^2-1}{R^2}\sum_{y,y'}\sigma_y\sigma_{y'}\sum_{u,v:u\neq v}M^{t_1}\mathbb{1}_x(v)M^{t_1}\mathbb{1}_x(u)M^{t_2}\mathbb{1}_y(v)M^{t_2}\mathbb{1}_{y'}(u)\\
&\le\sum_{y,y'}\sigma_y\sigma_{y'}\sum_{u,v:u\neq v}M^{t_1}\mathbb{1}_x(v)M^{t_1}\mathbb{1}_x(u)M^{t_2}\mathbb{1}_y(v)M^{t_2}\mathbb{1}_{y'}(u)\\
&+2\left(\frac{1}{Q}+\frac{1}{R^2}\right)\sum_{y,y'}\sum_{u,v:u\neq v}M^{t_1}\mathbb{1}_x(v)M^{t_1}\mathbb{1}_x(u)M^{t_2}\mathbb{1}_y(v)M^{t_2}\mathbb{1}_{y'}(u)\\
&\le\sum_{y,y'}\sigma_y\sigma_{y'}\sum_{u,v:u\neq v}M^{t_1}\mathbb{1}_x(v)M^{t_1}\mathbb{1}_x(u)M^{t_2}\mathbb{1}_y(v)M^{t_2}\mathbb{1}_{y'}(u)+2\left(\frac{1}{Q}+\frac{1}{R^2}\right)\left\langle M^{t_1}\mathbb{1}_x,\sum_{y\in S}M^{t_2}\mathbb{1}_y\right\rangle^2\\
&\le\sum_{y,y'}\sigma_y\sigma_{y'}\sum_{u,v:u\neq v}M^{t_1}\mathbb{1}_x(v)M^{t_1}\mathbb{1}_x(u)M^{t_2}\mathbb{1}_y(v)M^{t_2}\mathbb{1}_{y'}(u)+2\left(\frac{1}{Q}+\frac{1}{R^2}\right)\cdot\beta^2\cdot\frac{k^2}{n^2}\qquad\text{by Condition (}\mathbf{1}\text{)}\\
&\le\sum_{y,y'}\sigma_y\sigma_{y'}\sum_{u,v:u\neq v}M^{t_1}\mathbb{1}_x(v)M^{t_1}\mathbb{1}_x(u)M^{t_2}\mathbb{1}_y(v)M^{t_2}\mathbb{1}_{y'}(u)+\frac{2\rho\cdot\xi^2}{7}\cdot\frac{k^2}{n^2}\qquad\text{by assumption on }Q,R.
\end{aligned}
$$

合并所有项将所有不同类型的项合并后，我们得到

$$
\begin{aligned}
\mathbb{E}[Z^2]&=7\cdot\frac{\rho\cdot\xi^2}{7}\frac{k^2}{n^2}+\sum_y\sum_v M^{t_1}\mathbb{1}_x(v)^2M^{t_2}\mathbb{1}_y(v)^2+\sum_{y,y':y\neq y'}\sigma_y\sigma_{y'}\sum_v M^{t_1}\mathbb{1}_x(v)^2M^{t_2}\mathbb{1}_y(v)M^{t_2}\mathbb{1}_{y'}(v)\\
&+\sum_{y,y'}\sigma_y\sigma_{y'}\sum_{u,v:u\neq v}M^{t_1}\mathbb{1}_x(v)M^{t_1}\mathbb{1}_x(u)M^{t_2}\mathbb{1}_y(v)M^{t_2}\mathbb{1}_{y'}(u)\\
&=\rho\cdot\xi^2\frac{k^2}{n^2}+\sum_{y,y',u,v}\sigma_y\sigma_{y'}M^{t_1}\mathbb{1}_x(v)M^{t_1}\mathbb{1}_x(u)M^{t_2}\mathbb{1}_y(v)M^{t_2}\mathbb{1}_{y'}(u)\\
&=\rho\cdot\xi^2\frac{k^2}{n^2}+\left\langle M^{t_1}\mathbb{1}_x,\sum_y\sigma_yM^{t_2}\mathbb{1}_y\right\rangle^2.
\end{aligned}
\tag{78}
$$

合并方程（76）和（78），我们得到

$$\begin{aligned}\mathrm{Var}[Z] &= \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 \\ &\le \rho \cdot \xi^2 \frac{k^2}{n^2} + \left\langle M^{t_1}\mathbb{1}_x, \sum_y \sigma_y M^{t_2}\mathbb{1}_y \right\rangle^2 - \left\langle M^{t_1}\mathbb{1}_x, \sum_y \sigma_y M^{t_2}\mathbb{1}_y \right\rangle^2 \\ &= \rho \cdot \xi^2 \frac{k^2}{n^2}.\end{aligned}$$

根据切比雪夫不等式，我们有

$$\Pr\left[\left|Z - \left\langle M^{t_1}\mathbb{1}_x, \sum_l \sigma_y M^{t_2}\mathbb{1}_y \right\rangle\right| > \xi \cdot \frac{k}{n}\right] \le \frac{\rho \cdot \xi^2 \frac{k^2}{n^2}}{\xi^2 \frac{k^2}{n^2}} = \rho,$$

如要求的。 □

F 定理 2.12 和定理 2.13 的证明

在本节中，我们假设 k 与 n 的距离有一个正的下界，即 $k \le n^{0.999}$。我们首先证明定理 2.13，下面将重新陈述该定理。

定理 2.13。 任何一种算法，在能够访问一个(k, φ, ϵ)-可聚类图的情况下，以至少 $2/3$ 的概率，将 k 近似到一个 $2 - \Omega(1)$ 因子以内，都必须进行至少 $\binom{n}{k}^{1/2+\Omega(\epsilon)}$ 次查询。–

定理 2.13 阐述的是这样一类算法：它们以单边 $(2 - \Omega(1))$-近似 的方式输出簇数 k:

定义 F.1 (单侧 α-近似)。 设 *OBJ* 表示某个关注量的真实值。若一个算法输出满足以下条件的单侧 α-近似算法（针对 $\alpha \ge 1$）$\hat{}$*OBJ*:

$$OBJ \;\le\; \hat{OBJ} \;\le\; \alpha \cdot OBJ.$$

定理 2.13 是以下 [CKK+18] 中结果的一个推论。

定理 F.2 ([CKK+18]中的定理 5)。 存在一个输入分布，其支撑集为

1. *YES* 情形：两个 Ω(1) 扩展图的并，每个图各有 $n/2$ 个顶点，且每个图都诱导出一个稀疏度为 ϵ 的割。
2. *NO* 情形：一个具有 n 个顶点的 Ω(1) 扩展图

因此，任何以至少 $2/3$ 的概率区分 *YES* 情形和 *NO* 情形的算法，都必须至少进行 $n^{1/2+\Omega(\epsilon)}$ 次查询。

定理 *2.13* 的证明。 令 G 为 k 个从定理 F.2 中的困难分布采样得到的实例的互不相交的相同副本的并，每个副本均取 $\binom{n}{k}$ 个顶点。于是，查询 G 与查询该困难实例一一对应。

在 YES 情形中， G 由 $\binom{n}{2k}$ 个簇组成，每个簇包含 $2k$ 个顶点；特别地，它是一个（$2k$, Ω(1) ϵ）-可聚类图。在 NO 情形中， G 由 k 个互不相交的扩展图组成，每个扩展图包含 $\binom{n}{k}$ 个顶点；特别地，G 是（k, Ω(1) ϵ）-可聚类的。因此，任何算法 $\mathcal{A}$ 若计算 G 中簇数的（$2 - \Omega(1)$）-近似值，就能区分 YES 情形和 NO 情形，因而可用于区分一个包含 n/k 个顶点的 Ω(1)-扩展图与两个各包含 $n/2k$ 个顶点的 Ω(1)-扩展图。根据定理 F.2， $\mathcal{A}$ 必须至少进行 $(n/k)^{1/2+\Omega(\epsilon)}$ 次查询。 □

接下来，我们介绍算法 8，并证明定理 2.12 中形式化表述的性能保证。

定理 2.12. 设 C 为一个足够大的常数。假设 ϵ, φ 满足注 *4.2* 中的标准假设。对于每个 $\varepsilon \geq C \cdot \left(\epsilon/\varphi^2\right)^{1/4}$，ApproxK($\varepsilon$) （算法 *8*）运行时间为

$$O^*\left(\left(\frac{n}{k}\right)^{1/2+O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))} \cdot \frac{1}{\varepsilon^2}\right)$$

并以至少 0.99 的概率，找到一个关于 k 的$(1+\varepsilon)$乘法近似。

算法 8 ApproxK(ε)

$L \leftarrow 10^{-3} \cdot \left(\frac{\varphi^2}{\epsilon}\right)^{1/2}$ $x_1, \ldots, x_L \sim$均匀分布 $\left(V^L\right)$ **对于** 所有 $l \in [L]$ **执行** $\mathrm{SK}_{l,1}, \mathrm{SK}_{l,2} \leftarrow$ 对 进行两次独立试验(x_l)，并使用更多随机游走，见推论 F.4 **结束循环** **返回** $\frac{1}{L}\sum_{l\in[L]} \langle \mathrm{SK}_{l,1}, \mathrm{SK}_{l,2} \rangle$

引理 F.3。 设 $F = x_1, x_2, \ldots, x_L \sim$均匀分布$(V^L)$，其中 $\varepsilon \geq C \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/4}$，对于足够大的常数 C 和 $L = \frac{1}{2} \cdot 10^{-3} \cdot \left(\frac{\varphi^2}{\epsilon}\right)^{1/2}$。那么，以0.999 的概率，

$$\left|\frac{1}{L}\sum_{l\in[L]} \|f_{x_l}\|_2^2 - \frac{k}{n}\right| \leq \varepsilon \cdot \frac{k}{n}.$$

证明。 回想一下，根据引理 3.6，我们有 $|B_\delta| \leq O\left(n \cdot \frac{1}{\delta} \cdot \frac{\epsilon}{\varphi^2}\right)$。令 δ 为$(\epsilon/\varphi^2)_\kappa^{1/2}$，其中常数 κ 足够大，使得 $|B_\delta| \leq n \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/2}$。定义 $\mathcal{E}$ 为事件 $\mathcal{E} := \{F \cap B_\delta = \emptyset\}$。

$$\Pr_F[\mathcal{E}] = \left(1 - \frac{|B_\delta|}{n}\right)^L \geq 1 - L \cdot \frac{|B_\delta|}{n} \geq 1 - L \cdot \left(\frac{\epsilon}{\varphi^2}\right)^{1/2} \geq 1 - \frac{1}{2} \cdot 10^{-3},$$

其中，最后一个不等式由 $L = \frac{1}{2} \cdot 10^{-3} \cdot \left(\frac{\varphi^2}{\epsilon}\right)^{1/2}$ 的选取可得。

在事件 $\mathcal{E}$ 的条件下，随机变量 $x_l \mid \mathcal{E}$ 对应于原始样本 x_l，并限制在它们位于集合 B_δ之外的情形；等价地，可以将它们视为从 $V \setminus B_\delta$ 中抽取的样本。在此条件下，集合 $\{x_l \mid \mathcal{E}\}_{l\in[L]}$ 相互独立，且在 $V \setminus B_\delta$ 上均匀分布。我们有

$$\begin{aligned}\mathop{\mathbb{E}}_{\{x_l\}|\mathcal{E}}\left[\frac{1}{L}\sum_{l=1}^{L} \|f_{x_l}\|_2^2\right] &\leq \frac{1}{|V \setminus B_\delta|} \sum_{v\in V\setminus B_\delta} \frac{1 + 4\frac{\sqrt{\epsilon}}{\varphi} + 4\sqrt{\delta}}{|C_{i(v)}|} \qquad \text{by Remark 3.8} \\ &\leq \sum_{i\in[k]} \frac{|C_i \setminus B_\delta|}{|V \setminus B_\delta|} \cdot \frac{1 + 4\frac{\sqrt{\epsilon}}{\varphi} + 4\sqrt{\delta}}{|C_i|}.\end{aligned} \tag{79}$$

同样，

$$\begin{aligned}\mathop{\mathbb{E}}_{\{x_l\}|\mathcal{E}}\left[\frac{1}{L}\sum_{l=1}^{L} \|f_{x_l}\|_2^2\right] &\geq \frac{1}{|V \setminus B_\delta|} \sum_{v\in V\setminus B_\delta} \frac{1 - 4\frac{\sqrt{\epsilon}}{\varphi} - 4\sqrt{\delta}}{|C_{i(v)}|} \qquad \text{by Remark 3.8} \\ &\geq \sum_{i\in[k]} \frac{|C_i \setminus B_\delta|}{|V \setminus B_\delta|} \cdot \frac{1 - 4\frac{\sqrt{\epsilon}}{\varphi} - 4\sqrt{\delta}}{|C_i|}.\end{aligned} \tag{80}$$

为进一步精确这些上界和下界，我们必须先对 $\sum_{i\in[k]} \frac{|C_i\setminus B_\delta|}{|C_i|}$ 进行界定。令 $C_{j_{min}}$ 为规模最小的簇。注意，

$$\sum_{i\in[k]\setminus j_{min}} \frac{|C_i|}{|C_i|} - \frac{|B_\delta|}{|C_{j_{min}}|} \le \sum_{i\in[k]} \frac{|C_i\setminus B_\delta|}{|C_i|} \le \sum_{i\in[k]} \frac{|C_i|}{|C_i|} \le k.$$

由于 $|C_{j_{min}}| \ge n/(\eta k)$，其中 $\eta \in O(1)$ 是簇大小之间的最大比值，因此，对于足够大的常数 α，我们有 $\frac{|B_\delta|}{|C_{j_{min}}|} \le \alpha \cdot k \cdot \frac{1}{\delta} \cdot \frac{\epsilon}{\varphi^2}$ 和 $|B_\delta| \le \alpha \cdot n \cdot \frac{1}{\delta} \cdot \frac{\epsilon}{\varphi^2}$。由此可得，

$$k \cdot \left(1 - \alpha \cdot \frac{1}{\delta} \cdot \frac{\epsilon}{\varphi^2}\right) \le \sum_{i\in[k]} \frac{|C_i\setminus B_\delta|}{|C_i|} \le k.$$

此外，注意到 $n \cdot \left(1 - \alpha \cdot \frac{1}{\delta} \cdot \frac{\epsilon}{\varphi^2}\right) \le |V \setminus B_\delta| \le n$。将这些界代入公式（79）和公式（80），得到

$$\frac{k}{n} \cdot \left(1 - \frac{4\sqrt{\epsilon}}{\varphi} - 4\sqrt{\delta}\right) \cdot \left(1 - \alpha \cdot \frac{1}{\delta} \cdot \frac{\epsilon}{\varphi^2}\right) \le \mathop{\mathbb{E}}_{\{x_l\}|\mathcal{E}} \left[\sum_{l=1}^{L} \frac{1}{L} \|f_{x_l}\|_2^2\right] \le \frac{k}{n} \cdot \frac{1 + \frac{4\sqrt{\epsilon}}{\varphi} + 4\sqrt{\delta}}{1 - \alpha \cdot \frac{1}{\delta} \cdot \frac{\epsilon}{\varphi^2}}.$$

由于 $\delta = \frac{(\epsilon/\varphi^2)}{\kappa}^{1/2}$，左端和右端中的其他项与 $4\sqrt{\delta}$ 相比可以忽略：

$$\frac{k}{n} \cdot \left(1 - \left(\frac{\epsilon}{\varphi^2}\right)^{1/4}\right) \le \mathop{\mathbb{E}}_{\{x_l\}|\mathcal{E}} \left[\sum_{l=1}^{L} \frac{1}{L} \|f_{x_l}\|_2^2\right] \le \frac{k}{n} \cdot \left(1 + \left(\frac{\epsilon}{\varphi^2}\right)^{1/4}\right).$$

在应用 Hoeffding 不等式之前，还需注意，对于所有 $v \in V \setminus B_\delta$，根据注记 3.8，$\|f_v\|_2^2 = \Theta(k/n)$。因此，所有独立随机变量 $\|f_{x_1}\|_2^2, \ldots, \|f_{x_l}\|_2^2$ 的取值都位于一个有界区域内。由 Hoeffding 不等式可得

$$\Pr\left[\left|\frac{1}{L}\sum_{l\in[L]} \|f_{x_l}\|_2^2 - \frac{k}{n}\right| \ge \varepsilon\frac{k}{n}\middle|\mathcal{E}\right] \le \Pr\left[\left|\frac{1}{L}\sum_{l\in[L]} \|f_{x_l}\|_2^2 - \mathop{\mathbb{E}}_{\{x_l\}|\mathcal{E}} \left[\sum_{l=1}^{L} \frac{1}{L} \|f_{x_l}\|_2^2\right]\right| \ge \left(\varepsilon - \left(\frac{\epsilon}{\varphi^2}\right)^{1/4}\right)\frac{k}{n}\middle|\mathcal{E}\right] \le e^{-\Omega(\varepsilon'^2 \cdot L),}$$

其中，为简洁起见，我们使用了记号 $\varepsilon' = \varepsilon - \left(\frac{\epsilon}{\varphi^2}\right)^{1/4} \ge \frac{1}{2}\varepsilon$。为了确保 $e^{-\Omega(\varepsilon'^2 \cdot L)} \le \frac{1}{2} \cdot 10^{-3}$，只需选择 $L \ge \frac{C'}{\varepsilon^2}$，使其为足够大的常数 C'。注意，当设置为 $L = \frac{1}{2} \cdot 10^{-3} \cdot \left(\frac{\varphi^2}{\epsilon}\right)^{1/2}$ 时，该不等式成立。最后，

$$\Pr\left[\left|\frac{1}{L}\sum_{l\in[L]} \|f_{x_l}\|_2^2 - \frac{k}{n}\right| \ge \varepsilon\frac{k}{n}\right] \le \Pr\left[\left|\frac{1}{L}\sum_{l\in[L]} \|f_{x_l}\|_2^2 - \frac{k}{n}\right| \ge \varepsilon\frac{k}{n}\middle|\mathcal{E}\right] + 1 - \Pr[\mathcal{E}] \le \frac{1}{2}\cdot 10^{-3} + \frac{1}{2}\cdot 10^{-3} \le 10^{-3}.$$

□

现在我们想说明，SKETCH(x_1)…,SKETCH(x_L) 以高概率为 $\|f_{x_1}\|_2^2, \ldots, \|f_{x_L}\|_2^2$ 提供足够好的近似。不幸的是，我们对 SKETCH（算法 3）性能的保证还不够充分—— SKETCH 仅以恒定概率为 $\|f_x\|_2^2$ 提供常数近似。

不过请回想附录 C 中引理 4.24 的证明： SKETCH的成功概率和近似质量取决于我们运行多少次随机游走来生成 SKETCH。因此，我们只通过更新随机游走次数来修改 SKETCH：

推论 F.4. 令 $p > 0$ 为期望失败概率，并设 $T = \frac{1}{p \cdot \varepsilon^2}$, $\varepsilon \geq \left(\frac{\epsilon}{\varphi^2}\right)^1$

4. 将过程SKETCH （算法 *3*）修改为运行 T 倍数量的随机游走后，其运行时间为 $O^*\left(\binom{n}{k}^{1/2+O(\epsilon/\varphi^2}\log(1/\epsilon)\cdot(1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\cdot T\right)$，返回一个支撑集大小至多为 $O^*\left(\binom{n}{k}^{1/2+O(\epsilon/\varphi^2}\log(1/\epsilon)\cdot(1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\cdot T\right)$ 的向量，并满足以下保证:

令 δ 为$(\epsilon/\varphi^2)^{1/2}$ 与引理 *F.3* 中一样。对于每个 $x \in V \setminus B_\delta$，以至少 $1-p$ 的概率（关于 SKETCH的内部随机性），有

$$\left|\langle \textsc{Sketch}(x), \textsc{Sketch}(x)\rangle - \|f_x\|_2^2\right| \leq \varepsilon\frac{k}{n}.$$

上述每个 $\textsc{Sketch}(x)$都是一个独立实例。

证明。 $\textsc{Sketch}(x)$ 的运行时间比之前高 T 倍，因为我们唯一做出的调整就是随机游走的数量。类似地，$\textsc{Sketch}(x)$ 返回的向量，其支撑集至多比之前大 T 倍： $O^*\left(\binom{n}{k}^{1/2+O(\epsilon/\varphi^2}\log(1/\epsilon)\cdot(1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\cdot T\right)$。

性能保证的证明沿用了引理 4.24 的证明。与该证明一样，我们使用引理 4.21，并取参数设置 $\beta = O(1)$、$\gamma = O\left(\sqrt{\frac{n}{k}}\right)$（与引理 4.24 中相同），以及不同的 $\rho = \frac{p}{(t_\Delta+1)^2}$, $\xi = \frac{\varepsilon}{|c_{t_1}||c_{t_2}|(t_\Delta+1)^2} = O^*\left(\frac{\varepsilon}{1/\epsilon^{()t_\Delta}}\right)$，其中 c_{t_1} 和 c_{t_2} 分别是定理 2.10 中多项式 $p(x)$ 里 x^{t_1} 和 x^{t_2} 的系数。回顾引理 4.21，其中 R, Q 是随机游走的数量。取 $R, Q = O\left(\sqrt{\frac{n}{k}}\cdot\frac{1}{\rho\cdot\xi^2}\right) = O^*\left(\sqrt{\frac{n}{k}}\cdot\left(\frac{1}{\epsilon}\right)^{O(t_\Delta}\cdot\frac{1}{p\cdot\varepsilon^2}\right)$，而非之前使用的 $R, Q = O^*\left(\sqrt{\frac{n}{k}}\cdot\left(\frac{1}{\epsilon}\right)^{O(t_\Delta)}\right)$，并利用对所有 $x \in V \setminus B_\delta$ 都有 $\|f_x\|_2^2 = \Theta(\frac{k}{n})$ 这一事实（见注记 3.8），我们得到

$$\Pr_{\text{random walks}}\left[\left|\left\langle \sigma_x\widehat{p}_x^{t_1}, \sigma'_x\widehat{p}_x^{t_2}\right\rangle - \left\langle \sigma_x M^{t_1}\mathbb{1}_x, \sigma'_x M^{t_2}\mathbb{1}_x\right\rangle\right| \leq \frac{\varepsilon}{|c_{t_1}||c_{t_2}|(t_\Delta+1)^2}\cdot\frac{k}{n}\right] \geq 1 - \frac{p}{(t_\Delta+1)^2} \tag{81}$$

对于 SKETCH 使用的每一对随机游走长度 t_1， t_2。对 $(t_\Delta+1)^2$ 对 t_1， $t_2 \in [t_{\min}, t_{\min}+t_\Delta]$ 应用并集界，可得

$$\Pr_{\text{random walks}}\left[\left|\langle \textsc{Sketch}(x), \textsc{Sketch}(x)\rangle - \left\langle \sigma_x\sum_t c_t\cdot M^t\mathbb{1}_x, \sigma'_x\sum_t c_t\cdot M^t\mathbb{1}_x\right\rangle\right| > \varepsilon\cdot\frac{k}{n}\right] \leq p.$$

利用断言 C.1 中证明的一个界,

$$\left|\left\|\sum_t c_t M^t\mathbb{1}_x\right\|_2^2 - \|f_x\|_2^2\right| \leq \frac{2\epsilon}{\varphi^2}\|f_x\|_2^2 + n^{-8} \leq \varepsilon\frac{k}{n},$$

其中，最后一个不等式是通过令 $\varepsilon \geq \left(\frac{\epsilon}{\varphi^2}\right)^{1/4}$ 得到的。将两者结合并对 $\varepsilon \to \varepsilon/2$ 重新缩放，可得

$$\Pr_{\text{random walks}}\left[\left|\langle \textsc{Sketch}(x), \textsc{Sketch}(x)\rangle - \|f_x\|_2^2\right| > \varepsilon\cdot\frac{k}{n}\right] \leq p,$$

如所需。

□

定理 *2.12* 的证明。 定理 2.12 的证明结合了引理 F.3 和推论 F.4。由于我们将同时用到二者，记调用引理 F.3 时使用的精度参数为 ε_1，记推论 F.4 的精度参数和失败概率参数为 ε_2， p_2。

假设引理 F.3 的结论成立，并且对于所有试验对， SKETCH 推论 F.4 均成立。那么,

$$\left|\sum_{l\in[L]}\frac{1}{L}\langle \mathrm{SK}_{l,1},\mathrm{SK}_{l,2}\rangle-\frac{k}{n}\right|\le\left|\frac{1}{L}\sum_{l\in[L]}\|f_{x_l}\|_2^2-\frac{k}{n}\right|+\frac{1}{L}\sum_{l\in[L]}\left|\langle \mathrm{SK}_{l,1},\mathrm{SK}_{l,2}\rangle-\|f_{x_l}\|_2^2\right|\le\varepsilon_1\frac{k}{n}+\varepsilon_2\frac{k}{n}.$$

若令 $\varepsilon_1=\varepsilon_2=\varepsilon/2$，即可得到所需的近似保证。推论 F.4 中的条件 $\varepsilon_2\ge\left(\frac{\epsilon}{\varphi^2}\right)^{1/4}$ 是必要的，这一点可由 $\varepsilon_2=\frac{\varepsilon}{2}\ge\frac{C}{2}\cdot\left(\left(\frac{\epsilon}{\varphi^2}\right)^{1/4}\right)$ 得出。

成功概率。 算法 ApproxK 运行 $2L$ 个 Sketch的独立实例。如果每个实例以概率 $1-p_2$成功，那么所有实例均成功的概率为 $(1-p_2)^{2L}$。我们还要求引理 F.3 的陈述成立。因此，算法 8 的成功概率下界为

$$(1-p_2)^L-10^{-3}\ge 1-2p_2\cdot L-10^{-3}.$$

为了使用引理 F.3，我们需要验证 $\varepsilon_1\ge C\cdot\left(\left(\frac{\epsilon}{\varphi^2}\right)^{1/4}\right)$——这可由 ε 的类似性质推出。然后，$L=\frac{1}{2}\cdot 10^{-3}\cdot\left(\frac{\varphi^2}{\varepsilon}\right)^{1/2}$。令 $p_2=10^{-3}/(2L)$。于是 $p_2=\frac{1}{4}\left(\frac{\epsilon}{\varphi^2}\right)^{1/2}$，而且 ApproxK 具有至少为 0。999 的所需成功概率。

运行时间。 根据推论 F.4， Sketch 的每次试验需要 $O^*\left(\binom{n}{k}^{1/2+O(\epsilon/\varphi^2\log(1/\epsilon))}\cdot(1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\cdot\frac{1}{p_2\cdot\varepsilon_2^2}\right)$。计算Sketch的两次试验的内积需要 $O^*\left(\binom{n}{k}^{1/2+O(\epsilon/\varphi^2\log(1/\epsilon))}\cdot(1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\cdot\frac{1}{p_2\cdot\varepsilon_2^2}\right)$。因此， ApproxK 的总运行时间为

$$O^*\left(\left(\frac{n}{k}\right)^{1/2+O(\epsilon/\varphi^2\log(1/\epsilon))}\cdot(1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\cdot\frac{L}{p_2\cdot\varepsilon_2^2}\right)=O^*\left(\left(\frac{n}{k}\right)^{1/2+O(\epsilon/\varphi^2\log(1/\epsilon))}\cdot(1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\cdot\frac{1}{\varepsilon^2}\right)$$

□

G 定理 1.3 的证明 Sketch

本节将讨论如何修改主要分析，以得到以下重述的权衡。

定理 1.3 (聚类预言机的空间/查询权衡；非正式表述)。 对于每个 $\delta\in[0,1]$，都存在一个误分类率为 $\epsilon^{\Omega(1)}$ 的聚类预言机，它

- 的预处理时间为 $\approx\left(n^{1/2}k^{1/2}+k\cdot\binom{n}{k}^{1-\delta}\right)\cdot n^{O(\epsilon/\varphi^2\log(1/\epsilon))}$。
- 计算一个大小为 $\approx k\cdot\binom{n}{k}^{1-\delta}\cdot n^{O(\epsilon/\varphi^2\log(1/\epsilon))}$ 的数据结构
- 的查询时间为 $\approx\binom{n}{k}^{\delta}\cdot n^{O(\epsilon/\varphi^2\log(1/\epsilon))}$

因此，数据结构的大小与查询时间的乘积为 $\approx n^{1+O(\epsilon/\varphi^2\log(1/\epsilon))}$，**与簇数无关** k。

从较高层面来看，这种权衡之所以成立，是因为决定我们能多准确地估计量 $\left\langle M^t\mathbb{1}_x,\ \sum_{y\in S}\sigma_u M^t\mathbb{1}_y\right\rangle$ 的，是从一个顶点 x 出发的随机游走数量与从每个顶点 $y\in S$ 出发的随机游走数量的乘积（见引理 4.21 ）。因此，我们可以通过增加从代表顶点出发的随机游走数量来减少从查询顶点出发的随机游走数量，反之亦然。

现在我们描述所需的算法修改，然后概述分析中的相应变化。

算法描述。 不作任何修改地运行 PREPROCESSING(k)，以获得一个代表性顶点集合 R。对于每个 $y \in R$，按照定义 4.5 计算一棵草图树，但将算法 3 第 4 行中的随机游走次数改为

$$r = O^*\left((n/k)^{1-\delta} n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$$

给定查询顶点 x，在计算 SKETCH(x) 时，算法 ISCOVERED (算法 6) 应使用

$$q = O^*\left((n/k)^{\delta} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$$

算法 3 第 4 行的随机游走。

此外，修改算法 6 第 3 步中的测试。不要测试 $|\langle \text{SKETCH}(x)\ \text{SK}_j\rangle| \geq 0.5\,|\langle \text{SKETCH}(x)\ \text{SKETCH}(x)\rangle|$，而应测试 $|\langle \text{SKETCH}(x)\,\text{SK}_j\rangle| \geq 0.5\,\frac{1}{\eta}\frac{k}{n}$-这是必要的，因为当 $\delta < \frac{1}{2}$ 时，仅使用从 x 出发的 $q = O^*\left((n/k)^{\delta} \cdot n^{O(\epsilon/\varphi^2 \log(1/\epsilon))} \cdot (1/\epsilon)^{O(\log(\varphi^2/\epsilon))}\right)$ 次随机游走，不足以以足够高的精度计算 $\langle \text{SKETCH}(x)\ \text{SKETCH}(x)\rangle$。

算法的其他部分保持不变。

分析概述。 根据注 4.23，当我们运行修改后的随机游走次数时，引理 4.22 的保证对每个集合 S 以及相对于 S而言属于强典型 （根据定义 4.15）的每个顶点都成立。这意味着， SKETCH （引理 2.11）和 ISCOVERED （引理 4.25）的保证仅对强典型顶点成立（而不是对所有典型顶点成立）。因此，在定理 4.7 的证明中，应将算法可能失败的"坏"集合 B 定义为

$$B \coloneqq \{x \in V : i(x) \in \mathcal{F}\} \cup \{x \in V : x \text{ is } \textit{not strongly typical} \text{ with respect to } S\} \cup B_\delta,$$

其中 S 是算法 5 第 1 行中采样的顶点集合。换言之，集合 B 应包含所有相对于 S而言并非强典型 的顶点，而不只是那些非典型的顶点。根据引理 4.17，

$$|\{x \in V : x \text{ is } \textit{not strongly typical} \text{ with respect to } S\}| \leq O(n \cdot (\epsilon/\varphi^2)^{1/3} \log(\varphi^2/\epsilon)),$$

这与定理 4.7 中的界相同。因此，误分类率保持不变。

最后，我们讨论将算法 6 第 3 行中的阈值改为 $0.5\,\frac{1}{\eta}\frac{k}{n}$ 对分析的影响。直观地说，当 S 包含一个与 x 属于同一簇的顶点时， $|\langle \text{SKETCH}(x)\ \text{SK}_j\rangle|$ 至少为 $\|f_x\|_2^2$ ； 否则它接近于零。"接近"的程度由其他簇中顶点贡献的噪声决定，而引理 4.19 对此进行了控制。将式（**5**）中的 10^{-10} 替换为一个充分小的常数，可以使这种噪声对任意充分小的常数 c 至多为 $c\|f_x\|_2^2$ 。这一改动只会使非典型顶点的集合增加一个常数因子。因此，无须将 $|\langle \text{SKETCH}(x)\ \text{SK}_j\rangle|$ 与 $0.5\,|\langle \text{SKETCH}(x)\ \text{SKETCH}(x)\rangle|$ 比较，只需将其与 $\|f_x\|_2^2$ 的任意常数因子下界比较，例如 $0.5\,\frac{1}{\eta}\frac{k}{n}$。