arXiv:2603.11216v1 [cs.DS] 11 Mar 2026

噪声流式与分布式数据中的频率矩与失配歧义*

刘凯文 Qin Zhang 计算机科学系 计算机科学系 印第安纳大学
印第安纳大学
kaiwliu@iu.edu qzhangcs@iu.edu

2026年3月13日

摘要

我们提出了一个用于含噪数据集统计估计的新框架。在该框架中，我们重点研究频率矩（F_p）问题，并证明：只要近似比由一个依赖于数据的量来参数化，就可以在数据流模型中使用次线性空间、在协调器模型中使用次线性通信，对未知真实数据集的 F_p 进行近似。我们将这个量称为 F_p失配歧义。我们还建立了一组下界，这些下界在输入规模方面是紧的。我们的结果带来了若干有趣的见解：

– 在数据流模型中，有噪声情形下的 F_p 问题本质上比无噪声情形更难。具体而言，在无噪声情形下， F_2 可以在与输入规模相关的对数空间内得到近似；而在有噪声情形下，任何用于 F_2 的算法都需要多项式空间。– 与无噪情形形成鲜明对比的是，在协调器模型中，含噪条件下通常无法为 F_p 实现关于输入规模的多对数通信。然而，当 F_p 失配歧义低于某个阈值时，就可以实现完全独立于输入规模的通信。

1 引言

近重复项的存在长期以来一直是数据管理中的一项挑战。这类噪声可能源自多种因素，例如数据收集错误、表示不一致、重新格式化或压缩伪影，也可能源自固有的物理限制，如科学仪器的测量不精确性或量子系统的概率性。例如，

– 提交给搜索引擎或大语言模型的大量查询，其中不同的关键词组合可能表达相同的语义意图。– 带有近重复项的网页、文档或图像通常存储在多台服务器上，而我们的目标是计算这些对象的聚合统计量。

*K. Liu 部分得到了美国国家科学基金会 CCF-1844234 项目的资助；Q. Zhang 部分得到了美国国家科学基金会 CCF-1844234 项目和印第安纳大学卢迪学院教职奖学金的资助。

– 如果我们想要 以经典方式存储量子态，那么我们就必须容忍近重复项 [38] .

通常会在分析前进行彻底的数据清洗 [22, 16, 30, 15]。然而，在“大数据”场景中，这种方法并不可行，因为其目标是在不完整存储或传输数据集的情况下，计算数据集上的统计函数。

为应对这一挑战，一系列研究旨在直接对含噪数据集进行统计分析，在不显式清洗数据的情况下即时消除噪声 [36, 10, 11, 37]。然而，正如我们将在相关工作（章节1.1）中讨论的那样，现有针对含噪数据集进行统计估计的方法存在固有局限。本文提出了一种分析含噪数据集的新框架。随后，我们在两种成熟的大数据模型中研究 p阶频率矩问题（F_p）：

– 数据流模型 [1]：我们按顺序处理输入数据集 $\sigma = (\sigma_1, \ldots, \sigma_m)$中的元素，通过一遍或多遍处理来计算 $F_p(\sigma)$。主要目标是尽可能减少内存空间的使用。

– 协调器模型 [33]：我们有 k 个站点和一个中央协调者。数据集 σ 被划分到这些 k 个站点上，其中第 i 个站点持有 $\sigma^{(i)}$。这些 k 个站点和协调者希望通过通信共同计算 $F_p(\sigma)$。计算分轮进行。每一轮中，协调者向每个 k 个站点发送一条消息，随后各站点向协调者回复一条消息。最后，协调者输出答案。我们希望同时最小化通信成本和计算所需的轮次数。

我们为这两种模型设计算法，并建立相应的下界。本文得出以下结论：

1. 在噪声数据设定中，只要将近似保证参数化为一个依赖于数据的量（称为 F_p-失配歧义），就可以在数据流模型中使用次线性空间、在协调器模型中使用次线性通信量，来近似 未知的真实数据集的 F_p 。

2. 在数据流模型中，我们针对输入规模建立了有噪声情形下 F_p 的紧确空间界，表明有噪声时 F_p 问题从根本上比无噪情形更难。具体而言，尽管根据 Alon 、 Matias 和 Szegedy 的经典结果， F_2 在无噪声情形下可用对数空间近似，但在有噪声情形下，任何用于 F_2 的算法都需要多项式空间 [1], 。

3. 在协调器模型中，我们证明，与无噪声情形不同，在存在噪声时，通常不可能使 F_p 的通信量达到输入规模的多对数级。不过，当 F_p-失配歧义低于某个阈值时，可以实现与输入规模无关的通信量。

一种用于含噪数据统计估计的新框架。 我们首先介绍用于分析含噪数据的新框架。设 $\sigma = (\sigma_1, \ldots, \sigma_m)$ 是由观测项构成的含噪数据集。每个项目 σ_i 都是由底层元素 τ_i 添加未知噪声后生成的。真实元素 τ_i 取自隐藏全集 $U = \{u_1, \ldots, u_n\}$。我们不对噪声的形式或幅度作任何假设。因此，可能的观测项集合可以是无限的，并且严格包含真实全集 U。

我们假设可以访问一个预言机，对于任意一对观测项 σ_i 和 σ_j，它都能判断二者是否相似，记作 $\sigma_i \sim \sigma_j$。每个项目都与自身相似，而 U 中的元素两两不相似。请注意，由于噪声任意，相似关系可能同时产生假阳性和假阴性：即使两个项目源自不同的真实元素，也可能被判定为相似；即使两个项目源自同一个真实元素，也可能被判定为不相似。

我们将大小为 m 的数据集 σ 建模为图 $G^\sigma = ([m], E^\sigma)$，其中节点 i 表示 σ_i。边集 E^σ 定义为 $\{(i, j) \in [m]^2 \mid i \le j, \ \sigma_i \sim \sigma_j\}$。由于始终有 $\sigma_i \sim \sigma_i$，G^σ 中的每个节点都包含一个自环。对于每个 $i \in [m]$，令 $B_i^\sigma = \{j \in [m] \mid \sigma_j \sim \sigma_i\}$ 表示其对应项目与 σ_i 相似的节点集合（包括 σ_i 本身），并令 $d_i^\sigma = |B_i^\sigma|$ 表示该集合的大小（或者说，第 i 个节点的度）。

容易看出，对于长度为 m 的无噪声数据集 τ，$G^\tau = ([m], E^\tau)$ 包含一组不相交团，每个团都对应于 τ 中一个不同的全集元素。我们将这样的图 G^τ 称为簇图。[1]

对于统计函数 $f(\cdot)$ 和含噪数据集 $\sigma = (\sigma_1, \ldots\ldots, \sigma_m)$，我们将 $f(\sigma)$ 定义为 $f(\tau)$，其中 $\tau = (\tau_1, \ldots\ldots, \tau_m)$ 表示 σ 的真实值，而 τ_i 是 σ_i 的真实元素。由于 τ 是无噪声的，$f(\tau)$ 总是定义良好的。然而，τ 本身是未知的。我们的目标是设计一种算法，利用输入数据集 σ 近似计算 $f(\tau)$。

乍看之下，这项任务似乎不可能实现。然而，正如本文将以 F_p 为例所展示的那样，通过在近似比中引入参数 $\eta_f(\sigma, \tau)$，近似计算 $f(\tau)$ 是可行的。这个参数称为 f-失配歧义，用于刻画 σ 和 τ 关于统计函数 f 的差异程度。

我们首先正式定义无噪声数据集的频率矩。

定义1.1（频率矩） 设 $U = (u_1, \ldots, u_n)$ 为一个有限全集，$\tau = (\tau_1, \ldots, \tau_m) \in U^m$ 为一个无噪声数据集。对于每个元素 $u_j \in U$，定义其在 τ 中的频率为

$$f_j = |\{i \in [m] \mid \tau_i = u_j\}|.$$

τ 的第 p 阶频率矩定义为 $F_p(\tau) = \sum_{j=1}^n f_j^p$.

接下来，我们为 F_p 问题定义一个失配歧义参数。

定义1.2（F_p-失配歧义（$p \ge 1$）） 对于含噪数据集 $\sigma = (\sigma_1, \ldots, \sigma_m)$ 及其真实值 $\tau = (\tau_1, \ldots, \tau_m)$，$\sigma$ 相对于 τ 的 F_p-失配歧义定义为

$$\eta_p(\sigma, \tau) = \frac{1}{F_p(\tau)} \sum_{i \in [m]} \left(|B_i^\sigma \cup B_i^\tau|^{p-1} - |B_i^\sigma \cap B_i^\tau|^{p-1} \right). \tag{1}$$

在不会引起混淆时，我们简记 $\eta_p(\sigma, \tau)$ 为 η_p。

我们将　想就我们为 F_p选择的歧义定义作几点说明。

- 当 σ 本身无噪声时，对于所有 $i \in [m]$，$B_i^\sigma = B_i^\tau$，我们都有 $\eta_p = 0$。

[1]理解簇图的一种等价且直观的方式是：它是一个“无 P3”图。这意味着不存在长度为三的诱导路径。

- 当 $p = 1$ 时，我们总有 $\eta_1 = 0$。事实上，对于 $p = 1$ 的情况，我们只需统计数据流中的元素数量，这即使在有噪声情形下也很简单。
- 我们将存在于 G^σ 中、但不存在于 G^τ 中的边称为假阳性边；将存在于 G^τ 中、但不存在于 G^σ 中的边称为假阴性边。对于 $p = 2$， η_2 中的术语 $\sum_{i\in[m]}(|B_i^\sigma \cup B_i^\tau| - |B_i^\sigma \cap B_i^\tau|)$ 等于假阳性边与假阴性边数量之和的两倍。我们将假阳性边和假阴性边统称为不匹配，这也解释了为何将这一类参数称为“失配歧义”。
- 我们注意到，在所有可能的真实值 τ 中取最小值后， G^σ 中不匹配边的数量与 $G^{cor} = ([m], E^{cor})$ 的相关聚类代价一致：对于满足 $\sigma_i \sim \sigma_j$ 的每一对 (i, j)，在 E^{cor} 中计入一条“$+1$”边；对于满足$\sigma_i \not\sim \sigma_j$ 的每一对 (i, j)，计入一条“-1”边。[2]
- 我们将和 $\sum_{i\in[m]} \left(|B_i^\sigma \cup B_i^\tau|^{p-1} - |B_i^\sigma \cap B_i^\tau|^{p-1}\right)$ 按$F_p^{\,1}(\tau)$ 进行缩放，以衡量不匹配边对我们试图估计的量 $F_p(\tau)$的影响。引入该缩放因子是为了方便；它将加性误差转化为相对误差，以便于表达近似比。

f-失配歧义的适当定义可能因不同的统计问题而异。但一般而言，该参数应表示为输入数据集的图表示中相对于相应真实值的假阳性边和假阴性边的函数。

在本文中，我们主要关注失配歧义较小的情形；鉴于我们拥有良好的相似性预言机，我们认为这种情况在实践中很常见。另一方面，当数据集表现出较大的失配歧义时，标准统计度量可能失去其意义，并且可能无法仅基于噪声输入 σ来估计 $f(\tau)$——其中 τ表示真实值。

预备知识。在本文中，我们始终假设 p 是一个常数。写作 F_p 时，我们总是指 $F_p(\tau)$，其中 τ 是未知的真实值。为简洁起见，我们常将 $\eta_p(\sigma, \tau)$ 写作 η_p，因为从上下文中通常可以清楚地判断 σ，而未知的真实值 τ 始终是隐式定义的。

我们用 $[n]$ 表示 $\{1, \ldots, n\}$。除非另有说明，所有对数均以 2 为底。~写作 $x = a \pm b$ 时，我们指的是 $x \in [a - b, a + b]$。如果以至少 $(1 - \delta)$ 的概率有 $(1 - \epsilon)X \le X \le (1 + \epsilon)X$，则称 X 是 X 的 (ϵ, δ)-近似。我们假设数据集中的每个项目都能放入一个字中。

请注意，在无噪声情形下，上界通常以全集大小n 而非数据流长度 m来表示，这基于 n 和 m 呈多项式关系这一标准假设。然而，在有噪声情形下，全集大小并未明确定义，因为它可能没有上限。因此，我们假设算法已知 m，它可视为不同元素数量的上界（也因此是有效全集大小的上界），因为可以忽略数据流中未出现的元素。如果事先不知道 m，则需要额外进行一遍扫描（在流式模型中）或一轮通信（在协调器模型中）来获知 m。

[2]在相关聚类问题中，我们有一个完全图 $G = (V, E)$，其中 E 中的每条边都标记为“$+1$”或“-1”。目标是将 V 划分为若干簇，使错误聚类的边总数最少。如果一条边标记为“$+1$”，而其端点属于不同的簇；或者标记为“-1”，而其两个端点被分配到同一个簇，则称该边被错误聚类。

为便于比较我们在有噪声情形下得到的上界与无噪声情形下已有的结果，我们也用 m 表示无噪声情形下的全集大小。值得注意的是，在无噪声情形下针对 F_p 下界的困难输入构造中（例如 [4, 9]），始终有 $n = \Theta(m)$。

我们的结果。 我们在数据流模型中的算法结果可表述如下：

- 对于具有失配歧义 $\eta_p \leq 3^{-\frac{1}{(p!)}}$ 的 m 个元素的数据流，存在一种针对 F_p ($p \in \mathbb{Z}^+$) 的单遍 (($\epsilon + O(\eta_p)$) 0.01)-近似算法。该算法使用 $O\left(\frac{1}{\epsilon^2} m^{1-1/p}\right)$ 字的空间。

我们用以下下界结果补充算法结果：

- 对于任意常数 $C \geq 0$，任何针对 F_p ($p >$1) 的 O(1)-遍 (($\epsilon + C\eta_p$) 0.48)-近似算法，若处理包含 m 个元素的数据流，则至少需要 $\Omega\left(\frac{1}{\epsilon^{1/p}} m^{1-1/p}\right)$ 比特的空间。

这一下界结果与无噪声情形形成对比：在该情形下，对于 F_p ($p > 1$)，存在针对包含 m 个元素的数据流的单遍 (ϵ,0.01)-近似算法（例如 [8]），且仅使用$\tilde{O}\left(\frac{1}{\epsilon^2} m^{1-2/p}\right)$ 比特的空间。[3] 具体而言，在无噪声情形下，$p = 2$ 的空间复杂度为对数级，而在有噪声情形下则为多项式级。

在协调器模型中，我们得到以下算法结果：

- 对于分布在 k 个站点上的、包含 m 个元素且具有失配歧义 $\eta_p \leq 0.4$ 的数据集，存在一个两轮的 (($\epsilon + O(\eta_p)$) 0.01)-近似算法，用于 F_p ($p \geq 1$)。该算法使用$O\left(\frac{1}{\epsilon^2} k m^{1-1/p}\right)$ 通信字数。

对于下界，我们证明：

- 对于任意常数 $C \geq 0$，在分布于 $k \geq 10p(C+1)$个站点上的、包含 m 个元素的数据集上，任何用于 F_p ($p > 1$) 的 (($\epsilon + C\eta_p$) 0.48)-近似算法都必须至少使用 $\Omega\left(\frac{1}{\epsilon^{1/p}} m^{1-1/p}\right)$通信比特。

这一下界结果与无噪声情形形成鲜明对比：在无噪声情形下，存在仅使用 $O\left(\frac{1}{\epsilon^2} k^{p-1} \log^{O(1)} m\right)$ 通信比特 [23]的 (ϵ,0.01)-近似算法；特别是，其对输入规模 m的依赖是多对数级的。

对于数据流模型和协调器模型，我们关于 m 的上界和下界是一致的，区别仅在于成本是以通信字数还是比特来衡量。

最后，我们证明，在协调器模型中，当失配歧义足够小时，我们可以设计一种效率显著更高的算法；就 k 而言，其通信成本接近无噪声情形，并且独立于数据集大小 m。

- 对于分布在 k 个站点上的、包含 m 个元素且失配歧义为 $\eta_p \leq \frac{\epsilon^p}{4^{p+1} \cdot k^{p-1}}$ 的数据集，存在一个三轮 (($\epsilon + O(\eta_p)$) 0.01)-近似算法，用于 F_p ($p \geq 2$)。该算法使用 $O\left(\frac{1}{\epsilon^{p+1}} k^p\right)$ 个通信字。

对于 $p = 2$，上述算法中的失配歧义阈值约为 ϵ^2/k，该算法的通信成本为 $O\left(k^2/\epsilon^3\right)$。而在我们针对

[3]为简便起见，我们在 $\tilde{O}(\cdot)$ 表示法中隐藏多对数因子。

协调器模型的下界证明中，困难输入的失配歧义约为 ϵ/k，由此得到 $\Omega\left(\sqrt{m/\epsilon}\right)$ 的通信下界。综合来看，这些结果揭示了协调器模型中通信成本随失配歧义变化的大致相变。

1.1 相关工作

文献中的若干研究方向与我们的工作密切相关。

无噪声数据上的频率矩。 该 F_p 问题已在无噪声情形下得到广泛研究，涵盖数据流模型 [31, 1, 24, 5, 32, 3, 7, 2, 19, 34, 4, 9, 20, 6, 21]和协调器模型 [23, 13, 35, 17, 29, 25]。在数据流模型中，对于 $p > 2$，已知的最佳空间上界为 $O\left(\frac{1}{\epsilon^2} n^{1-2/p} \log^2 n\right)$ 比特 [19], 其中 n 表示全集大小，且假设数据流长度满足 $m = n^{\Theta(1)}$。该上界在 $\log n$ 因子内是紧的 [31]。在协调器模型中，对于 $p \geq 2$，存在一种两轮算法，其通信成本为 $O\left(\frac{1}{\epsilon^2} k^{p-1} \log^{O(1)} n\right)$ 比特 [23],，且该结果在 $\log^{O(1)} n$ 因子内是紧的 [35]。

含噪数据上的统计估计。 关于在数据流模型中处理含噪数据集的文献相当稀少。 [10] 和 [11] 的作者研究了不同元素问题（F_0）以及 ℓ_0-采样，但其结果仅限于常数维欧几里得空间以及支持高效局部敏感哈希的度量空间。最近的一项工作 [37] 将范围扩展到一般度量空间（或更一般地，任何配备0/1相似度函数的域）。然而，这些工作中使用的数据歧义模型是专门为 F_0问题量身定制的——更确切地说，它被定义为图 G^σ 的最小基数不相交团划分与最近簇图之间的差异——并且不能自然地推广到其他统计问题。

在协调器模型中，关于含噪数据统计估计的唯一已知工作是 [36]。然而，它仅关注其图表示构成簇图的数据集。

不确定数据上的流式算法。 一系列工作研究了针对不确定/概率数据的流式算法 [27, 26, 12, 28, 39], ，其中数据噪声通过“可能世界”语义来刻画。在该框架中，每个元素都关联一个关于其可能出现情况的概率分布，因此适用于拼写错误等噪声来源。然而，这种方法仅在论域固定且事先已知时才可行，而这在现实世界应用中往往并不成立。例如，如果一个数据项目是从未被观测到的原始副本经扰动得到的图像，那么就不可能设计出一个总能返回正确真实元素的预言机。再举一个例子，对大语言模型的两次查询可能产生语义几乎相同但表层形式略有不同的答案。在这种情况下，“真实元素”的概念是抽象的，难以明确化，而且可能的语义数量可以是无界的。因此，我们基于相似性预言机的框架适用于更广泛的一类含噪数据。

2 算法

本章节介绍了在数据流和协调器模型中针对含噪数据集的 F_p 算法。

我们用 $\mathrm{mult}(i_1, \ldots\ldots, i_p)$ $(i_1, \ldots\ldots, i_p \in [m])$ 表示多项式系数，其定义为：$\mathrm{mult}(i_1, \ldots\ldots, i_p) = \frac{p!}{\prod_{k\in[m]}(\sum_{j\in[p]} \mathbf{1}\{i_j=k\})!}$。换言之，$\mathrm{mult}(i_1, \ldots\ldots, i_p)$ 表示由 $\{\sigma_{i_1}, \ldots, \sigma_{i_p}\}$ 构成的排列数量。

2.1 流式算法

我们首先介绍一种用于 F_p 的单遍流式算法，其近似保证由 η_p 参数化。

思路与技术概述。 我们称节点元组 $(i_1, \ldots\ldots, i_p)$ 为 G^σ 中的一个 p 团，当且仅当对所有 $j,\ k \in [p],\ \sigma_{i_j} \sim \sigma_{i_k}$ 均成立；若团中 p 节点的顺序会影响结果，则称该 p团为有序团。在我们的设定中，节点被视为与自身相似，因此团中允许包含重复节点。

请注意，对于真实数据流 τ，F_p恰好等于 $\left|\mathcal{K}_p^\tau\right|$，其中 $\mathcal{K}_p^\tau$是 G^τ 中有序 p-团的集合 G^τ。[4] 对于含噪数据流 σ，令 $\mathcal{K}_p^\sigma$为 G^σ 中有序 p-团的集合 G^σ。我们的主要观察是，对于任意真实值 τ，只要 η_p足够小，就有 $\left|\mathcal{K}_p^\sigma\right| \approx \left|\mathcal{K}_p^\tau\right| = F_p$。因此，如果我们能够准确估计 $\left|\mathcal{K}_p^\sigma\right|$，就能很好地估计 F_p。

现在的问题是如何在单遍中估计 $\left|\mathcal{K}_p^\sigma\right|$。一个自然的想法是从 $[m]^p$ 中抽取均匀随机的 p-元组，然后统计其中有多少个属于 $\mathcal{K}_p^\sigma$。然而，这种方法存在一个问题：$\left|\mathcal{K}_p^\sigma\right|$ 可能小至 m，从而导致样本属于 $\mathcal{K}_p^\sigma$的概率很低。另一方面，在数据流模型中直接计数 $\left|\mathcal{K}_p^\sigma\right|$ 颇具挑战，因为元组中的节点是有序的，而这种顺序通常与节点在数据流中出现的顺序并不一致。因此，我们转而关注递增有序的 p-团，即元组中节点的顺序与它们在数据流中的顺序相匹配的团。

我们设计了一种采样与计数过程，记为 $\mathcal{A}$，它利用了这样一个观察：每个按递增顺序排列的 j-团，都可以由一个按递增顺序排列的 $(j-1)$-团逐步构造而成。从概念上讲，这一生成过程可以表示为一棵树，根节点对应一个 0-团，其子节点则是在 0-团中再添加一个节点所形成的 1-团；如此继续，直至 p 层。树的叶节点恰好枚举出所有按递增顺序排列的 p-团。随后，我们从根节点到叶节点进行随机游走，以采样一条路径，并根据选中该路径的概率和适当的多项式系数，得出一个估计值 X。我们可以证明，X 满足 $\mathbf{E}[X] = |\mathcal{K}_p^\sigma|$。

上述方法的关键步骤是确定最终近似所需的无偏估计数量。设 $X_1, \ldots, X_t$ 是通过对 $\mathcal{A}$ 进行 t 次独立运行而得到的估计值，其中对每个 $j \in [t]$，都有 $\mathbf{E}[X_j] = \left|\mathcal{K}_p^\sigma\right|$。设 $\bar{X} = \frac{1}{t}\sum_{j\in[t]} X_j$。如果 σ 无噪声，则不难证明，只需 $t = O\left(\frac{1}{\epsilon^2}\bar{m}^{1-1/p}\right)$ 个估计值，就足以确保 $\bar{X}$ 紧密集中于其期望值 $\bar{}$ 附近。然而，在有噪声情形下，最坏情况下的 X 可能表现出很高的不稳定性。例如，考虑 $p=2$ 且 G^σ 是一张星图的情形。在这种情况下，$\mathcal{K}_2^\sigma$ 由所有相似节点组成的有序对构成，其中约三分之二的节点对包含中心节点。如果中心节点在数据流中最先到达，那么它以 $\left(1-\frac{1}{m}\right)$ 的概率不会被采样。因此，$\mathcal{K}_2^\sigma$ 中约三分之二的节点对不会被计数，这使我们无法在 $t = o(m)$ 时，以 0.99 的概率获得 1.01-近似。

我们能够证明的是，当 η_p 较小时，只需使用 $\bar{}$ 次线性以下数量的样本，就可以准确估计 $\left|\mathcal{K}_p^\sigma\right|$。首先，不难通过马尔可夫给出 X 的上界

[4] 例如，对于 $p=3$ 和 $G^\tau = (V, E)$，其中 $V = \{1,2\}$ 和 $E = \{(1,1), (1,2), (2,2)\}$，那么 $\mathcal{K}_p^\tau$ 包含 $(1,1,1)$，$(1,1,2)$，$(1,2,1)$，$(2,1,1)$，$(1,2,2)$，$(2,1,2)$，$(2,2,1)$ 以及 $(2,2,2)$。

¯不等式。事实上，由于 X 是独立随机变量之和，我们可以应用 [18] 中的一个不等式¯，得到更好的界。难点在于建立 X的下界。¯标准方法是对 X 的方差（或者每个 X_j （$j \in [t]$）的方差）进行界定，然后应用切比雪夫不等式。然而，如果直接分析 X_j 的方差而不考虑失配歧义 η_p，方差可能会非常高。再来看星图，它相对于任何真实值都具有较高的 η_p 。简单计算可得 $\mathrm{E}[X_j] = \Theta(m)$ 和 $\mathsf{Var}[X_j] = \Omega(m^p)$ 。因此，我们必须使用 $t = \Omega(m^{p-2})$ 份 X_j ，才能将 X 的¯方差降至 $O(1)$。所以，分析时必须考虑 η_p 。难点在于，¯构造 X 无法纳入 η_p，因为这是算法未知的量。

我们通过将每个随机变量 X_j ($j \in [t]$) 与另一个随机变量 $Y_j \triangleq \min\left\{X_j,\ m(p!)\ \left(d^{\tau}_{I_1}\right)^{p-1}\right\}$，联系起来，绕过上述问题，其中 I_1 在 $\{1,\ldots,m\}$中均匀随机分布。构造这样一个随机变量 Y_j 的直觉如下：回想一下，如果我们将 $\mathcal{A}$应用于真实值 τ，其图表示 G^{τ} 是一个簇图，那么我们可以得到关于 X_j方差的更紧上界，因此只需要 $t = o(m)$次 $\mathcal{A}$运行。现在，如果 η_p 较小，我们预计 $\mathcal{K}^{\sigma}_p$ 和 $\mathcal{K}^{\tau}_p$会有显著重叠，因此$\left|\mathcal{K}^{\sigma}_p\right|$和$\left|\mathcal{K}^{\sigma}_p \cap \mathcal{K}^{\tau}_p\right|$彼此接近。直观地说，我们可以将带有 $\mathbf{E}[Y_j] \approx \left|\mathcal{K}^{\sigma}_p \cap \mathcal{K}^{\tau}_p\right|$的 Y_j理解为接近带有 $\mathbf{E}[X_j] = \left|\mathcal{K}^{\sigma}_p\right|$的 X_j，但方差更小。因此，我们可以先为 Y_j 建立一个下界，这也为 X_j提供了一个下界。

算法与分析。 算法详见算法2，其中将算法1作为子程序。在较高层次上，算法 1旨在获得有序 p-团数量的无偏估计量，该估计量可作为 G^{σ}的低偏差估计 F_p。具体而言，算法在主循环（第 p-团（第 2–14 行），随后使用适当的多项式系数对结果进行缩放（第15 行）。通过并行多次执行算法 1，算法 2 采用最小均值聚合（第 3–4 行），为 F_p 生成最终的近似值。

关于算法2，我们有如下结果。

定理 2.1 对于任意常数 $p \in \mathbb{Z}^+$，给定长度为 m 且满足 $\eta_p \leq_{1\,3(p!)}$ 的含噪输入数据流，算法 2使用单遍计算 $((\epsilon + O(\eta_p)\,0.01)$-近似值 F_p，并使用 $O\left(\frac{1}{\epsilon^2} m^{1-1/p}\right)$ 字的空间。

—

在本章节的其余部分中，我们将证明定理 2.1。我们将使用以下不等式。

引理 2.2（[18]）设 $Z_1, \ldots, Z_t$ 为任意非负独立随机变量，其期望分别为 $\mu_1, \ldots, \mu_t$，其中对于所有 i，均有 $\mu_i \leq 1$ 。那么，对于任意 $\delta > 0$，

$$\mathbf{Pr}\left[\sum_{i=1}^{t} Z_i < \left(\sum_{i=1}^{t} \mu_i\right) + \delta\right] > \min\left(\frac{\delta}{1+\delta}, \frac{1}{13}\right).$$

注意，当 $p = 1$ 时，算法 2 的输出始终为 m，这恰好是 F_1。因此，我们重点讨论整数 $p \geq 2$。

我们首先引入以下概念。

定义 2.3（有序 p-团）对于长度为 m 的数据流 σ ，将 G^{σ} 中的有序p-团集合定义为 $\mathcal{K}^{\sigma}_p = \left\{(i_1, \ldots, i_p) \in [m]^p \mid \forall (j,k) \in [p]^2, \sigma_{i_j} \sim \sigma_{i_k}\right\}$.

算法 1： 流式鲁棒-F_p-单样本

Input: a noisy data stream σ
Output: an estimate of F_p

$r_1, \ldots, r_p, i_1, \ldots, i_p \leftarrow 0$, $a_1, \ldots, a_p \leftarrow$ null
foreach *incoming* σ_j **do**
  $r_1 \leftarrow r_1 + 1$
  with probability $\frac{1}{r_1}$, $(a_1, i_1, r_2) \leftarrow (\sigma_j, j, 0)$
  if $a_1 \sim \sigma_j$ **then**
    $r_2 \leftarrow r_2 + 1$
    with probability $\frac{1}{r_2}$, $(a_2, i_2, r_3) \leftarrow (\sigma_j, j, 0)$
    if $a_2 \sim \sigma_j$ **then**
      $r_3 \leftarrow r_3 + 1$
      with probability $\frac{1}{r_3}$, $(a_3, i_3, r_4) \leftarrow (\sigma_j, j, 0)$
      $\ldots$
      if $a_{p-1} \sim \sigma_j$ **then**
        $r_p \leftarrow r_p + 1$
        with probability $\frac{1}{r_p}$, $(a_p, i_p) \leftarrow (\sigma_j, j)$
返回 mu 1，lt.(i.., ^{p} i)$_{k\in[p]}$ ^{r_k}。

算法 2： 流式鲁棒-F_p

输入：含噪数据流 σ，大小为 m，参数 ϵ输出：((ϵ + O(ηp)) 0.01)-近似，对应于 F_p1 $\ell \leftarrow 100$, $t \leftarrow \frac{10^6}{\epsilon^2}(p!)m^{1-1/p}$ 2 获得 ℓt 估计 $\{X_{i,j}\}_{i\in[\ell],j\in[t]}$，通过并行运行 ℓt 个算法 1 实例 3 对于 $i \in [\ell]$，令 $\bar{X}_i \leftarrow \frac{1}{t}\sum_{j\in[t]} X_{i,j}$ 4 返回 $\min\{X_1, \ldots, X_\ell\}$。

事实 2.4 对于无噪声数据流 τ, $\left|\mathcal{K}_p^\tau\right| = F_p$。

证明：设 n_τ 为 τ 中不同元素的数量，$\{V_1, \ldots, V_{n_\tau}\}$ 为 G^τ 中团的集合。容易看出，对于任意 $K_p \in \mathcal{K}_p^\tau$，$K_p$ 中所有 p 节点都必须属于某个 $k \in [n_\tau]$ 对应的同一个团 V_k。因此，有序 p-团的总数为 $\sum_{k\in[n_\tau]} |V_k|^p = F_p$。□

下面的引理表明，对于含噪数据流 σ 及其对应的真实值 τ，如果 η_p 很小，那么 $\mathcal{K}_p^\sigma$ 与 $\mathcal{K}_p^\tau$ 有显著重叠。

引理 2.5 对于分别表示含噪数据流 σ 及其对应真实值 τ 的图 $G^\sigma = ([m], E^\sigma)$ 和 $G^\tau = ([m], E^\tau)$，有

$$\left|\mathcal{K}_p^\sigma \setminus \mathcal{K}_p^\tau\right| \le \eta_p F_p, \quad and \quad \left|\mathcal{K}_p^\tau \setminus \mathcal{K}_p^\sigma\right| \le \frac{p}{2} \cdot \eta_p F_p.$$

证明：我们首先对 $\mathcal{K}_p^\sigma \setminus \mathcal{K}_p^\tau$ 的大小进行界定。考虑 $\mathcal{K}_p^\sigma \setminus \mathcal{K}_p^\tau$ 中满足 $i_1 = i$ 的有序 p-团 $K_p = (i_1, \ldots\ldots, i_p)$。必定存在一个 $k \in [p]$，使得 $i_k \in B_i^\sigma \setminus B_i^\tau$。因此，

$\mathcal{K}_p^\sigma \setminus \mathcal{K}_p^\tau$ 中满足 $i_1 = i$ 的有序 p-团至多有 $|B_i^\sigma|^{p-1} - |B_i^\sigma \cap B_i^\tau|^{p-1} \le |B_i^\sigma \cup B_i^\tau|^{p-1} - |B_i^\sigma \cap B_i^\tau|^{p-1}$. 。对 $i \in [m]$ 求和可得

$$\left|\mathcal{K}_p^\sigma \setminus \mathcal{K}_p^\tau\right| \le \sum_{i\in[m]} \left(|B_i^\sigma \cup B_i^\tau|^{p-1} - |B_i^\sigma \cap B_i^\tau|^{p-1}\right) = \eta_p F_p.$$

接下来，我们界定 $\mathcal{K}_p^\tau \setminus \mathcal{K}_p^\sigma$ 的大小。令 n_i 为 $\mathcal{K}_p^\tau$ 中包含 i 且至少包含一个来自 $B_i^\tau \setminus B_i^\sigma$ 的节点的有序 p-团的数目。我们有

$$n_i \le p\left(|B_i^\tau|^{p-1} - |B_i^\sigma \cap B_i^\tau|^{p-1}\right) \le p\left(|B_i^\sigma \cup B_i^\tau|^{p-1} - |B_i^\sigma \cap B_i^\tau|^{p-1}\right).$$

注意，对于每个 $K_p = (i_1, \ldots, i_p) \in \mathcal{K}_p^\tau \setminus \mathcal{K}_p^\sigma$，存在 $x, y \in [p]$ 使得 $i_x \ne i_y$ 且 $(i_x, i_y) \notin E^\sigma$，这意味着 K_p 在 n_{i_x} 和 n_{i_y} 中都被计数。对 $i \in [m]$ 求和，我们有

$$\left|\mathcal{K}_p^\tau \setminus \mathcal{K}_p^\sigma\right| \le \frac{1}{2}\sum_{i\in[m]} n_i \le \frac{p}{2} \cdot \sum_{i\in[m]} \left(|B_i^\sigma \cup B_i^\tau|^{p-1} - |B_i^\sigma \cap B_i^\tau|^{p-1}\right) = \frac{p}{2} \cdot \eta_p F_p.$$

□

事实 2.4 与引理 2.5 表明我们可以估计 F_p，其途径是估计 $\left|\mathcal{K}_p^\sigma\right|$。然而，在亚线性内存空间下的数据流模型中直接估计该量似乎较为困难。因此，我们将注意力转向一个密切相关的集合。

定义 2.6（递增有序 p-团）对于数据流 σ，将 G^σ 中递增有序的 p-团集合定义为 $\vec{\mathcal{K}}_p^\sigma = \{(i_1, \ldots, i_p) \in \mathcal{K}_p^\sigma \mid i_1 \le \ldots \le i_p\}$。

注意，对于每个递增有序的 p-团 $\vec{K}_p = (i_1, \ldots, i_p) \in \vec{\mathcal{K}}_p^\sigma$，$\mathcal{K}_p^\sigma$ 中恰好有 $\mathrm{mult}(i_1, \ldots, i_p)$ 个有序 p-团，它们由相同的节点多重集构成。

为便于后续分析，我们引入以下概念。

定义 2.7（尾集）对于数据流 σ 和 $j \in [p]$，将 G^σ 中递增有序的 j-团 $(i_1, \ldots, i_j)$ 的尾集定义为 $\mathit{tail}(i_1, \ldots, i_j) = \{i \in [m] \mid i \ge i_j \wedge (\forall k \in [j], \sigma_i \sim \sigma_{i_k})\}$.

用字来说，$\mathrm{tail}(i_1, \ldots, i_j)$ 包含数据流中在 i_j（包括 i_j）之后出现的所有节点 $i \in [m]$，这些节点与 j-团 $(i_1, \ldots, i_j)$ 一起构成一个递增有序的 $(j+1)$-团。

下述引理表明，算法 1 输出一个关于 $\left|\mathcal{K}_p^\sigma\right|$ 的无偏估计量。设 $I_1, \ldots, I_p$ 对应于 $i_1, \ldots, i_p$ 在算法 1 中的随机节点，并设 $R_1, \ldots, R_p$ 对应于 $r_1, \ldots, r_p$ 在算法 1 中的尾计数。注意，$R_1 = m$ 实际上是一个固定值，而其余则为真正的随机变量。

引理 2.8 设 $X = \mathit{mult}(I_1, \ldots, I_p) \prod_{k\in[p]} R_k$ 为算法 *1* 的输出。则有 $\mathbf{E}[X] = \left|\mathcal{K}_p^\sigma\right|$。

证明：容易看出，在算法 1中，I_1 从 $[m]$ 中均匀采样，且 I_k 从 $\mathrm{tail}(I_1, \ldots, I_{k-1})$ 中均匀采样，对于 $k = 2, \ldots, p$。令 $\vec{K}_p = (i_1, \ldots, i_p) \in \vec{\mathcal{K}}_p^\sigma$ 为一个递增有序的 p-团。我们有

$$\begin{aligned}
\mathbf{Pr}[I_1 = i_1, \ldots, I_p = i_p] &= \mathbf{Pr}[I_1 = i_1] \cdot \prod_{k=2}^{p} \mathbf{Pr}\left[I_k = i_k \mid I_1 = i_1, \ldots, I_{k-1} = i_{k-1}\right] \\
&= \frac{1}{m}\prod_{k=2}^{p} \frac{1}{|\mathrm{tail}(i_1, \ldots, i_{k-1})|}. \qquad (2)
\end{aligned}$$

注意到 $R_1 = m$，且 $R_k = |\text{tail}(I_1, \ldots, I_{k-1})|$ 对于 $k = 2, \ldots, p$，且 $(I_1, \ldots, I_p)$ 始终为一个递增有序的 p-团。由此可得

$$\begin{aligned}
\mathbf{E}[X] &= \sum_{(i_1,\ldots,i_p)\in\vec{\mathcal{K}}_p^\sigma} \mathbf{Pr}[I_1 = i_1, \ldots, I_p = i_p] \cdot \mathbf{E}[X \mid I_1 = i_1, \ldots, I_p = i_p] \\
&\overset{(2)}{=} \sum_{(i_1,\ldots,i_p)\in\vec{\mathcal{K}}_p^\sigma} \frac{1}{m} \prod_{k=2}^{p} \frac{1}{|\text{tail}(i_1, \ldots, i_{k-1})|} \cdot \text{mult}(i_1, \ldots, i_p) \cdot m \prod_{k=2}^{p} |\text{tail}(i_1, \ldots, i_{k-1})| \\
&= \sum_{(i_1,\ldots,i_p)\in\vec{\mathcal{K}}_p^\sigma} \text{mult}(i_1, \ldots, i_p) = \left|\mathcal{K}_p^\sigma\right|.
\end{aligned}$$

□

设 $X_1, \ldots, X_t$ 表示 t 算法 1的独立输出，并设 $\bar{X} = \frac{1}{t}\sum_{i\in[t]} X_i$。接下来的两个引理表明，当失配歧义 η_p 较小时，X 会以较高概率紧密集中在 F_p 附近。

引理 2.9（$\bar{X}$ 不太大） $\mathbf{Pr}\left[\bar{X} < (1 + \epsilon + \eta_p) \cdot F_p\right] > \frac{1}{13}$.

证明：引理 2.8给出 $\mathbf{E}[\bar{X}] = \left|\mathcal{K}_p^\sigma\right|$。应用引理 2.2（令 $Z_i = X_i/|\mathcal{K}_p^\sigma|$ 且 $\delta = 1$），可得

$$\mathbf{Pr}\left[\bar{X} < \left(1 + \frac{1}{t}\right)\left|\mathcal{K}_p^\sigma\right|\right] > \frac{1}{13}. \tag{3}$$

根据事实 2.4 和引理 2.5,

$$\left|\mathcal{K}_p^\sigma\right| \le \left|\mathcal{K}_p^\tau\right| + \left|\mathcal{K}_p^\sigma \setminus \mathcal{K}_p^\tau\right| \le (1 + \eta_p)F_p. \tag{4}$$

该引理由（3）、（4）以及 $\frac{1}{t} \ll \frac{\epsilon}{2}$ 这一事实可得。□

引理 2.10（$\bar{X}$ 不太小） $\mathbf{Pr}\left[\bar{X} \ge (1 - \epsilon - 2(p!)\eta_p) \cdot F_p\right] > 1 - 2 \cdot 10^{-5}$

Lem

证明：对于每个 $j \in [t]$，令 $Y_j = \min\left\{X_j,\ m(p!)\left|B_{I_1}^\tau\right|^{p-1}\right\}$. 我们有以下断言，其证明稍后给出。第一个断言指出，在 η_p 较小的情况下，每个 Y_j 的期望与 F_p 相差不大；第二个断言则给出了每个 Y_j的方差上界。

断言 2.11 对于任意 $j \in [t]$，$(1 - 2(p!)\eta_p)F_p \le \mathbf{E}[Y_j] \le (1 + \eta_p)F_p$.

断言 2.12 对于任意 $j \in [t]$，$\mathbf{Var}[Y_j] \le (\mathbf{E}[Y_j])^2 \cdot \frac{p!}{1-2(p!)\eta_p} \cdot m^{1-1/p}$.

我们现在有

$$\begin{aligned}
\mathbf{Pr}\left[\bar{X} < (1-\epsilon-2(p!)\eta_p)\,F_p\right] &= \mathbf{Pr}\left[\frac{1}{t}\sum_{j\in[t]} X_j - (1-2(p!)\eta_p)F_p < -\epsilon F_p\right] && (5)\\
&\le \mathbf{Pr}\left[\frac{1}{t}\sum_{j\in[t]} Y_j - \mathbf{E}[Y_j] < -\epsilon F_p\right] && (6)\\
&\le \frac{\mathbf{Var}[Y_j]}{t\epsilon^2 F_p^2} && (7)\\
&\le \frac{(1+\eta_p)^2\mathbf{Var}[Y_j]}{t\epsilon^2(\mathbf{E}[Y_j])^2} && (8)\\
&\le \frac{(1+\eta_p)^2}{t\epsilon^2}\cdot\frac{p!\cdot m^{1-1/p}}{1-2(p!)\eta_p} && (9)\\
&\le \frac{1}{t}\cdot\frac{12(p!)m^{1-1/p}}{\epsilon^2} && (10)\\
&\le 2\cdot 10^{-5},
\end{aligned}$$

从（5）到（6），我们利用了这一事实 $X_j \ge Y_j$ 以及断言2.11的第一个不等式。从（7）到（8），我们应用断言2.11的第二个不等式。从（8）到（9），我们应用断言2.12。从（9）到（10），我们使用假设$\eta_p < {}_3{}^1{}_{(p!)}$。最后一个不等式源于我们对 t 的选择（算法2第 1 行）。□

现在我们可以证明定理2.1了。回顾一下，我们在算法2中设置了 $\ell = 100$。令$X_{\min} = \min\left\{\bar{X}_1,\ldots,\bar{X}_\ell\right\}$为算法2的输出。根据引理2.9和2.10，我们有

$$\begin{aligned}
\mathbf{Pr}\left[X_{\min} < (1+\eta_p+\epsilon)F_p\right] &\ge 1-\left(1-\frac{1}{13}\right)^{\ell} > 0.999, \quad \text{and}\\
\mathbf{Pr}\left[X_{\min} \ge (1-\epsilon-2(p!)\eta_p)F_p\right] &\ge 1-\ell\cdot 2\cdot 10^{-5} = 0.998.
\end{aligned}$$

该定理可由并集界推出。

最后，我们给出引理2.10中两个断言的证明。

证明：[命题的证明 2.11] 我们对每个 $j\in[t]$ 都进行证明，因此省略 j 在 X_j 和 Y_j中的下标，并为方便起见，将它们记作 X 和 Y。我们首先考察 X 与 Y 之间的差异，

$$\mathbf{E}[X-Y] = \sum_{i\in[m]}\mathbf{Pr}[I_1=i]\cdot\mathbf{E}[X-Y\mid I_1=i] = \frac{1}{m}\sum_{i\in[m]}\mathbf{E}[X-Y\mid I_1=i].$$

在以 $I_1=i$ 为条件时，我们有

$$\begin{aligned}
X &= \mathrm{mult}(i,I_2,\ldots,I_p)\cdot\prod_{k\in[p]}R_k \le p!\cdot\prod_{k\in[p]}R_k\\
&= p!\cdot m\cdot\prod_{k=2}^{p}|\mathrm{tail}(i,\ldots,I_{k-1})| && (11)\\
&\le p!\cdot m\cdot|\mathrm{tail}(i)|^{p-1} && (12)\\
&\le m(p!)|B_i^\sigma|^{p-1}, && (13)
\end{aligned}$$

其中从（11）到（12），我们利用如下事实：$\mathsf{tail}(i, I_2, \ldots, I_{p-1}) \subseteq \mathrm{tail}(i, I_2, \ldots, I_{p-2}) \subseteq \cdots \subseteq \mathrm{tail}(i)$。该步骤从（12）到（13）可得自 $\mathsf{tail}(i) \subseteq B_i^\sigma$。

因此我们有

$$\mathbf{E}[X-Y \mid I_1 = i] = \mathbf{E}\left[\max\left\{X - m(p!)|B_i^\tau|^{p-1}, 0\right\} \;\middle|\; I_1 = i\right] \le m(p!)\cdot\max\left\{|B_i^\sigma|^{p-1} - |B_i^\tau|^{p-1}, 0\right\}.$$

注意到 $\mathbf{Pr}[I_1 = i] = \frac{1}{m}$，我们有

$$\begin{aligned}
\mathbf{E}[X-Y] &\le p! \cdot \sum_{i\in[m]} \max\left\{|B_i^\sigma|^{p-1} - |B_i^\tau|^{p-1}, 0\right\} \\
&\le p! \cdot \sum_{i\in[m]} \left(|B_i^\sigma \cup B_i^\tau|^{p-1} - |B_i^\tau \cap B_i^\sigma|^{p-1}\right) \\
&= p! \cdot \eta_p F_p. \qquad (14)
\end{aligned}$$

因此，

$$\begin{aligned}
\mathbf{E}[Y] &= \mathbf{E}[X] - \mathbf{E}[X-Y] \\
&\ge \left|\mathcal{K}_p^\sigma\right| - p! \cdot \eta_p F_p \quad \text{(by (14) and Lemma 2.8)} \\
&\ge \left|\mathcal{K}_p^\tau\right| - \left|\mathcal{K}_p^\tau \setminus \mathcal{K}_p^\sigma\right| - p! \cdot \eta_p F_p \\
&\ge (1 - 2(p!)\eta_p) F_p. \quad \text{(by Fact 2.4 and Lemma 2.5)}
\end{aligned}$$

另一方面，由（4）和引理 2.8，我们有 $\mathbf{E}[Y] \le \mathbf{E}[X] = \left|\mathcal{K}_p^\sigma\right| \le (1+\eta_p)F_p$. 命题得证。 □

证明：[命题 2.12] 令 n_τ 表示真实值 τ中的不同元素个数，$\{V_1, \ldots, V_{n_\tau}\}$ 表示 G^τ中的团集合。我们对每个 $j \in [t]$证明，因此为方便起见，省略 j在 Y_j 中的下标。

根据 Y 的定义，我们有 $Y \le m(p!) \cdot \max_{i\in[m]} |B_i^\tau|^{p-1} = m(p!) \cdot \max_{k\in[n_\tau]} |V_k|^{p-1}$，这意味着

$$\mathbf{Var}[Y] \le \mathbf{E}[Y^2] \le m(p!) \cdot \max_{k\in[n_\tau]} |V_k|^{p-1} \cdot \mathbf{E}[Y]. \qquad (15)$$

注意到

$$\max_{k\in[n_\tau]} |V_k|^{p-1} = \left(\max_{k\in[n_\tau]} |V_k|^p\right)^{1-1/p} \le \left(\sum_{k\in[n_\tau]} |V_k|^p\right)^{1-1/p} = F_p^{1-1/p}. \qquad (16)$$

并且由 H¨older 不等式，

$$m = \sum_{k\in[n_\tau]} |V_k| \le n_\tau^{1-1/p} \left(\sum_{k\in[n_\tau]} |V_k|^p\right)^{1/p} = n_\tau^{1-1/p} F_p^{1/p}. \qquad (17)$$

结合（15）、（16）、（17）以及断言2.11的第一个不等式，我们有

$$\frac{\mathbf{Var}[Y]}{\mathbf{E}[Y]^2} \le \frac{p! \cdot m \cdot \max_{k\in[n_\tau]} |V_k|^{p-1}}{\mathbf{E}[Y]} \le \frac{p! \cdot n_\tau^{1-1/p} F_p^{1/p} \cdot F_p^{1-1/p}}{(1-2(p!)\eta_p)F_p} \le \frac{p! \cdot m^{1-1/p}}{1-2(p!)\eta_p}.$$

□

2.2 分布式算法

我们注意到，我们的流式算法无法直接转化为轮次高效分布式算法，因为从有序 p-团中采样至少需要 p 轮次。设计一种在通信和轮次两方面都高效的分布式算法，关键在于如下事实：对于真实数据集 τ，量 $F_p(\tau)$与G^τ的第$(p-1)$阶度矩一致，其定义为 $M_{p-1}^\tau \triangleq \sum_{i\in[m]} |B_i^\tau|^{p-1}$.当 $\eta_p(\sigma,\ \tau)$较小时，G^σ的第$(p-1)$阶度矩$M_{p-1}^\sigma \triangleq \sum_{i\in[m]} |B_i^\sigma|^{p-1}$预期接近 M_{p-1}^τ。因此，对 M_{p-1}^σ的准确估计可以给出 $F_p(\tau)$的良好近似。

在协调器模型中，获取 M_{p-1}^σ 的无偏估计相对 straightforward：我们可以均匀采样节点，并向每个站点查询其度数。然而，与流式场景类似，这种朴素的采样策略存在不稳定性。幸运的是，为数据流模型开发的分析技术同样可以应用于此场景。

我们用于在协调器模型中近似 F_p 的两轮算法如算法 3所述。在第一轮中，站点和协调器随机采样一组节点。在第11行，协调器尝试将该样本从“不放回”转换为“有放回”。在第二轮中，参与者计算每个采样节点的（$p-1$）阶度矩。最后，协调器使用最小均值聚合输出最终的近似值。

关于算法 3，我们有如下保证。

定理 2.13 对于任意常数 $p \geq 1$，给定一个大小为 m 且 $\eta_p \leq 0.4$ 划分在 k 个站点之间的输入数据集，在协调器模型中，算法 *3* 使用两轮和 $O\left(\frac{1}{\epsilon^2} k m^{1-1/p}\right)$ 通信字数，计算 F_p的 $((\epsilon + O(\eta_p))0.01)$-近似值。

尽管算法 3 与算法 2有很大不同，但两者的分析有相似之处，尤其是在处理随机变量不稳定性方面。

现在我们证明定理 2.13。首先，根据 Chernoff 界，我们有 $t\ell \leq |S'| \leq 3t\ell$ ，其概率为 $1-o(1)$，这意味着算法 3 以概率 $o(1)$ 失败。其次，容易看出，算法 3 只使用两轮，其通信复杂度有如下界$O(|S'| + k|S|) = O(kt\ell) = O\left(\frac{1}{\epsilon^2} k m^{1-1/p}\right)$。接下来需要证明算法 3的正确性。

我们先给出以下简单事实。

事实 2.14 对于无噪声数据集 τ, $F_p = \sum_{i\in[m]} (d_i^\tau)^{p-1}$。

证明：令 n_τ为 τ中不同元素的数量，令 $\{V_1, \ldots, V_{n_\tau}\}$ 为 G^τ中的团集合。于是有

$$F_p(\tau) = \sum_{k\in[n_\tau]} |V_k|^p = \sum_{k\in[n_\tau]} \sum_{i\in V_k} |V_k|^{p-1} = \sum_{k\in[n_\tau]} \sum_{i\in V_k} |B_i^\tau|^{p-1} = \sum_{i\in[m]} |B_i^\tau|^{p-1}.$$

□

我们有以下引理。

引理 2.15 对于数据集 σ，令 $I \in S$ 为从算法 *3* 中采样得到的一个节点，且 $X = m(d_I)^{p-1}$。我们有 $\mathbf{E}[X] = \sum_{i\in[m]} (d_i)^{p-1}$。

算法 3：分布式鲁棒-F_p输入：大小为 m 的含噪数据集 σ，划分至 k 个站点，其中第 κ 个站点持有$\sigma^{(\kappa)}$；参数 ϵ输出： F_p1 $t \leftarrow \frac{10^6 m^{1-1/p}}{\epsilon^2}$， $\ell \leftarrow 100$2 的一个 $((\epsilon + O(\eta_p))$ 0.01)-近似 m // 第一轮：3 协调器：将 m 发送给所有站点。4 对于站点 $\kappa \leftarrow 1, \ldots, k$，执行 5 $S^{(\kappa)} \leftarrow \emptyset$ /* 第 κ 个站点的采样节点集合 */ 6 对每个 $q \in V^{(\kappa)}$ 执行 7 以概率 $\frac{2t\ell}{m}$，$S^{(\kappa)} \leftarrow \overline{S^{(\kappa)} \cup \{q\}}$8 将 $S^{(\kappa)}$ 发送给协调器。9// 第二轮：10 /* 将“无放回采样”转换为“有放回采样” */ 11 协调器： $S' \leftarrow \bigcup_{\kappa \in [k]} S^{(\kappa)}$。如果 $|S'| < t\ell$，协调器则输出失败。否则，从 S' 中计算出样本 S，使得 (1) $|S| = t\ell$，且 (2) S 中的每个元素都是从 $[m]$ 中有放回均匀采样得到的。这可以使用 [14, 第 3 节2]中描述的方法完成。将 S 发送给所有站点。12 对于站点 $\kappa \leftarrow 1, \ldots, k$，执行 13 对每个 $i \in S$ 执行14 $d_i^{(\kappa)} \leftarrow \left|V^{(\kappa)} \cap B_i^{\sigma}\right|$ /* 第 κ 个站点上节点 i 的局部度 */ 15 将 $\{d_i^{(\kappa)}\}_{i \in S}$ 发送给协调器。16 // 输出：17 协调器：对每个 $i \in S$ 计算 $d_i \leftarrow \sum_{\kappa \in [k]} d_i^{(\kappa)}$。将 S 划分为 ℓ 组$\{S_1, \ldots, S_\ell\}$，每组包含 t 个样本。对于每个 $j \in [\ell]$，计算 $\bar{X}_j \leftarrow \frac{m}{t} \sum_{i \in S_j} (d_i)^{p-1}$。 18 返回 $\min\{X_1, \ldots, X_\ell\}$。

证明：由于 I 是从 $[m]$中均匀采样得到的，我们有

$$\mathbf{E}[X] = \sum_{i \in [m]} \mathbf{Pr}[I = i] \cdot m(d_i)^{p-1} = \sum_{i \in [m]} (d_i)^{p-1}.$$

□

令 $X_1, \ldots, X_t$ 为 X 中定义于引理 2.15 和 $\bar{X} = \frac{1}{t} \sum_{i=1}^{t} X_i$的独立同分布副本。接下来的两个引理表明，$X$ 不会过大或过小。

引理 2.16（$\bar{X}$ 不会过大）Pr $\left[\bar{X} < (1 + \epsilon + \eta_p) \cdot F_p\right] > \frac{1}{13}$.

证明： 根据引理 2.15，我们有 $\mathbf{E}[\bar{X}] = \sum_{i \in [m]} (d_i)^{p-1}$。对 2.2 应用引理 $X_1, \ldots, X_t$，可得

$$\mathbf{Pr}\left[\bar{X} < \left(1 + \frac{1}{t}\right) \sum_{i \in [m]} (d_i)^{p-1}\right] > \frac{1}{13}. \tag{18}$$

根据事实 2.14，$F_p = \sum_{i\in[m]}(d_i^\tau)^{p-1}$。因此我们有

$$\begin{aligned}\sum_{i\in[m]}(d_i)^{p-1} &\le F_p + \left|\sum_{i\in[m]}(d_i)^{p-1} - \sum_{i\in[m]}(d_i^\tau)^{p-1}\right| \\ &\le F_p + \sum_{i\in[m]}\left(|B_i^\sigma \cup B_i^\tau|^{p-1} - |B_i^\sigma \cap B_i^\tau|^{p-1}\right) \\ &= (1+\eta_p)F_p. \end{aligned} \tag{19}$$

该引理由（18）、（19）以及 $\frac{1}{t} \ll \frac{\epsilon}{2}$ 这一事实得出。 □

引理 2.17（$\bar{X}$ 不太小） $\mathbf{Pr}\left[\bar{X} < (1-\epsilon-2\eta_p)\cdot F_p\right] < 10^{-5}$.

证明：对于任意 $j \in [t]$，令 $Y_j = \min\left\{X_j, m(d_I^\tau)^{p-1}\right\}$。下面给出几个断言，其证明将在本章节末尾给出。

命题 2.18 对于任意 $j \in [t]$，$(1-2\eta_p)F_p \le \mathbf{E}[Y_j] \le F_p$。

断言 2.19 对于任意 $j \in [t]$，方差$[Y_j]$ $\frac{(\mathbf{E}[Y_j])^2}{(\mathbf{E}[Y_j])^2} \le \frac{m^{1-1/p}}{1-2\eta_p}$。

现在我们有

$$\begin{aligned}\mathbf{Pr}\left[\bar{X} < (1-\epsilon-2\eta_p)\,F_p\right] &= \mathbf{Pr}\left[\frac{1}{t}\sum_{j\in[t]}X_j - (1-2\eta_p)F_p < -\epsilon F_p\right] &(20)\\ &\le \mathbf{Pr}\left[\frac{1}{t}\sum_{j\in[t]}Y_j - \mathbf{E}[Y_j] < -\epsilon F_p\right] &(21)\\ &\le \frac{\mathbf{Var}[Y_j]}{t\epsilon^2 F_p^2} &(22)\\ &\le \frac{\mathbf{Var}[Y_j]}{t\epsilon^2(\mathbf{E}[Y_j])^2} &(23)\\ &\le \frac{1}{t\epsilon^2}\cdot\frac{m^{1-1/p}}{1-2\eta_p} &(24)\\ &\le 10^{-5}, &(25)\end{aligned}$$

从（20）到（21），我们使用了以下事实：$X_j \ge Y_j$ 根据 Y_j 的定义以及命题2.18的第一个不等式。从（21）到（22），我们使用切比雪夫不等式。从（22）到（23），我们使用命题2.18的第二个不等式。从（23）到（24），我们应用命题2.19。最后一个不等式由我们的假设 $\eta_p \le 0.4$ 以及我们对取值 t（算法1 的第3行）作出的选择推出。□

现在我们可以证明定理 2.13。回想一下，我们在算法 $\ell = 100$ 中设置了 3。令$X_{\min} = \min\left(\bar{X}_1, \ldots, \bar{X}_\ell\right)$ 为算法 3 的输出。根据引理 2.16 和 2.17，我们有

$$\begin{aligned}\mathbf{Pr}\left[X_{\min} < (1+\epsilon+\eta_p)F_p\right] &\ge 1 - \left(1-\frac{1}{13}\right)^\ell > 0.999, \quad \text{and}\\ \mathbf{Pr}\left[X_{\min} < (1-\epsilon-2\eta_p)F_p\right] &\le \ell\cdot 10^{-5} = 0.001.\end{aligned}$$

该定理由并集界得出。

最后，我们证明引理 2.17 的证明中用到的两个结论。我们的证明对每个$j \in [t]$都成立。因此，为方便起见，我们省略 j 在 X_j 和 Y_j中的下标。

证明：[结论 2.18] 的证明。我们先考察 X 与 Y之间的差异，

$$\begin{aligned}
\mathbf{E}[X-Y] &= \sum_{i\in[m]} \frac{1}{m} \cdot \mathbf{E}[X-Y \mid I=i] = \sum_{i\in[m]} \max\left\{(d_i)^{p-1} - (d_i^\tau)^{p-1}, 0\right\} \\
&\le \sum_{i\in[m]} \left(|B_i^\sigma \cup B_i^\tau|^{p-1} - |B_i^\sigma \cap B_i^\tau|^{p-1}\right) \\
&= \eta_p F_p. \qquad (26)
\end{aligned}$$

因此，我们有

$$\begin{aligned}
\mathbf{E}[Y] &= \mathbf{E}[X] - \mathbf{E}[X-Y] \\
&\ge \sum_{i\in[m]} (d_i)^{p-1} - \eta_p F_p \qquad \text{(by (26) and Lemma 2.15)} \\
&\ge F_p - \left|\sum_{i\in[m]} (d_i)^{p-1} - \sum_{i\in[m]} (d_i^\tau)^{p-1}\right| - \eta_p F_p \quad \text{(by Fact 2.14)} \\
&\ge (1-\eta_p)F_p - \sum_{i\in[m]} \left(|B_i^\sigma \cup B_i^\tau|^{p-1} - |B_i^\sigma \cap B_i^\tau|^{p-1}\right) \\
&= (1-2\eta_p)F_p.
\end{aligned}$$

对于另一个方向，根据事实2.14 我们有

$$\mathbf{E}[Y] = \frac{1}{m}\sum_{i\in[m]} m \cdot \min\left\{(d_i)^{p-1}, (d_i^\tau)^{p-1}\right\} \le \sum_{i\in[m]} (d_i^\tau)^{p-1} = F_p.$$

□

证明：[结论 2.19] 令 n_τ表示 τ中不同元素的数量，$\{V_1, \ldots, V_{n_\tau}\}$ 表示 G^τ中的团集合。根据 Y的定义，我们有

$$Y \le m \cdot \max_{i\in[m]} |B_i^\tau|^{p-1} = m \cdot \max_{k\in[n_\tau]} |V_k|^{p-1},$$

这意味着

$$\mathbf{Var}[Y] \le \mathbf{E}[Y^2] \le m \cdot \max_{k\in[n_\tau]} |V_k|^{p-1} \cdot \mathbf{E}[Y]. \qquad (27)$$

注意到

$$\max_{k\in[n_\tau]} |V_k|^{p-1} = \left(\max_{k\in[n_\tau]} |V_k|^p\right)^{1-1/p} \le \left(\sum_{k\in[n_\tau]} |V_k|^p\right)^{1-1/p} = F_p^{1-1/p}. \qquad (28)$$

根据 H¨older 不等式,

$$m = \sum_{k\in[n_\tau]} |V_k| \le n_\tau^{1-1/p} \left(\sum_{k\in[n_\tau]} |V_k|^p \right)^{1/p} = n_\tau^{1-1/p} F_p^{1/p}. \tag{29}$$

结合式（27）、（28）、（29）以及命题 2.18的第一个不等式，可得

$$\frac{\mathbf{Var}[Y]}{(\mathbf{E}[Y])^2} \le \frac{m \cdot \max_{k\in[n_\tau]} |V_k|^{p-1}}{\mathbf{E}[Y]} \le \frac{n_\tau^{1-1/p} F_p^{1/p} \cdot F_p^{1-1/p}}{(1-2\eta_p)F_p} \le \frac{m^{1-1/p}}{1-2\eta_p}.$$

□

3 下界

我们现在探讨下　界，首先概述关键思路和证明技巧。

思路与技术概述。 我们利用经典的多方通信问题不相交性问题$_{n,\ k}$ 来证明下界。在不相交性问题$_{n,\ k}$中，有 k 个参与者。每位参与者 κ ($\kappa \in [k]$) 都会得到一个向量 $X^{(\kappa)} = \left(X_1^{(\kappa)}, \ldots, X_n^{(\kappa)}\right) \in \{0,1\}^n$，它表示 $[n]$ 的一个子集。我们交替使用向量和集合的表示方式。不相交性问题$_{n,\ k}$的输入 $\{X^{(1)}, \ldots, X^{(k)}\}$ 满足以下两种情况之一：

1. *NO* 实例：所有 k 个子集两两不相交。也就是说，对于任意 $\kappa \neq \kappa'$, $X^{(\kappa)} \cap X^{(\kappa')} = \emptyset$。

2. *YES* 实例：它们有一个唯一的公共元素，其余元素则互不相交。也就是说，$\left|\bigcap_{\kappa\in[k]} X^{(\kappa)}\right| = 1$ 并且对于任意 $\kappa \neq \kappa'$, $X^{(\kappa)} \cap X^{(\kappa')} = \cap_{\kappa\in[k]} X^{(\kappa)}$。

问题是让 k 参与者通过通信判断输入是 YES 实例还是 NO 实例。DISJ$_{n,\ k}$ 最初是在黑板模型中研究的，在该模型中，每当一位玩家发送消息时，其他所有（$k-1$）玩家都能看到。显然，在黑板模型中证明的任何下界也适用于协调器模型；在协调器模型中，k 站点充当 k玩家，并通过协调器彼此通信。关于 DISJ$_{n,\ k}$，已知如下结果。若一个算法以概率至多 δ出错，则称该算法为 δ-错误算法。

定理 3.1 ([9]) 任何针对不相交性问题、错误率为0.49的算法，在黑板模型中都需要$\Omega\left(k\, n \log k\right)$通信比特。

上述不相交性问题$_{n,\ k}$ 被用于证明无噪声数据流模型 [9]中该问题的 $\tilde{\Omega}\left(m^{1-2/p}\right)$ 空间下界。标准归约如下：给定一个用于解决 F_p的(ϵ, δ)-近似流式算法 $\mathcal{A}$，我们可以按以下方式解决参数为 $n = \Theta(m)$和 $k = ((1+3\epsilon)n)^{1/p}$ 的不相交性问题$_{n,\ k}$：每位参与者 $\kappa \in [k]$ 通过纳入每个 $i \in X^{(\kappa)}$ 来构造一个元素集合 $\sigma^{(\kappa)}$。最终数据流 σ 是所有 $\sigma^{(\kappa)}$的串接，即$\sigma = \sigma^{(1)} \circ \ldots \circ \sigma^{(k)}$。$k$ 位参与者可以如下模拟流式算法 $\mathcal{A}$ 来解决不相交性问题$_{n,\ k}$：第一位参与者在 $\sigma^{(1)}$ 上运行 $\mathcal{A}$，并将最终内存配置 M_1 写到黑板上。第二位参与者从 M_1开始，在 $\sigma^{(2)}$ 上继续运行 $\mathcal{A}$，然后将最终内存配置 M_2 写到黑板上，依此类推。最后，最后一位参与者在

流式算法输出一个大于 (F_p的估计$1+\epsilon)n$时输出 YES，否则输出 NO。如果允许多遍处理，那么最后一位参与者可以将其最终内存配置写到黑板上，并从第一位参与者开始重新进行模拟。如果流式算法 $\mathcal{A}$ 使用 $O(1)$ 遍且占用 s比特空间，那么 k 位参与者解决不相交性问题$_{n,\ k}$ 所需的通信比特至多为 $O(sk)$。根据定理 3.1，有 $sk=\Omega\left(_{k}{}^{n}\log k\right)$，因此，$s=\Omega\left(_{k^2}{}^{n}\log k\right)=\Omega\left(^{m^{1-2/p}}\log m\right)$。

然而，这种证明策略并不普遍适用于协调器模型，因为该归约需要 $k=\Omega\left(m^{1/p}\right)$。事实上，在协调器模型中，存在使用 $\tilde{O}\left(\frac{k^{p-1}}{\epsilon^2}\right)$ 通信比特 [17, 23]的 $(\epsilon, 0.01)$-近似算法来处理 F_p，且仅需 k 个参与者。我们的关键观察是，在噪声数据设定中，只需常数个参与者即可完成类似的归约。此外，该证明仅要求数据集具有较小的 F_p-失配歧义度，$\eta_p=O(\epsilon)$；在此条件下，$(\epsilon+O(\eta_p)\ \delta)$-近似等价于 $(O(\epsilon)\ \delta)$-近似。

更准确地说，对于每个元素 $i\in[n]$，每位参与者 $\kappa\in[k]$都会添加 t/k 个两两不相似的元素（参数 t 的具体取值将在后文给出）；用 $Q_i^{(\kappa)}$ 表示这些元素。我们还要求：若两个参与者 κ和 κ' 在不相交性问题$_n$实例中共享一个元素 i，那么对于任意一对元素 $u\in Q_i^{(\kappa)}$ 和 $v\in Q^{(\kappa')}{}_i$，均有 $u\sim v$。反之，若他们不共享任何元素，则他们创建的所有元素都两两不相似。值得注意的是，由于传递性约束，在无噪声设定中无法满足这一要求：若 $Q_i^{(\kappa)}$ 中的所有元素都与 $Q^{(\kappa')}{}_i$ 中的元素相似，那么根据传递性，$Q_i^{(\kappa)}$ 中的所有元素必然彼此相似。相比之下，我们将证明，在噪声数据设定中，可以构造一个显式数据集和一个恰当的相似度函数，使两项要求以 0.99 的概率同时成立。

我们给出两个具体的构造。第一个更加自然——其相似性预言机由一个自然的距离函数诱导而来。更准确地说，令 $s=\Theta\left(^{t^2}\log_{k^2}{}^{t}\right)$ 为一个参数。对于每个玩家~~$\kappa\in$~~$[k]$ 以及每个 $i\in[n]$，我们考虑 DISJ$_{n,\ k}$ 实例中的两种情况：

1. 如果 $X_i^{(\kappa)}=1$，玩家随机将集合 $\{(i,\ k+1、1).....(i,\ k+1,\ s)\}$ 划分为 t/k 个大小相等的子集，并令项目 $q_i^{(\kappa)}{}_{,\ j}$ 为第 j个划分。不同玩家使用的随机性彼此独立。

2. 如果 $X^{(\kappa)}_{\ i}=0$，玩家任意地将集合 $\{(i,\ \kappa、1).....(i,\ \kappa,\ s)\}$ 划分为 t/k 个大小相等的子集，并令项目 $q_i^{(\kappa)}{}_{,\ j}$ 为第 j个划分。

接着我们令 $Q_i^{(\kappa)}=\left\{q_{i\,1}^{(\kappa)},,\ldots,q_i^{(\kappa)}{}_{,\frac{t}{k}}\right\}$。

这个构造满足我们所需的所有性质。唯一的问题是，每个项目都需要 $O\left(m^{1/p}\log m\right)$ 比特来描述，这导致上界与下界之间存在显著的字与比特间隙。因此，我们提出第二种略显人为的构造，它仍满足所有必要的性质，同时每个项目只需用对数数量的比特即可描述。

现在，让我们看一下不相交性问题$_{n,k}$的 YES 实例。设 i^* 为所有参与者所共享的元素。容易看出，由 i^* 中的 k 名参与者在 σ_Y中创建的元素集合构成一个完全 k部图，它离一个团并不远。根据我们对 F_p-失配歧义η_p的定义，存在一个无噪声数据集 τ_Y，使得当它被用作真实值时，我们所

从不相交性问题$_{n,k}$ 的 YES 实例为 F_p 构造的结果具有 $\eta_p(\sigma_Y, \tau_Y) \leq \frac{10\epsilon p}{k}$，通过选择 $k = 10p(C+1)$（一个常数！），其至多为 $\frac{\epsilon}{C+1}$。

另一方面，对于 DISJ$_{n,k}$ 的 NO 实例，所得数据集 σ_N for F_p 的图表示仅由单节点组成，对应于一个无噪声数据集 τ_N with$\eta_p(\sigma_N, \tau_N) = 0$。通过选择参数 $t = (10\epsilon m)^{1/p}$ 和 $n = \frac{m}{t}$，我们可以证明 $(1 - \epsilon - C\eta_p(\sigma_Y, \tau_Y))F_p(\tau_Y) > (1 + \epsilon + C\eta_p(\sigma_N, \tau_N))F_p(\tau_N)$。因此，任何 $((\epsilon + C\eta_p)\ \delta)$-近似流式算法，只要用于求解 F_p，就能够以至多 $(\delta + 0.01)$ 的错误概率求解 DISJ$_{n,k}$ 问题，其中术语 0.01 计入了输入归约的错误概率。因此，定理 3.1 意味着，在含噪数据流中对 F_p 进行 $(($ $\epsilon + C\eta)$ 0.48)-近似需要 $\Omega\left(\frac{1}{\epsilon^{1/p}} m^{1-1/p}\right)$ 的空间下界。

使用相同的方法，我们可以证明协调器模型中的通信下界为 $\Omega\left(\frac{1}{\epsilon^{1/p}} m^{1-1/p}\right)$ 比特。

下界。 我们证明了以下关于含噪数据集上 F_p-估计的多方通信下界。

定理 3.2 对于任意常数 $p > 1$、$C \geq 0$ 和 $\epsilon \in (0、13)$，任何用于计算任意含噪数据集上 F_p 的 $((\epsilon + C\eta_p)\ 0.48)$-近似算法（数据集最多包含 m 个元素，并分布在 $k \geq 10p(C+1)$ 个参与者之间），在黑板模型中都需要 $\Omega\left(\frac{1}{\epsilon^{1/p}} m^{1-1/p}\right)$ 比特的通信。

该定理直接推出两个推论，一个针对数据流模型，另一个针对协调器模型。在数据流模型与 k-方通信之间的标准对应关系中（参见技术概述），我们令参与者数量 $k = 10p(C+1) = \Theta(1)$。

推论 3.3 对于任意常数 $p > 1$、$C \geq 0$ 和 $\epsilon \in (0、13)$，任何用于计算 F_p 的 $O(1)$ 遍历 $(($ $\epsilon + C\eta_p)\ 0.48)$-近似算法，在任意长度至多为 m 的含噪数据流上，都需要$\Omega\left(\frac{1}{\epsilon^{1/p}} m^{1-1/p}\right)$ 比特空间。

推论 3.4 对于任意常数 $p > 1$、$C \geq 0$ 和 $\epsilon \in (0、13)$，任何用于在协调器模型中计算 F_p 的$(($ $\epsilon + C\eta_p)\ 0.48)$-近似算法，在任意至多含有 m 个元素、分布于 $k \geq 10p(C+1)$ 个站点的含噪数据集上，都需要$\Omega\left(\frac{1}{\epsilon^{1/p}} m^{1-1/p}\right)$ 通信比特，与轮次数无关。

现在我们证明定理 3.2。约定用“元素”指代不相交性问题 DISJ$_{n,k}$输入中的对象，用“项目”指代 F_p 输入归约之后的对象。

令 $t = (10\epsilon m)^{1/p}$，且 $n = \frac{m}{t} = (10\epsilon)^{-1/p} m^{1-1/p}$。

为简单起见，我们假设 $\epsilon \in (0, 13)$，但通过适当调整 t的常数系数，证明对任意常数 $\epsilon \in (0, 1)$ 都成立。

输入归约。 我们将 DISJ$_{n,k}$中的 $k = 10p(C+1)$ 设为集合。每个玩家 $\kappa \in [k]$ 针对每个元素 $i \in [n]$创建 $\frac{t}{k}$ 个项目 $Q_i^{(\kappa)} = \left\{q_{i\,1}^{(\kappa)}, \ldots, q_{i\ \frac{t}{k}}^{(\kappa)}\right\}$，用于 F_p。令 $\sigma = \cup_{i\in[n]} \cup_{\kappa\in[k]} Q_i^{(\kappa)}$ 为 F_p的归约后输入。σ 中的每个项目 $q_{i\ ,j}^{(\kappa)}$都对应于图表示 G^σ中的一个节点 (i, j, κ)。

在介绍 σ 的详细构造之前，我们首先说明其所需满足的性质。具体构造将在本章节末尾给出。

性质 3.5 经过输入归约后的 F_p 输入 σ 具有以下性质。

1. 对任意 $\kappa \in [k]$ 和 $i \in [n]$，都有 $q_i^{(\kappa)}{}_{,\ j} \not\sim q_i^{(\kappa)}{}_{j'}$，对于任意 $j, j' \in [\frac{t}{k}]$ $(j \neq j')$。换言之，同一参与者基于同一元素创建的项目两两不相似。

2. 对任意 $i \in [n]$ 以及任意 $\kappa, \kappa' \in [k]$ $(\kappa \neq \kappa')$ 且满足 $X_i^{(\kappa)} = X^{(\kappa')}{}_i = 1$，都有 $q_i^{(\kappa)}{}_{,\ j} \sim q_i^{(\kappa')}{}_{,\ j'}$ 对于任意 $j, j' \in [\frac{t}{k}]$。换言之，如果两个参与者 κ, κ' 共享一个元素 i（即 $X_i^{(\kappa)} = X^{(\kappa')}{}_i = 1$），那么参与者 κ 基于 i创建的任何项目，都与参与者 κ' 基于 i创建的任何项目相似。

3. 对于任意 $i \in [n]$ 以及任意 κ，$\kappa' \in [k]$ $(\kappa \neq \kappa')$ 满足 $X_i^{(\kappa)} = 0$ 或 $X^{(\kappa')}{}_i = 0$，我们有 $q_i^{(\kappa)}{}_{,\ j} \not\sim q_i^{(\kappa')}{}_{j'}$，对于任意 j、$j' \in [\frac{t}{k}]$。按字面意思，由一个零项（即 $X_i^{(\kappa)} = 0$）创建的项目与由同一元素创建的任何其他元素都不相似。

4. 对于任意 i，$i' \in [n]$ $(i \neq i')$，我们有 $q_i^{(\kappa)}{}_{,j} \sim q_i^{(\kappa')}{}_{,j'}$ 对于任意 j、$j' \in [\frac{t}{k}]$ 以及 κ、$\kappa' \in [k]$。按字面意思，由不同元素创建的元素是两两不相似的。

容易看出，每个不相交性问题$_{n,k}$ 仅由一个玩家持有的元素 $i \in [n]$ 被映射为 G^σ中的 $\frac{t}{k} \cdot k = t$个单节点，而由所有参与者持有的元素（若存在）则被映射为 G^σ 中的一个完全k部图，且这 k 名参与者各自持有一个大小为$\frac{t}{k}$ 的独立集。

失配歧义： σ。现在，我们考察当 σ 由 DISJ$_n$ 中的 *YES* 实例或 *NO* 实例生成时所具有的性质k。

设 σ_N 是从 NO DISJ$_{n,\ k}$ 实例经输入归约后得到的 F_p 数据集。考虑与 σ_N 相符的真实值 τ_N，我们有

$$\eta_p(\sigma_N, \tau_N) = 0, \quad \text{and} \quad F_p(\tau_N) = n \cdot t = m. \tag{30}$$

设 σ_Y 是从 YES DISJ$_{n,\ k}$ 实例经输入归约后得到的 F_p 数据集。我们考虑如下真实值数据集 τ_Y，其具有 G^{τ_Y}：对于所有 k 参与者都持有的元素 $i^* \in [n]$，我们创建由节点 $\{(i^*,\ j,\ \kappa) \mid j \in [\frac{t}{k}],\ \kappa \in [k]\}$ 构成的团。对于每个 $i \in [n] \backslash \{i^*\}$，我们创建一个由单节点 $\{(i,\ j,\ \kappa) \mid j \in [\frac{t}{k}], \kappa \in [k]\}$ 构成的集合。

我们有

$$F_p(\tau_Y) = (n-1) \cdot t + t^p \in ((1+9\epsilon)m, (1+10\epsilon)m). \tag{31}$$

为了计算 $\eta_p(\sigma_Y,\ \tau_Y)$，我们注意到，$\sigma_Y$和 τ_Y之间唯一的区别来自由 $\{(i^*,\ j,\ \kappa) \mid j \in [\frac{t}{k}],\ \kappa \in [k]\}$索引的元素；也就是说，$G^{\sigma_Y}$中的完全 k部图与 G^{τ_Y}中的团。我们有

$$\begin{aligned}
\eta_p(\sigma_Y, \tau_Y) &= \frac{1}{F_p(\tau_Y)} \sum_{\kappa \in [k]} \sum_{j \in [\frac{t}{k}]} \left(\left| B^\sigma_{(i^*,j,\kappa)} \cup B^\tau_{(i^*,j,\kappa)} \right|^{p-1} - \left| B^\sigma_{(i^*,j,\kappa)} \cap B^\tau_{(i^*,j,\kappa)} \right|^{p-1} \right) \\
&\overset{(31)}{\leq} \frac{1}{m} \cdot k \cdot \frac{t}{k} \cdot \left(t^{p-1} - \left(t - \frac{t}{k} + 1 \right)^{p-1} \right) \\
&\leq \frac{1}{m} \cdot t^p \cdot \left(1 - \left(1 - \frac{1}{k} \right)^{p-1} \right) \leq \frac{1}{m} \cdot t^p \cdot \left(1 - \left(1 - \frac{1}{k} \right)^{p} \right) \\
&\leq 10\epsilon \cdot \frac{p}{k} \qquad (32) \\
&\leq \frac{\epsilon}{C+1}, \qquad (33)
\end{aligned}$$

其中，最后一个不等式成立是因为我们已设定 $k = 10p(C+1)$。

将 DISJ$_{n,k}$ 归约到 F_p。假设我们有一个针对 F_p 的 $(\epsilon + C\eta_p, \delta)$ 近似算法 $\mathcal{A}$，我们按如下方式构造一个针对 DISJ$_{n,k}$ 的算法 $\mathcal{A}'$：给定 DISJ$_{n,k}$ 的输入 I，我们使用输入归约$\tilde{}$创建 F_p 的输入 σ。然后运行 $\mathcal{A}$ 处理 σ，并得到估计值 $\tilde{F}_p$。如果 I 是 YES 实例，那么根据 (31) 和 (33)，以概率 $(1-\delta-o(1))$（额外的 $o(1)$ 误差来自随机化输入归约，稍后我们给出显式构造时会清楚地看到这一点），

$$\tilde{F}_p \geq (1-\epsilon-C\eta_p(\sigma_Y,\tau_Y))F_p(\tau_Y) \geq (1-2\epsilon)(1+9\epsilon)m > (1+\epsilon)m.$$

如果 I 是 NO 实例，那么根据 (30)，我们有概率为 $(1-\delta-o(1)$，

$$\tilde{F}_p \leq (1+\epsilon+C\eta_p(\sigma_N,\tau_N))F_p(\tau_N) \leq (1+\epsilon)m.$$

$\tilde{}$因此，通过检查 $F_p > (1+\epsilon)m$，我们可以判定 DISJ$_{n,\ k}(I) = 1$ 还是 0。由此，任何针对 F_p 的 $((\epsilon + C\eta_p)\ \delta)$ 近似算法都可以用于以概率 $(1-\delta-o(1))$ 求解 DISJ$_{n,\ k}$。根据定理 3.1，任何针对 F_p 的 $((\epsilon + C\eta_p)\ 0.48)$ 近似算法都至少需要 $k^{n}{}_{\log k} = \Omega\left(\frac{1}{\epsilon^{1/p}} m^{1-1/p}\right)$ 个通信比特。

3.1 σ的显式构造

现在我们给出一些 $\sigma = \cup_{i\in[n]} \cup_{\kappa\in[k]} Q_i^{(\kappa)}$ 的显式构造，其中 $Q_i^{(\kappa)} = \left\{q_{i\ 1}^{(\kappa)}, \ , \ \ldots\ldots, \ q_i^{(\kappa)}, \ {}_{k}^{t}\right\}$，满足性质3.5。

我们先介绍一种构造，其中相似度由一种自然的距离函数——集合对称差——定义。这种方法的缺点是，每个项目（一个集合）需要$b = O\left(m^{1/p}\log m\right)$ 比特来表示，因此就 m 而言，上界与下界之间存在 b 的差距，前提是二者都以比特衡量。[5] 为了解决这一问题，我们提出第二种构造，其中每个项目只需用 $O(\log m)$比特表示，因此就 m而言，上界与下界几乎紧致。其代价是，这种情况下的相似度函数并非源自自然的距离函数。尽管如此，只要能够在 σ上定义二元相似度函数，我们的算法就适用于任意数据集 σ。

基于集合对称差的构造。令 $s = {}^{3t^2}{}_{\log_{k^2}}{}^{t}$。对于每个玩家$\kappa \in [k]$ 以及每个 $i \in [n]$，我们考虑两种情形：

1. 如果 $X_i^{(\kappa)} = 1$，该玩家将集合 $\{(i,k+1,1)\ldots,(i,k+1,s)\}$ 随机划分为大小相等的 ${}_{k}^{t}$个子集，并令项目 $q_i^{(\kappa)}{}_{,j}$ 为第 j 个划分。不同参与者所使用的随机性是独立的。

2. 如果 $X_i^{(\kappa)} = 0$，该玩家将集合 $\{(i,\kappa,1)\ldots,(i,\kappa,s)\}$ 任意划分为大小相等的 ${}_{k}^{t}$ 个子集，并令项目 $q_i^{(\kappa)}{}_{,j}$ 为第 j 个划分。

每个项目都可以用 $O\left(s/{}_{k}^{t}\cdot\log(nks)\right) = O\left(m^{1/p}\log m\right)$ 比特来描述。现在我们检查 σ 是否满足性质3.5中的四项要求。

[5]在无噪声情形中，只要我们能够使用哈希函数对每个项目进行子采样，项目大小就不会影响基于子采样的算法的性能。然而，在有噪声情形中，我们必须完整存储所有元素，因为相似度函数可能需要访问它们的完整描述。

首先，由于我们构造中的集合划分，显然，对于任意 $\kappa \in [k]$、 $i \in [n]$ 和j， $j' \in [\frac{t}{k}]$，都有 $\left|q_i^{(\kappa)}{}_{,\ j} \cap q_i^{(\kappa)}{}_{,\ j'}\right| = 0$。

其次，假设两个参与者 κ， $\kappa' \in [k]$ $(\kappa \neq \kappa')$ 共享同一个元素 i^*。令 $z = \frac{sk}{t}$。对于j、$j' \in [\frac{t}{k}]$，我们有

$$\begin{aligned}
\mathbf{Pr}\left[\left|q_{i,j}^{(\kappa)} \cap q_{i,j'}^{(\kappa)}\right| \geq 1\right] &= 1 - \frac{\binom{s-z}{z}}{\binom{s}{z}} = 1 - \frac{(s-z)!/(z!(s-2z)!)}{s!/(z!(s-z)!)} \\
&= 1 - \frac{(s-z)(s-z-1)\cdots(s-2z+1)}{s(s-1)\cdots(s-z+1)} \\
&= 1 - \prod_{a=0}^{z-1}\left(1 - \frac{z}{s-a}\right) \\
&= 1 - \exp\left(\sum_{a=0}^{z-1}\ln\left(1 - \frac{z}{s-a}\right)\right) \\
&\geq 1 - \exp\left(\sum_{a=0}^{z-1} -\frac{z}{s-a}\right) \\
&\geq 1 - \exp\left(-\frac{z^2}{s}\right) = 1 - \frac{1}{t^3}.
\end{aligned}$$

根据并集界，以至少 $1 - t^2 \cdot \frac{1}{t^3} = 1 - o(1)$ 的概率，对于所有 κ， $\kappa' \in [k]$ $(\kappa \neq \kappa')$ 和j， $j' \in [\frac{t}{k}]$，我们有 $\left|q_{i^*}^{(\kappa)}{}_{,\ j} \cap q_{i^*}^{(\kappa')}{}_{,\ j'}\right| \geq 1$。

第三，对于任意 $i \in [n]$ 和任意满足 $\kappa \neq \kappa'$ 的 κ， $\kappa' \in [k]$，若 $X_i^{(\kappa)} = 0$ 或 $X^{(\kappa')}{}_i = 0$，则我们的构造保证 $q_i^{(\kappa)}{}_{,\ j}$ 中所有元组的第二个条目均不同于 $q^{(\kappa')}{}_{i,\ j'}$ 中的条目，对任意 j， $j' \in [\frac{t}{k}]$ 均如此。因此， $\left|q_i^{(\kappa)}{}_{,\ j} \cap q_i^{(\kappa')}{}_{,\ j'}\right| = 0$。

最后，对于任意 i， $i' \in [n]$ $(i \neq i')$，显然，在我们的构造中， $q_i^{(\kappa)}{}_{,\ j}$ 中所有元组的首项都不同于 $q^{(\kappa')}{}_{i,\ j'}$ 中的首项，对任意 j， $j' \in [\frac{t}{k}]$ 和 κ， $\kappa' \in [k]$ 均如此。因此， $\left|q_i^{(\kappa)}{}_{,\ j} \cap q_{i'}^{(\kappa')}{}_{,\ j'}\right| = 0$。

现在，对于两个集合 u， $v \in \sigma$，若定义 $u \sim v$ 当 $|u \cap v| \geq 1$ (或，对称差$|u \oplus v| \leq \frac{2sk}{t} - 2$)，且 $u \not\sim v$ 否则，那么由此得到的鲁棒 F_p 输入能够满足性质 3.5 中所有要求，概率为 $(1 - o(1))$。

一种紧凑物品表示的构造。现在我们给出一种构造，其中每个项目都可以用 $O(\log m)$ 比特来描述。事实上，这种构造比前一种更简单，但其相似度函数并非源自距离度量。

对于每个玩家 $\kappa \in [k]$ 和每个 $i \in [n]$，我们再次考虑两种情况：

1. 若 $X_i^{(\kappa)} = 1$，对于每个 $j \in [\frac{t}{k}]$，$q_i^{(\kappa)}{}_{,\ j} = (i,\ k+1,\ jn + Y_\kappa)$，其中 Y_κ 是从 $\{1, \ldots, n\}$ 中均匀随机抽取的。2. 若 $X_i^{(\kappa)} = 0$，对于每个 $j \in [\frac{t}{k}]$，$q_i^{(\kappa)}{}_{,\ j} = (i,\ \kappa,\ jn + Y_\kappa)$，其中 Y_κ 是从 $\{1, \ldots, n\}$ 中均匀随机抽取的。

符号	描述
$V^{(\kappa)}$	站点 $\kappa \in [k]$ 持有的 $[m]$ 的子集。
$d_i^{(\kappa)} \triangleq \left\|V^{(\kappa)} \cap B_i^{\sigma}\right\|$	站点 κ 持有的与 σ_i 相似的节点数量。
$d_i \triangleq \sum_{\kappa\in[k]} d_i^{(\kappa)} = \|B_i^{\sigma}\|$	σ_i的全局度。
$d_{i,\text{局部}} \triangleq d_i^{(\kappa(i))}$	σ_i的局部度，其中 $\kappa(i) \in [k]$ 是持有 σ_i 的站点。
$D^{(\kappa)}_{\text{局部}} \triangleq \sum_{i\in V^{(\kappa)}} (d_{i,\text{local}})^{p-1}$ $D_{\text{local}} \triangleq \sum_{\kappa\in[k]} D^{(\kappa)}_{\text{局部}} = \sum_{i\in[m]} (d_{i,\text{local}})^{p-1}$	$V^{(\kappa)}$的局部度矩之和。 $[m]$的局部度矩之和。
$d_{i,\text{local}}^{\tau} \triangleq \left\|B_i^{\tau} \cap V^{(\kappa(i))}\right\|$	τ_i的局部度。
$d_i^{\tau} \triangleq \|B_i^{\tau}\|$	τ_i的全局度。

表 1：符号汇总。

We 则将两个项目 $(a,\ b,\ c)$ 和 $(a',\ b',\ c')$ 之间的相似度定义如下：

$$(a,b,c) \sim (a',b',c') \ = \ \begin{cases} \text{true}, & \text{if } (a=a') \wedge (b=b') \wedge (c=c'), \\ \text{true}, & \text{if } (a=a') \wedge (b=b') \wedge (c \neq c' \pmod n), \\ \text{false}, & \text{otherwise.} \end{cases}$$

现在我们来检验性质 3.5中的四项要求。第三项和第四项要求很容易满足，因为在这两种情况下，相关条件都确保所考虑的两个元素的第一项或第二项不同。第一项要求得以满足，因为 Y_κ 在第 κ个站点中所有节点的构造都相同。对于第二项要求，考虑元素集合 $Q_{i^*} = \{q_{i^*,j}^{(\kappa)} \mid \kappa \in [k], j \in [\frac{t}{k}]\} = \{(i^*, k+1, jm+Y_\kappa) \mid \kappa \in [k], j \in [\frac{t}{k}]\}$，其中 i^*是 DISJ 的一个 YES 实例中的唯一公共元素$_{n,k}$。由于所有 Y_κ 都是从 $\{1,\ldots,n\}$中随机抽取的，$\{Y_1,\ldots,Y_k\}$ 中存在一对相同元素的概率至多为$\frac{k^2}{n} = o(1)$。因此，以概率 $(1-o(1)$，替代构造也满足性质3.5中的所有要求。

4 低歧义情形下的改进分布式算法

如引言中所述，第 3 节中提出的算法3 在 m方面的通信成本渐近最优。然而，我们发现，当失配歧义 η_p 低于某个阈值时，可以设计出通信成本独立于 m的算法。本节针对具有较小 F_p-失配歧义的数据集，在协调器模型下提出一种分布式算法，其通信成本为 poly $\left(k,\ \frac{1}{\epsilon}\right)$ 字。

思路与技术概述。 为便于参考，我们在表1 中总结了一组将在本节余下部分中使用的符号。

与一般情形（第 2.2 节）相同，我们的目标是估计第 $p-1$ 阶度矩 M_{p-1}^{σ}，当 η_p 较小时，该值接近 F_p 。直观地说，我们可以将 η_p 视为对手可用的预算，对手通过 G^{τ} 添加或删除边来扰动真实图。

例如，考虑一个由（$m - m^{1/p}$）个孤立节点和一个大小为 $m^{1/p}$ 的团组成的图，其中这些节点均匀分布在 k 个站点上。当 $\eta_p = \Theta(\epsilon)$ 时，对手可以利用其预算，通过删除两个端点位于同一站点的所有边，将该团转换为一个 k 部图。因此，每个站点都无法在本地确定某个节点究竟属于团，还是仅仅是一个孤立节点——然而，由于团节点对量 M_{p-1}^{σ} 的贡献很大，这一区分至关重要。

对于 η_p 较小的数据集，我们的核心思路是：当对手没有足够的预算来消除所有局部结构信息时，我们便可以利用剩余的局部结构，通过少量通信识别高度数节点。请注意，高度数节点对 M_{p-1}^{σ} 的值贡献占主导。

我们先考虑无噪情形，看看局部信息如何发挥作用。设 σ 是一个无噪声数据集，令 V 表示 G^{σ} 的最大团。对于每个 $\kappa \in [k]$，令 $V^{(\kappa)}$ 表示站点 κ 持有的 V 中的节点。令 $\alpha = \frac{k}{\epsilon}$ 为一个阈值。如果 $\frac{|V|}{|V^{\kappa}()|} > \alpha$，那么 $V_{(\kappa)}$ 很小，且站点 κ 可能无法在没有通信的情况下识别出 $V^{(\kappa)}$ 属于一个大团 V。但这并不是什么大问题，因为 $V_{(\kappa)}$ 的贡献至多为 V 贡献的 $\frac{1}{\alpha}$ 倍，因此可以忽略。另一方面，如果 $\frac{|V|}{|V^{\kappa}()|} \leq \alpha$，那么在 $|V|$ 很大的情况下，$|V_{(\kappa)}|$ 也很大。在这种情况下，站点κ 无需任何通信即可识别出 $V^{(\kappa)}$ 属于一个大团。因此，要估计M_{p-1}^{σ}，只需估计大型 $V_{(\kappa)}$ 的贡献。

为实现这一思想，我们按照以下概率对节点进行采样：$\left\{ {}^{(d_{i,\mathrm{local}})}{}_{D_{\mathrm{local}}}{}^{p-1} \right\}_{i \in [m]}$（即，与其局部度的（$p-1$）次幂成正比；本概述中$d_{i,\ \mathrm{local}}$、$D_{\mathrm{local}}$以及其他新引入符号的定义见表1）。对于每个采样节点 I，我们查询所有 k个站点以获得其全局度 d_I。容易验证，$Y = D_{\mathrm{local}} \cdot \left(\frac{d_I}{d_{I,\mathrm{local}}} \right)^{p-1}$ 是 M_{p-1}^{σ}的无偏估计量。然而，由于方差较大，该估计量不稳定。为解决这一问题，我们引入截断估计量 $X = Y \cdot \mathbf{1}\left\{ {}^{d_{I,\ \mathrm{local}}}_{d_I} \leq \alpha \right\}$。$X$ 是 M_{p-1}^{σ}的有偏估计量，其偏差来自全局与局部度数比超过α的节点（即来自小 $V_{(\kappa)}$的节点），这部分节点至多占 M_{p-1}^{σ} 的 $\frac{k}{\alpha}(= \epsilon)$。我们还可以证明，$O\left(\frac{\alpha^{p-1}}{\epsilon^2} \right) = O\left(\frac{k^{p-1}}{\epsilon^{p+1}} \right)$ 个样本足以对$\mathbf{E}[X]$进行准确近似。

现在考虑数据集具有较小歧义的情形。若要使本方法失效，对手必须从较大的 $V^{(\kappa)}$中选取一些节点，并通过添加或删除边来改变它们的全局与局部度比，以便将其推到阈值 α之上。为此，我们证明对手最有效的策略是删除局部边。作为一个简单的例子，考虑 $p = 2$ 以及满足$d_i^{\tau}/d_{i,\ \mathrm{local}}^{\tau} = \alpha/2$的节点 i。要将$d_i^{\tau}/d_{i,\ \mathrm{local}}^{\tau}$ 提高到 α，对手需要删除大约$d_{i,\ \mathrm{local}}^{\tau}/2 = d_i^{\tau}/\alpha$ 条与节点 i 相关联的局部边，从而将节点 i从低全局与局部度比节点的集合中剔除。由于节点 i 对 F_2 的贡献为 $d_i \leq d_i^{\tau}$，该操作至多使 X 的偏差增加 d_i^{τ}。因此，在 $\eta_2 F_2$次边插入/删除的预算下，对手至多只能使输入的偏差增加 $\alpha\eta_2 F_2$，而当 $\eta_2 \leq \frac{\epsilon}{\alpha} = O\left(\frac{\epsilon^2}{k} \right)$时，该增量至多为 ϵF_2。我们可以将这一论证推广到一般情形以及任意 $p \geq 2$，从而表明当$\eta_p \leq O\left(\frac{\epsilon^p}{k^{p-1}} \right)$时，$X$ 的偏差至多为 ϵF_p。

算法与分析。算法的描述见算法 4。下面用文字简要说明该算法。第一轮中，每个站点将其节点的局部（$p-1$）阶度矩之和发送给

协调器，协调器将这些值汇总，得到所有 m 节点的 D_{local}（$p-1$）阶总度矩。第二轮中，站点与协调器共同对每个节点进行采样，采样概率与其局部（$p-1$）阶度矩成正比。第三轮中，计算每个采样节点的精确全局度。最后，协调器利用 D_{local} 以及采样节点的局部和全局（$p-1$）阶度矩，构造最终的截断估计量。

算法 4： 分布式鲁棒-F_p-低噪声

输入： 一个大小为 m 、划分到 k 个站点的含噪数据集 σ，其中第 κ 个站点持有 $\sigma^{(\kappa)}$；一个参数 ϵ

输出： F 的 $((\epsilon + O(\eta_p))\ 0.01)$-近似 p

$t \leftarrow \frac{36 \cdot 4^p \cdot k^{p-1}}{\epsilon^{p+1}}$， $\theta \leftarrow 2\left(\frac{4k}{\epsilon}\right)^{p-1}$

// 第一轮：

协调器：向所有站点发送“开始”。

对于站点 $\kappa \leftarrow 1, \ldots, k$ **do**

  $D^{(\kappa)}_{\text{本地}} \leftarrow \sum_{i \in V^{(\kappa)}} (d_{i,\text{local}})^{p-1}$

  发送 $D^{(\kappa)}_{\text{本地}}$ 发送给协调器

// 第二轮：

协调器：计算 $D_{\text{local}} \leftarrow \sum_{\kappa \in [k]} D^{(\kappa)}$ 本地。然后，从 $[k]$ 中抽取 t 独立同分布样本 根据概率 $\left\{ \frac{D^{(\kappa)}}{D_{\text{local}}} \right\}_{\kappa \in [k]}$。令 s_κ 为抽取 κ的次数。

对每个 $\kappa \in [k]$，将 s_κ 发送到站点 κ。

用于站点 $\kappa \leftarrow 1, \ldots, k$ **do**

  从中抽取 s_κ 个独立同分布样本 $V^{(\kappa)}$ 根据概率 $\left\{ \frac{(d_{i,\ \text{local}}\ ^{p-1})_{\text{本地}}}{D^{(\kappa)}} \right\}_{i \in V^{(\kappa)}}$

  将所有采样节点发送给协调器

// 第三轮：

协调器：令 S 为采样节点的集合。将 S 发送给所有站点。

对于 站点 $\kappa \leftarrow 1, \ldots, k$ **do**

  对每个 $i \in S$ **执行**

    $d_i^{(\kappa)} \leftarrow \left| V^{(\kappa)} \cap B_i^\sigma \right|$

  将 $\{d_i^{(\kappa)}\}_{i \in S}$ 发送给协调器

// 输出：

协调器：对每个 $i \in S$ 计算 $d_i \leftarrow \sum_{\kappa \in [k]} \mathbb{1}\left\{ \frac{d_i^{(\kappa)} d_{\varphi}}{d_{i,\ \text{本地}}} \right\}$

返回 $\frac{1}{t} \sum_{i \in S} \left(D_{\text{local}} \cdot \left(\frac{d_i}{d_{i,\text{local}}} \right)^{p-1} \cdot \mathbb{1}\left\{ \left(\frac{d_i}{d_{i,\text{local}}} \right)^{p-1} \leq \theta \right\} \right)$。

关于算法4，我们有如下结果。

定理 4.1 对于任意常数 $p \geq 2$ 以及一个含噪的、大小为m 的数据集 σ，其中 $\eta_p \leq \frac{\epsilon^p}{4^{p+1} \cdot k^{p-1}}$，算法 *4* 可以计算出 $((\epsilon + O(\eta_p))0.01)$-近似 F_p，在协调器模型中使用 k 个站点，通过三轮次和 $O\left(\frac{k^p}{\epsilon^{p+1}}\right)$ 个通信字数。

在本章节的其余部分，我们证明定理4.1。轮复杂度可由算

法4。通信成本主要来自 S 站点与协调器之间的通信，所用 $O(kt) = O(k^p/\epsilon^{p+1})$ 字。在本章节的其余部分，我们证明该算法的正确性。

下面的引理表明，算法 4 第 20 行中的每个被加数与 F_p 的偏差都不大。

引理 4.2 对于数据集 σ，令 $I \in S$ 为从算法 4 中采样的节点，且 $X = D_{local} \cdot \left(\frac{d_I}{d_{I,local}}\right)^{p-1} \cdot \mathbf{1}\left\{\left(\frac{d_I}{d_{I,local}}\right)^{p-1} \le \theta\right\}$。有 $\left(1-2\eta_p-\frac{\epsilon}{2}\right)F_p \le \mathbf{E}[X] \le (1+\eta_p)F_p$.

证明：我们从第二个不等式开始。令 $p_i = \mathbf{Pr}[I=i]$。有

$$\begin{aligned} p_i &= \mathbf{Pr}[\text{site } \kappa(i) \text{ is sampled}] \cdot \mathbf{Pr}[i \text{ is sampled} \mid \text{site } \kappa(i) \text{ is sampled}] \\ &= \frac{D_{\text{local}}^{(\kappa(i))}}{D_{\text{local}}} \cdot \frac{(d_{i,\text{local}})^{p-1}}{D_{\text{local}}^{(\kappa(i))}} = \frac{(d_{i,\text{local}})^{p-1}}{D_{\text{local}}}. \end{aligned}$$

令 $Y = D_{\text{local}} \cdot \left(\frac{d_I}{d_{I,\text{local}}}\right)^{p-1}$。根据 X的定义，有

$$\mathbf{E}[X] \le \mathbf{E}[Y] = \sum_{i\in[m]} p_i \cdot D_{\text{local}} \cdot \left(\frac{d_i}{d_{i,\text{local}}}\right)^{p-1} = \sum_{i\in[m]} (d_i)^{p-1}. \tag{34}$$

注意到

$$\left|\sum_{i\in[m]} (d_i)^{p-1} - \sum_{i\in[m]} (d_i^\tau)^{p-1}\right| \le \sum_{i\in[m]} \left(|B_i^\sigma \cup B_i^\tau|^{p-1} - |B_i^\sigma \cap B_i^\tau|^{p-1}\right) = \eta_p F_p. \tag{35}$$

根据（35）和事实 2.14，有

$$(1-\eta_p)F_p \le \sum_{i\in[m]} (d_i)^{p-1} \le (1+\eta_p)F_p. \tag{36}$$

结合（34）和（36），可得 $\mathbf{E}[X] \le (1+\eta_p)F_p$。

现在我们证明第一个不等式。首先，我们对 X 与 Y 之间的差值进行界定。

$$\begin{aligned} \mathbf{E}[Y-X] &= \sum_{i\in[m]} p_i \cdot D_{\text{local}} \cdot \left(\frac{d_i}{d_{i,\text{local}}}\right)^{p-1} \cdot \mathbf{1}\left\{\left(\frac{d_i}{d_{i,\text{local}}}\right)^{p-1} > \theta\right\} \\ &= \sum_{i\in[m]} (d_i)^{p-1} \cdot \mathbf{1}\left\{\left(\frac{d_i}{d_{i,\text{local}}}\right)^{p-1} > \theta\right\} \\ &\le \sum_{i\in[m]} |B_i^\sigma \cup B_i^\tau|^{p-1} \cdot \mathbf{1}\left\{\left(\frac{d_i}{d_{i,\text{local}}}\right)^{p-1} > \theta\right\} \\ &\le \underbrace{\sum_{i\in[m]} \left(|B_i^\sigma \cup B_i^\tau|^{p-1} - (d_i^\tau)^{p-1}\right)}_{①} + \underbrace{\sum_{i\in[m]} (d_i^\tau)^{p-1} \cdot \mathbf{1}\left\{\left(\frac{d_i}{d_{i,\text{local}}}\right)^{p-1} > \theta\right\}}_{②}. \end{aligned}$$

对于第一部分，

$$① \leq \sum_{i\in[m]} \left(|B_i^\sigma \cup B_i^\tau|^{p-1} - |B_i^\sigma \cap B_i^\tau|^{p-1} \right) = \eta_p F_p. \tag{37}$$

对于第二部分，我们使用阈值 $\alpha = \left(\frac{\theta}{2}\right)^{\frac{1}{p-1}}$ 将 $[m]$ 划分为两个集合：

$$H = \left\{ i \in [m] \;\middle|\; \frac{d_i^\tau}{d_{i,\text{local}}^\tau} > \alpha \right\} \quad \text{and} \quad L = \left\{ i \in [m] \;\middle|\; \frac{d_i^\tau}{d_{i,\text{local}}^\tau} \leq \alpha \right\}.$$

我们利用以下两个断言，分别界定 H 和 L 中元素的贡献。其证明稍后给出。

Claim 4.3

$$\sum_{i\in H} (d_i^\tau)^{p-1} \leq \frac{kF_p}{\alpha}.$$

Claim 4.4

$$\sum_{i\in L} (d_i^\tau)^{p-1} \cdot \mathbf{1}\left\{ \left(\frac{d_i}{d_{i,local}} \right)^{p-1} > \theta \right\} \leq \theta \eta_p F_p.$$

根据断言 4.3和断言 4.4，我们有

$$\begin{aligned} ② \;&\leq\; \sum_{i\in H} (d_i^\tau)^{p-1} + \sum_{i\in L} (d_i^\tau)^{p-1} \cdot \mathbf{1}\left\{ \left(\frac{d_i}{d_{i,\text{local}}} \right)^{p-1} > \theta \right\} \\ &\leq\; \frac{kF_p}{\alpha} + \theta\eta_p F_p \leq \frac{\epsilon F_p}{2}, \end{aligned} \tag{38}$$

其中，在最后一个不等式中，我们使用了 $\theta = 2\left(\frac{4k}{\epsilon}\right)^{p-1}$，$\alpha = \left(\frac{\theta}{2}\right)^{\frac{1}{p-1}} = \frac{4k}{\epsilon}$，以及我们的假设 $\eta_p \leq \frac{\epsilon^p}{4^{p+1}\cdot k^{p-1}}$。结合（37）和（38），我们有

$$\mathbf{E}[Y - X] \leq \eta_p F_p + \frac{\epsilon F_p}{2}. \tag{39}$$

根据（34）、（36）和（39），可得

$$\mathbf{E}[X] = \mathbf{E}[Y] - \mathbf{E}[Y - X] \geq \left(1 - 2\eta_p - \frac{\epsilon}{2}\right) F_p.$$

□

最后，我们证明这两个断言。

证明：[证明断言4.3] 设 n_τ为 τ中不同元素的个数，且 $\{V_1,\ldots,V_{n_\tau}\}$为 G^τ中的团集合。注意到

$$\begin{aligned}
\sum_{i\in H}(d_i^\tau)^{p-1} &= \sum_{\kappa\in[k]}\sum_{j\in[n_\tau]}\sum_{i\in V^{(\kappa)}\cap V_j}(d_i^\tau)^{p-1}\cdot\mathbf{1}\left\{\frac{d_i^\tau}{d_{i,\text{local}}^\tau}>\alpha\right\} &(40)\\
&= \sum_{\kappa\in[k]}\sum_{j\in[n_\tau]}\sum_{i\in V^{(\kappa)}\cap V_j}|V_j|^{p-1}\cdot\mathbf{1}\left\{\frac{|V_j|}{\left|V^{(\kappa)}\cap V_j\right|}>\alpha\right\} &(41)\\
&= \sum_{\kappa\in[k]}\sum_{j\in[n_\tau]}|V_j|^{p-1}\cdot\left|V^{(\kappa)}\cap V_j\right|\cdot\mathbf{1}\left\{\frac{|V_j|}{\left|V^{(\kappa)}\cap V_j\right|}>\alpha\right\}\\
&< \sum_{\kappa\in[k]}\sum_{j\in[n_\tau]}\frac{|V_j|^p}{\alpha}=\sum_{\kappa\in[k]}\frac{F_p}{\alpha}=\frac{kF_p}{\alpha},
\end{aligned}$$

其中，从（40）到（41），我们利用了以下事实：若 $i\in V_j\cap V^{(\kappa)}$，则 $d_i^\tau=|B_i^\tau|=|V_j|$ 和 $d_{i,\text{local}}^\tau=\left|B_i^\tau\cap V^{(\kappa)}\right|=\left|V_j\cap V^{(\kappa)}\right|$。□

证明：[命题的证明 4。4] 对于任意 $i\in L$ 满足 $\left(\frac{d_i}{d_{i,\text{local}}}\right)^{p-1}>\theta$，有

$$\begin{aligned}
\theta &< \frac{(d_i)^{p-1}}{(d_{i,\ \text{local}})^{p-1}} &(42)\\
&\le \frac{|B_i^\sigma\cup B_i^\tau|^{p-1}}{\left|B_i^\sigma\cap B_i^\tau\cap V^{(\kappa(i))}\right|^{p-1}} &(43)\\
&= \frac{(d_i^\tau)^{p-1}+\left(|B_i^\sigma\cup B_i^\tau|^{p-1}-(d_i^\tau)^{p-1}\right)}{(d_{i,\text{local}}^\tau)^{p-1}-\left((d_{i,\text{local}}^\tau)^{p-1}-\left|B_i^\sigma\cap B_i^\tau\cap V^{(\kappa(i))}\right|^{p-1}\right)}, &(44)
\end{aligned}$$

其中，从（42）到（43）我们使用 $d_i=|B_i^\sigma|\le|B_i^\sigma\cup B_i^\tau|$ 和 $d_{i,\ \text{local}}=\left|B_i^\sigma\cap V^{(\kappa(i))}\right|\ge\left|B_i^\sigma\cap B_i^\tau\cap V^{(\kappa(i))}\right|$。对（44）中的各项重新排列可得

$$\begin{aligned}
\theta(d_{i,\text{local}}^\tau)^{p-1}-(d_i^\tau)^{p-1} &\le \left(|B_i^\sigma\cup B_i^\tau|^{p-1}-(d_i^\tau)^{p-1}\right)+\theta\left((d_{i,\text{local}}^\tau)^{p-1}-\left|B_i^\sigma\cap B_i^\tau\cap V^{(\kappa(i))}\right|^{p-1}\right)\\
&\le \left(|B_i^\sigma\cup B_i^\tau|^{p-1}-(d_i^\tau)^{p-1}\right)+\theta\left((d_i^\tau)^{p-1}-|B_i^\sigma\cap B_i^\tau|^{p-1}\right)\\
&\le \theta\left(|B_i^\sigma\cup B_i^\tau|^{p-1}-|B_i^\sigma\cap B_i^\tau|^{p-1}\right), &(45)
\end{aligned}$$

其中，在第二个不等式中，我们利用了这样一个事实：对于任意的 $p\ge2$ 以及任意三个集合 A、B、C，只要满足 $B\subseteq A$，就总有 $|A|^{p-1}-|B|^{p-1}\ge|A\cap C|^{p-1}-|B\cap C|^{p-1}$。最后一个不等式成立是因为 $\theta\ge1$。

另一方面，由于 $i\in L$和 $\alpha=\left(\frac{\theta}{2}\right)^{\frac{1}{p-1}}$，我们有

$$\theta(d_{i,\text{local}}^\tau)^{p-1}-(d_i^\tau)^{p-1}\ge\theta\left(\frac{d_i^\tau}{\alpha}\right)^{p-1}-(d_i^\tau)^{p-1}=(d_i^\tau)^{p-1}. \tag{46}$$

结合式（45）和式（46），对于任意满足 $i \in L$ 的 $\left(\frac{d_i}{d_{i,\text{local}}}\right)^{p-1} > \theta$,

$$(d_i^\tau)^{p-1} \le \theta \left(|B_i^\sigma \cup B_i^\tau|^{p-1} - |B_i^\sigma \cap B_i^\tau|^{p-1} \right).$$

由此可得

$$\sum_{i \in L} (d_i^\tau)^{p-1} \cdot \mathbf{1}\left\{ \left(\frac{d_i}{d_{i,\text{local}}}\right)^{p-1} > \theta \right\} \le \sum_{i \in [m]} \theta \left(|B_i^\sigma \cup B_i^\tau|^{p-1} - |B_i^\sigma \cap B_i^\tau|^{p-1} \right) = \theta \eta_p F_p.$$

□

引理 4.5 设 $\bar{X} = \frac{1}{t} \sum_{i \in [t]} X_i$，其中 X_i 的定义与引理 *4.2* 中对 S 的第 i 个样本的定义相同。则以至少 2 3 的概率，$(1 - 2\eta_p - \epsilon) F_p \le \bar{X} \le (1 + \eta_p + \epsilon) F_p$.

证明：我们首先给出 X_i 方差的上界。

$$\begin{aligned}
\mathbf{Var}[X_i] &\le \mathbf{E}[X_i^2] \\
&\le \theta \cdot D_{\text{local}} \cdot \mathbf{E}[X_i] \quad (\text{since } X_i \le \theta \cdot D_{\text{local}}) \\
&= \theta \cdot \left(\sum_{i \in [m]} (d_{i,\text{local}})^{p-1} \right) \cdot \mathbf{E}[X_i] \le \theta \cdot \left(\sum_{i \in [m]} (d_i)^{p-1} \right) \cdot \mathbf{E}[X_i] \\
&\le \theta \cdot (1 + \eta_p) F_p \cdot \mathbf{E}[X_i] \quad (\text{by } (36)) \\
&\le \theta \cdot \frac{1 + \eta_p}{1 - 2\eta_p - \frac{\epsilon}{2}} \cdot \mathbf{E}[X_i]^2 \quad (\text{by Lemma } 4.2) \\
&\le 6\theta \cdot \mathbf{E}[X_i]^2. \quad (\text{by assumption } \eta_p \le \epsilon^p / (4^{p+1} \cdot k^{p-1}))
\end{aligned}$$

根据切比雪夫不等式，

$$\mathbf{Pr}\left[\left| \bar{X} - \mathbf{E}[\bar{X}] \right| > \frac{\epsilon}{2} \cdot \mathbf{E}[\bar{X}] \right] \le \frac{4}{t\epsilon^2} \cdot \frac{\mathbf{Var}(X_1)}{\mathbf{E}[X_1]^2} \le \frac{24\theta}{t\epsilon^2} = \frac{1}{3}.$$

根据引理 4.2，以 2 3 的概率，

$$\begin{aligned}
\bar{X} &\le \left(1 + \frac{\epsilon}{2}\right)(1 + \eta_p) F_p \le (1 + \epsilon + \eta_p) F_p, \quad \text{and} \\
\bar{X} &\ge \left(1 - \frac{\epsilon}{2}\right)\left(1 - 2\eta_p - \frac{\epsilon}{2}\right) F_p \ge (1 - \epsilon - 2\eta_p) F_p,
\end{aligned}$$

至此，证明完成。 □

最后，定理4.1 可直接由引理 4.5得出。注意，我们总可以对算法 4应用中位数技巧，将其错误概率降至 *0.01*，而不改变渐近通信复杂度和轮复杂度。

注释 4.6 我们注意到，在存在噪声的情况下，协调器模型中无噪声数据 $\binom{k^{p-1}}{\epsilon^2}$ 所需 F_p 比特的下界 *[35]*仍然成立。因此，就 k而言，定理 *4.1*中的通信开销仅比最优值高 k 倍。

注释 4.7 在我们的下界构造中，由不相交性问题的一个 *YES* 实例得到的 F_p输入n, k其失配歧义至多为 $\frac{c_2\epsilon}{k}$，其中 $c_2 = 10p$为常数。对于 $p = 2$，定理*4.1* 中的歧义阈值最高可达 $\frac{c_1\epsilon^2}{k}$，其中 $c_1 = \frac{1}{4^{p+1}}$为常数。因此， F_2 的通信成本相变发生在 $[\frac{c_1\epsilon^2}{k}, \frac{c_2\epsilon}{k}]$范围内。

5 总结性说明

在这项工作中，我们引入参数 F_p-失配歧义，并利用它为含噪数据上的 F_p 问题设计适用于数据流模型和协调器模型的次线性算法。继这项工作之后，仍有若干方向有待探索。首先，我们的上界与下界之间仍存在 ϵ差距，需要进一步弥合。

其次，在协调器模型中，完整刻画失配歧义与通信成本之间的权衡将很有意义。具体而言，备注 4.7中提到的通信成本相变是平滑的还是突变的?

第三，值得探讨这样一种情形：相似性预言机 f 接收两个元素 σ_i 和 σ_j，并输出 $[0,1]$ 中的一个实数，而非 $\{0,1\}$ 中的一个布尔值来表示它们相似还是不相似。按照惯例，对于所有 $i \in [m]$, 我们令 $f(i,\ i) = 1$。定义 g： $[m] \times [m] \to \{0,1\}$，使得当 σ_i 和 σ_j 对应于同一个真实元素时， $g(i,\ j) = 1$；否则， $g(i,\ j) = 0$。随后，我们可以将 F_p-失配歧义的定义推广如下：

$$\eta_p^{\mathrm{wt}}(\sigma,\tau) = \frac{1}{F_p(\tau)} \sum_{i\in[m]} \left[\left(\sum_{j\in[m]} \max\{f(i,j), g(i,j)\} \right)^{p-1} - \left(\sum_{j\in[m]} \min\{f(i,j), g(i,j)\} \right)^{p-1} \right].$$

我们留下一个开放问题：当以 η_p^{wt}作为近似比中的参数时，是否能够开发出高效的数据流算法与分布式算法来估计 F_p。

最后，在我们的失配歧义框架下研究其他统计问题将会很有意义。

参考文献

[1] Noga Alon、Yossi Matias 和 Mario Szegedy。频率矩近似的空间复杂度。《计算机系统科学杂志》，58(1):137–147，1999年。[2] Alexandr Andoni。利用最大稳定性计算高阶频率矩。发表于*ICASSP*，第6364–6368页，2017年。[3] Alexandr Andoni、Robert Krauthgamer 和 Krzysztof Onak。通过精确采样实现流式算法。发表于*FOCS*，第363–372页，2011年。[4] Ziv Bar-Yossef、T. S. Jayram、Ravi Kumar 和 D. Sivakumar。用于数据流与通信复杂度的信息统计方法。《计算机系统科学杂志》，68(4):702–732，2004年。[5] Lakshminath Bhuvanagiri、Sumit Ganguly、Deepanjan Kesh 和 Chandan Saha。用于估计数据流频率矩的更简单算法。发表于*SODA*，第708–713页，2006年。[6] Mark Braverman 和 Or Zamir。频率矩估计的最优性。发表于*STOC*，第360–370页，2025年。[7] Vladimir Braverman 和 Rafail Ostrovsky。频率矩的递归草图技术。*CoRR*，abs/1011.2571，2010年。

[8] Vladimir Braverman 和 Rafail Ostrovsky。推广 Indyk 和 Woodruff 的分层方法：数据流中基于频率的向量的递归草图。发表于 Prasad Raghavendra、Sofya Raskhodnikova、Klaus Jansen 和 Jos´e D. P. Rolim 编，*APPROX-RANDOM*，计算机科学讲义第8096卷，第58–70页。Springer，2013年。[9] Amit Chakrabarti、Subhash Khot 和 Xiaodong Sun。集合不相交性问题的多方通信复杂度近乎最优的下界。发表于*CCC*，第107–117页。IEEE计算机学会，2003年。.[10] Di Chen 和 Qin Zhang。稳健不同元素的流式算法。发表于 Fatma Ozcan、Georgia Koutrika 和 Sam Madden 编，*SIGMOD*，第1433–1447页。ACM，2016年。[11] Jiecao Chen 和 Qin Zhang。含近重复项的流式数据的不同元素采样。发表于 Jan Van den Bussche 和 Marcelo Arenas 编，*PODS*，第369–382页。ACM，2018年。[12]Graham Cormode 和 Minos N. Garofalakis。概率数据流的草图技术。发表于 Chee Yong Chan、Beng Chin Ooi 和 Aoying Zhou 编，*SIGMOD*，第281–292页。ACM，2007年。[13] Graham Cormode、S. Muthukrishnan 和 Ke Yi。分布式函数监控算法。发表于*SODA*，第1076–1085页，2008年。[14] Graham Cormode、S. Muthukrishnan、Ke Yi 和 Qin Zhang。对分布式数据流进行连续采样。《*ACM*杂志》，59(2):10:1–10:25，2012年。[15] Xin Luna Dong 和 Felix Naumann。数据融合：解决数据冲突以实现集成。《*VLDB*基金会会刊》，2(2):1654–1655，2009年。[16] Ahmed K. Elmagarmid、Panagiotis G. Ipeirotis 和 Vassilios S. Verykios。重复记录检测：综述。*IEEE Trans. Knowl. Data Eng.*，19(1):1–16，2007年。[17] Hossein Esfandiari、Praneeth Kacham、Vahab Mirrokni、David P. Woodruff 和 Peilin Zhong。协调器模型及更广泛场景下经典函数的最优通信界。发表于 Bojan Mohar、Igor Shinkar 和 Ryan O'Donnell 编，*STOC*，第1911–1922页。ACM，2024年。[18] Uriel Feige。关于无界方差的独立随机变量之和，以及图平均度的估计。发表于 L´aszl´o Babai 编，第*36*届*ACM*年度计算理论研讨会论文集，芝加哥，伊利诺伊州，美国，*2004*年*6*月*13*日至*16*日，第594–603页。ACM，2004年。[19] Sumit Ganguly。高阶频率矩的多项式估计量。*CoRR*，abs/1104.4552，2011年。[20]Sumit Ganguly。数据流高阶矩估计的下界。*CoRR*，abs/1201.0253，2012年。[21] Sumit Ganguly 和 David P. Woodruff。高概率频率矩草图。发表于 Ioan- nis Chatzigiannakis、Christos Kaklamanis、D´aniel Marx 和 Donald Sannella 编，*ICALP*，LIPIcs第107卷，第58:1–58:15页。Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik，2018年。

[22] Thomas N Herzog、Fritz J Scheuren 和 William E Winkler。《数据质量与记录链接技术》，第1卷。Springer，2007年。[23] Zengfeng Huang、Zhongzheng Xiong、Xiaoyi Zhu 和 Zhewei Wei。分布式模型中重频项和频率矩的简单最优算法。载于 Michal Kouck´y 和 Nikhil Bansal（编），《第57届ACM计算理论年会论文集，STOC 2025，布拉格，捷克，2025年6月23日至27日》，第371–382页。ACM，2025年。[24] Piotr Indyk 和 David P. Woodruff。数据流频率矩的最优近似。载于《STOC》，第202–208页，2005年。[25] Rajesh Jayaram 和 David P. Woodruff。迈向流式模型和分布式模型中的最优矩估计。《ACM算法汇刊》，19(3):27:1–27:35，2023年。[26] T. S. Jayram、Satyen Kale 和 Erik Vee。概率数据的高效聚合算法。载于 Nikhil Bansal、Kirk Pruhs 和 Clifford Stein（编），《SODA》，第346–355页。SIAM，2007年。[27] T. S. Jayram、Andrew McGregor、S. Muthukrishnan 和 Erik Vee。概率数据流上的统计聚合估计。载于 Leonid Libkin（编），《PODS》，第243–252页。ACM，2007年。[28] Cheqing Jin、Ke Yi、Lei Chen、Jeffrey Xu Yu 和 Xuemin Lin。不确定数据流上的滑动窗口Top-k查询。《VLDB基金会会刊》，1(1):301–312，2008年。[29] Ravi Kannan、Santosh S. Vempala 和 David P. Woodruff。分布式数据的主成分分析与高阶相关性。载于《COLT》，《JMLR研讨会与会议论文集》第35卷，第1040–1057页，2014年。[30] Nick Koudas、Sunita Sarawagi 和 Divesh Srivastava。记录链接：相似度度量与算法。载于《SIGMOD》，第802–803页。ACM，2006年。[31] Yi Li 和 David P. Woodruff。小误差下高频矩估计的紧下界。载于 Prasad Raghavendra、Sofya Raskhodnikova、Klaus Jansen 和 Jos´e D. P. Rolim（编），《APPROX-RANDOM》，《计算机科学讲义》第8096卷，第623–638页。Springer，2013年。[32] Morteza Monemizadeh 和 David P. Woodruff。具有相对误差的单遍lp采样及其应用。载于《SODA》，第1143–1160页，2010年。[33] Jeff M. Phillips、Elad Verbin 和 Qin Zhang。手持输入多方通信复杂度的下界：简明证明。载于 Yuval Rabani（编），《SODA》，第486–501页。SIAM，2012年。[34] David P. Woodruff。所有频率矩的最优空间下界。载于《SODA》，第167–175页，2004年。[35] David P. Woodruff 和 Qin Zhang。分布式函数监测的紧界。载于 Howard J. Karloff 和 Toniann Pitassi（编），《STOC》，第941–960页。ACM，2012年。

[36] Qin Zhang。分布式含噪数据集上的通信高效计算。载于《SPAA》，第313–322页，2015年。[37] Qin Zhang。一般度量空间中含近重复项的流式数据的稳健统计分析。《ACM管理数据学报》，3(2):111:1–111:25，2025年。[38] Qin Zhang 和 Mohsen Heidari。量子数据摘要。载于 Sudeepa Roy 和 Ahmet Kara（编），《ICDT》，《LIPIcs》第328卷，第16:1–16:19页。Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik，2025年。[39] Qin Zhang、Feifei Li 和 Ke Yi。在概率数据中寻找频繁项。载于 Jason Tsong- Li Wang（编），《SIGMOD》，第819–832页。ACM，2008年。