# Frequency Moments in Noisy Streaming and Distributed Data under Mismatch Ambiguity

KAIWEN LIU, Indiana University, USA
QIN ZHANG, Indiana University, USA

We propose a novel framework for statistical estimation on noisy datasets. Within this framework, we focus on the frequency moments ($F_p$) problem and demonstrate that it is possible to approximate $F_p$ of the unknown ground-truth dataset using sublinear space in the data stream model and sublinear communication in the coordinator model, provided that the approximation ratio is parameterized by a data-dependent quantity, which we call the $F_p$-mismatch-ambiguity. We also establish a set of lower bounds, which are tight in terms of the input size. Our results yield several interesting insights:

- In the data stream model, the $F_p$ problem is inherently more difficult in the noisy setting than in the noiseless one. In particular, while $F_2$ can be approximated in logarithmic space in terms of the input size in the noiseless setting, any algorithm for $F_2$ in the noisy setting requires polynomial space.
- In the coordinator model, in sharp contrast to the noiseless case, achieving polylogarithmic communication in the input size is generally impossible for $F_p$ under noise. However, when the $F_p$ mismatch ambiguity falls below a certain threshold, it becomes possible to achieve communication that is entirely independent of the input size.

CCS Concepts: • **Theory of computation → Streaming, sublinear and near linear time algorithms**.

Additional Key Words and Phrases: data streams, near-duplicates, frequency moments, mismatch ambiguity

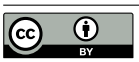
**ACM Reference Format:**
Kaiwen Liu and Qin Zhang. 2026. Frequency Moments in Noisy Streaming and Distributed Data under Mismatch Ambiguity. *Proc. ACM Manag. Data* 4, 2 (PODS), Article 105 (May 2026), 24 pages. https://doi.org/10.1145/3801901

## 1 Introduction

The presence of near-duplicates has long been a challenge in data management. Such noise may arise from diverse sources, such as data collection errors, representational inconsistencies, reformatting or compression artifacts, as well as inherent physical limitations like the imprecision of scientific instruments or the probabilistic nature of quantum systems. For example,

- A large number of queries submitted to a search engine or a large language model, where different combinations of keywords may convey the same semantic intent.
- Webpages, documents or images with near-duplicates are often stored across multiple servers, and our goal is to compute aggregate statistics over these objects.
- If we want to store quantum states classically, then we have to tolerate near-duplicates [39].

K. Liu is partly supported by NSF CCF-1844234.
Q. Zhang is partly supported by NSF CCF-1844234 and IU Luddy Faculty Fellowship.
Authors' Contact Information: Kaiwen Liu, kaiwliu@iu.edu, Indiana University, Bloomington, IN, USA; Qin Zhang, qzhangcs@iu.edu, Indiana University, Bloomington, IN, USA.




ACM 2836-6573/2026/5-ART105
https://doi.org/10.1145/3801901

Typically, a thorough data cleaning step is performed before analysis [15, 16, 22, 30]. However, this approach is infeasible in "big data" settings, where the objective is to compute statistical functions over the dataset without storing or communicating it in its entirety.

To address this challenge, a line of research aims to conduct statistical analysis directly on noisy datasets, mitigating the noise on the fly without explicitly cleaning the data [10, 11, 37, 38]. However, as we shall discuss in the related work (Section 1.1), existing approaches to statistical estimations on noisy datasets exhibit inherent limitations. In this work, we introduce a new framework for analyzing noisy datasets. We then study the $p$-th frequency moments problem ($F_p$) within two established big data models:

- The data stream model [1]: We sequentially process the items in input dataset $\sigma = (\sigma_1, \ldots, \sigma_m)$ in one or multiple passes to compute $F_p(\sigma)$. The main goal is to minimize memory space usage.
- The coordinator model [34]: We have $k$ sites and one central coordinator. The dataset $\sigma$ is partitioned across the $k$ sites, with the $i$-th site holding $\sigma^{(i)}$. The $k$ sites and coordinator want to jointly compute $F_p(\sigma)$ via communication. The computation proceeds in rounds. In each round, the coordinator sends a message to each of the $k$ sites, which then respond to the coordinator with a message. At the end, the coordinator outputs the answer. We would like to minimize both the communication cost and the number of rounds of the computation.

We design algorithms for both models and establish corresponding lower bounds. This paper delivers the following messages:

(1) In the noisy data setting, it is possible to approximate $F_p$ of the *unknown* ground-truth dataset using sublinear space in the data stream model and sublinear communication in the coordinator model, provided that the approximation guarantee is parameterized by a data-dependent quantity, referred to as the *$F_p$-mismatch-ambiguity*.
(2) In the data stream model, we establish tight space bounds for $F_p$ with respect to the input size in the noisy setting, showing that the $F_p$ problem is fundamentally more difficult in the presence of noise than in the noiseless case. In particular, whereas $F_2$ can be approximated in logarithmic space in the noiseless setting by the classical result of Alon, Matias, and Szegedy [1], in the noisy setting any algorithm for $F_2$ requires polynomial space.
(3) In the coordinator model, we show that, unlike in the noiseless setting, achieving polylogarithmic communication in the input size is generally impossible for $F_p$ in the presence of noise. Nevertheless, when the $F_p$-mismatch-ambiguity is below a certain threshold, communication that is *independent* of the input size becomes achievable.

**A Novel Framework for Statistical Estimations on Noisy Data.** We start by introducing our new framework for analyzing noisy data. Let $\sigma = (\sigma_1, \ldots, \sigma_m)$ be a noisy dataset of observed items. Each item $\sigma_i$ is generated from an underlying element $\tau_i$ by the addition of unknown noise. The ground-truth elements $\tau_i$ are drawn from a hidden universe $U = \{u_1, \ldots, u_n\}$. We impose no assumptions on the form or magnitude of the noise. As a result, the set of possible observed items may be infinite and strictly contains the ground-truth universe $U$.

We assume access to an oracle that, for any pair of observed items $\sigma_i$ and $\sigma_j$, determines whether they are similar, denoted $\sigma_i \sim \sigma_j$. Each item is similar to itself, and elements in $U$ are pairwise dissimilar. Note that, because the noise is arbitrary, the similarity relation may produce both false positives and false negatives: two items may be declared similar even if they originate from different ground-truth elements, and two items may be declared dissimilar even if they originate from the same ground-truth element.

We model a dataset $\sigma$ of size $m$ as a graph $G^\sigma = ([m], E^\sigma)$, with node $i$ representing $\sigma_i$. The edge set $E^\sigma$ is defined as $\{(i,j) \in [m]^2 \mid i \le j, \sigma_i \sim \sigma_j\}$. Since we always have $\sigma_i \sim \sigma_i$, each node in $G^\sigma$ contains a self-loop. For each $i \in [m]$, let $B_i^\sigma = \{j \in [m] \mid \sigma_j \sim \sigma_i\}$ be the set of nodes whose corresponding items are similar to $\sigma_i$ (including $\sigma_i$ itself), and let $d_i^\sigma = \left|B_i^\sigma\right|$ be its size (or, the degree of the $i$-th node).

It is easy to see that for a noiseless dataset $\tau$ of length $m$, $G^\tau = ([m], E^\tau)$ contains a set of disjoint cliques, each corresponding to a distinct universe element in $\tau$. We refer to such a graph $G^\tau$ as a *cluster graph*.[1]

For a statistical function $f(\cdot)$ and a noisy dataset $\sigma = (\sigma_1, \ldots, \sigma_m)$, we define $f(\sigma)$ as $f(\tau)$, where $\tau = (\tau_1, \ldots, \tau_m)$ represents the ground truth of $\sigma$, and $\tau_i$ is the ground truth element of $\sigma_i$. Since $\tau$ is noiseless, $f(\tau)$ is always well defined. However, $\tau$ itself is unknown. Our goal is to design an algorithm that approximates $f(\tau)$ using input dataset $\sigma$.

At first glance, this task might seem impossible to achieve. However, as we will demonstrate in this paper using $F_p$ as an example, it is feasible to approximate $f(\tau)$ by introducing a parameter $\eta_f(\sigma, \tau)$ into the approximation ratio. This parameter, called the *$f$-mismatch-ambiguity*, captures how much $\sigma$ and $\tau$ differ with respect to the statistical function $f$.

We start by formally defining frequency moments for a noiseless dataset.

Definition 1.1 (Frequency Moments). *Let $U = (u_1, \ldots, u_n)$ be a finite universe and $\tau = (\tau_1, \ldots, \tau_m) \in U^m$ be a noiseless dataset. For each element $u_j \in U$, define its frequency in $\tau$ as*

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

*The $p$-th frequency moment of $\tau$ is defined as $F_p(\tau) = \sum_{j=1}^n f_j^p$.*

We next define a mismatch-ambiguity parameter for the $F_p$ problem.

Definition 1.2 ($F_p$-mismatch-ambiguity ($p \ge 1$)). *For a noisy dataset $\sigma = (\sigma_1, \ldots, \sigma_m)$ and its ground truth $\tau = (\tau_1, \ldots, \tau_m)$, the $F_p$-mismatch-ambiguity of $\sigma$ with respect to $\tau$ is defined as*

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

*When there is no confusion, we simply write $\eta_p(\sigma, \tau)$ as $\eta_p$.*

We would like to make a few remarks regarding our choice of ambiguity definition for $F_p$.

- When $\sigma$ itself is noiseless, we have for all $i \in [m]$, $B_i^\sigma = B_i^\tau$, and consequently $\eta_p = 0$.
- When $p = 1$, we always have $\eta_1 = 0$. Indeed, for the case when $p = 1$, we just need to count the number of items in the data stream, which is trivial even in the noisy setting.
- We call an edge present in $G^\sigma$ but absent in $G^\tau$ a *false positive* edge, and an edge present in $G^\tau$ but absent in $G^\sigma$ a *false negative* edge. For $p = 2$, the term $\sum_{i \in [m]} \left(\left|B_i^\sigma \cup B_i^\tau\right| - \left|B_i^\sigma \cap B_i^\tau\right|\right)$ in $\eta_2$ is equal to twice of the sum of the false positive and false negative edges. We refer to the false positive and false negative edges as *mismatches*, which motivates the name "mismatch ambiguity" for this family of parameters.
- We observe that the minimum number of mismatch edges in $G^\sigma$, taken over all possible ground truths $\tau$, coincides with the correlation clustering cost of $G^{cor} = ([m], E^{cor})$, where each pair $(i,j)$ satisfying $\sigma_i \sim \sigma_j$ contributes a '+1' edge to $E^{cor}$, and each pair $(i,j)$ with $\sigma_i \nsim \sigma_j$ contributes a '$-1$' edge.[2]

[1] An equivalent and intuitive way to understand a cluster graph is that it is a "P3-free" graph. This means that there is no induced path of length three.

[2] In the correlation clustering problem, we have a complete graph $G = (V, E)$, where each edge in $E$ is labeled either '+1' or '$-1$'. The objective is to partition $V$ into clusters so as to minimize the total number of misclustered edges. An edge is

- We scale the sum $\sum_{i\in[m]}\left(\left|B_i^\sigma \cup B_i^\tau\right|^{p-1} - \left|B_i^\sigma \cap B_i^\tau\right|^{p-1}\right)$ by $\frac{1}{F_p(\tau)}$ to measure the impact of the mismatch edges on the quantity $F_p(\tau)$ we try to estimate. The scaling factor is introduced for convenience; it transforms an additive error into a relative one for the ease of expressing the approximation ratio.

The appropriate definition of $f$-mismatch-ambiguity may vary across different statistical problems. In general, however, this parameter should be expressed as a function of the false positive and false negative edges in the graph representation of the input dataset with respect to the corresponding ground truth.

In this paper, we are mainly interested in the case where the mismatch ambiguity is small, which we believe is common in practice given that we have a good similarity oracle. On the other hand, when the dataset exhibits large mismatch ambiguity, standard statistical measures may lose their significance, and it may become infeasible to estimate $f(\tau)$—where $\tau$ denotes the ground truth—based solely on the noisy input $\sigma$.

**Preliminaries.** In this paper, we always assume that $p$ is a constant. When we write $F_p$, we always mean $F_p(\tau)$, where $\tau$ is the unknown ground truth. We often write $\eta_p(\sigma,\tau)$ as $\eta_p$ for brevity, as $\sigma$ is typically clear from context and $\tau$, the unknown ground truth, is always implicitly defined.

We use $[n]$ to denote $\{1,\dots,n\}$. All logarithms are taken to base 2, unless stated otherwise. When we write $x = a \pm b$, we mean $x \in [a-b, a+b]$. We say $\tilde{X}$ is an $(\epsilon,\delta)$-approximation of $X$ if with probability at least $(1-\delta)$, we have $(1-\epsilon)X \le \tilde{X} \le (1+\epsilon)X$. We assume each item in the dataset fits one word.

Note that in the noiseless setting, upper bounds are typically stated in terms of the universe size $n$ rather than the stream length $m$, under the standard assumption that $n$ and $m$ are polynomially related. In the noisy setting, however, the universe size is not explicitly defined, as it may be unbounded. Accordingly, we assume that the algorithm is given $m$, which can be regarded as an upper bound on the number of distinct items (and hence the effective universe size), since elements absent from the stream can be ignored. If $m$ is not known in advance, an additional pass (in the streaming model) or communication round (in the coordinator model) is required to learn $m$.

To facilitate comparison between our bounds in the noisy setting and the existing results in the noiseless setting, we also use $m$ to denote the universe size in the noiseless setting. Notably, in the hard input constructions for the $F_p$ lower bound in the noiseless setting (e.g., [4, 9]), it consistently holds that $n = \Theta(m)$.

**Our Results.** Our algorithmic result in the data stream model can be stated as follows:

- There is a one-pass $((\epsilon + O(\eta_p)), 0.01)$-approximation algorithm for $F_p$ $(p \in \mathbb{Z}^+)$ over a data stream of $m$ items with mismatch ambiguity $\eta_p \le \frac{1}{3(p!)}$. The algorithm uses $O\left(\frac{1}{\epsilon^2} m^{1-1/p}\right)$ words of space.

We supplement the algorithmic result with the following lower bound:

- For any constant $C \ge 0$, any $O(1)$-pass $((\epsilon + C\eta_p), 0.48)$-approximation algorithm for $F_p$ $(p > 1)$ over a data stream of $m$ items must use at least $\Omega\left(\frac{1}{\epsilon^{1/p}} m^{1-1/p}\right)$ bits of space.

This lower bound result contrasts with the noiseless setting, where for $F_p$ $(p > 1)$, there exists one-pass $(\epsilon, 0.01)$-approximation algorithms (e.g., [8]) for data streams containing $m$ items, using

considered misclustered if it is labeled '+1' and its end nodes belong to different clusters, or if it is labeled '−1' and both end nodes are assigned to the same cluster.

$\tilde{O}\left(\frac{1}{\epsilon^2}m^{1-2/p}\right)$ bits of space.[3] In particular, the space complexity for $p = 2$ is logarithmic in the noiseless setting, but is polynomial in $m$ in the noisy setting.

In the coordinator model, we obtain the following algorithmic result:

- There is a two-round $((\epsilon + O(\eta_p)), 0.01)$-approximation algorithm for $F_p$ $(p \geq 1)$ over a dataset of $m$ items with mismatch ambiguity $\eta_p \leq 0.4$ distributed across $k$ sites. The algorithm uses $O\left(\frac{1}{\epsilon^2}km^{1-1/p}\right)$ words of communication.

For the lower bound, we show:

- For any constant $C \geq 0$, any $((\epsilon + C\eta_p), 0.48)$-approximation algorithm for $F_p$ $(p > 1)$ over a dataset of $m$ items distributed across $k \geq 10p(C + 1)$ sites must use at least $\Omega\left(\frac{1}{\epsilon^{1/p}}m^{1-1/p}\right)$ bits of communication.

This lower bound result is in stark contrast with the noiseless setting, where $(\epsilon, 0.01)$-approximation algorithms for $F_p$ exist using only $O\left(\frac{1}{\epsilon^2}k^{p-1}\log^{O(1)} m\right)$ bits of communication [23]; in particular, the dependency on the input size $m$ is polylogarithmic.

For both the data stream model and the coordinator model, our upper and lower bounds match in terms of $m$, differing only in whether the cost is measured in words or in bits.

Finally, we show that in the coordinator model, when the mismatch ambiguity is sufficiently small, we can design an algorithm that is significantly more efficient, achieving communication cost close to that in the noiseless setting in terms of $k$ and *independent* of the dataset size $m$.

- There is a three-round $((\epsilon + O(\eta_p)), 0.01)$-approximation algorithm for $F_p$ $(p \geq 2)$ over a dataset of $m$ items with mismatch ambiguity $\eta_p \leq \frac{\epsilon^p}{4^{p+1}\cdot k^{p-1}}$ distributed across $k$ sites. The algorithm uses $O\left(\frac{1}{\epsilon^{p+1}}k^p\right)$ words of communication.

For $p = 2$, the mismatch ambiguity threshold in the above algorithm is in the order of $\epsilon^2/k$, and the algorithm achieves a communication cost of $O\left(k^2/\epsilon^3\right)$. While in our lower bound proof for the coordinator model, the hard input has a mismatch ambiguity in the order of $\epsilon/k$, which yields an $\Omega\left(\sqrt{m/\epsilon}\right)$ communication lower bound. Taken together, these results reveal a rough phase transition in the communication cost as a function of the mismatch ambiguity in the coordinator model.

### 1.1 Related Work

Several research directions in the literature are closely related to our work.

**Frequency Moments on Noiseless Data.** The $F_p$ problem has been extensively studied in the noiseless setting, both in the data stream model [1–7, 9, 19–21, 24, 31, 33, 35] and the coordinator model [13, 17, 23, 25, 29, 36]. In the data stream model, for $p > 2$, the best known space upper bound is $O\left(\frac{1}{\epsilon^2}n^{1-2/p}\log^2 n\right)$ bits [19], where $n$ denotes the universe size and the stream length is assumed to satisfy $m = n^{\Theta(1)}$. This bound is tight up to a $\log n$ factor [31]. In the coordinator model, for $p \geq 2$, there is a two-round algorithm that achieves a communication cost of $O\left(\frac{1}{\epsilon^2}k^{p-1}\log^{O(1)} n\right)$ bits [23], which is tight up to a $\log^{O(1)} n$ factor [36].

**Statistical Estimations on Noisy Data.** The literature on handling noisy datasets in the data stream model is rather sparse. The authors of [10] and [11] investigated the distinct elements problem ($F_0$) and $\ell_0$-sampling, but their results are restricted to constant-dimensional Euclidean

[3]For simplicity, we hide polylogarithmic factors in the $\tilde{O}(\cdot)$ notation.

spaces and metric spaces that support efficient locality-sensitive hashing. A recent work [38] broadens the scope to general metric spaces (or more generally, any domain equipped with a 0/1 similarity function). However, the data ambiguity model used in all these works is tailored specifically to the $F_0$ problem—more precisely, it is defined as the difference between the minimum-cardinality disjoint clique partition of the graph $G^\sigma$ and the closest cluster graph—and does not extend naturally to other statistical problems.

In the coordinator model, the only known work on statistical estimation on noisy data is [37]. However, it focuses exclusively on datasets whose graph representations form cluster graphs.

**Streaming Algorithms on Uncertain Data.** A line of work has studied streaming algorithms for uncertain/probabilistic data [12, 26–28, 40], where data noise is captured through the "possible-worlds" semantics. In this framework, each element is associated with a probability distribution over its possible occurrences, making it suited for noise sources such as misspellings. However, such method is only feasible when the universe is fixed and known in advance, which often does not hold in real-world applications. For example, if a data item is an image perturbed from an original copy that is never observed, then it is impossible to design an oracle that always returns the correct ground-truth element. As another example, two queries to a large language model may yield answers with nearly identical semantics but slightly different surface forms. In this case, the notion of a "ground-truth element" is conceptual and difficult to make explicit, and the number of possible semantics can be unbounded. Therefore, our similarity-oracle based framework is suitable for a much broader class of noisy data.

## 2 The Algorithms

In this section, we present algorithms for $F_p$ on noisy datasets in the data stream and coordinator models.

We use $\mathrm{mult}(i_1, \ldots, i_p)$ $(i_1, \ldots, i_p \in [m])$ to denote the multinomial coefficient, defined as: $\mathrm{mult}(i_1, \ldots, i_p) = \frac{p!}{\prod_{k\in[m]}(\sum_{j\in[p]} \mathbf{1}\{i_j=k\})!}$ . In words, $\mathrm{mult}(i_1, \ldots, i_p)$ is the number of permutations that can be formed by $\{\sigma_{i_1}, \ldots, \sigma_{i_p}\}$.

### 2.1 The Streaming Algorithm

We begin with a one-pass streaming algorithm for $F_p$ with an approximation guarantee parameterized by $\eta_p$.

**Ideas and Technical Overview.** We say a tuple of nodes $(i_1, \ldots, i_p)$ is a $p$-clique in $G^\sigma$ if, for all $j, k \in [p]$, $\sigma_{i_j} \sim \sigma_{i_k}$, and a $p$-clique is *ordered* if the sequence of its $p$ nodes matters. As nodes in our setting are considered similar to themselves, we allow cliques to include duplicate nodes.

We note that for the ground truth data stream $\tau$, $F_p$ is precisely $\left|\mathcal{K}_p^\tau\right|$, where $\mathcal{K}_p^\tau$ is the set of ordered $p$-cliques in $G^\tau$.[4] For a noisy data stream $\sigma$, let $\mathcal{K}_p^\sigma$ be the set of ordered $p$-cliques in $G^\sigma$. Our main observation is that for *any* ground truth $\tau$ such that $\eta_p$ is small, we have $\left|\mathcal{K}_p^\sigma\right| \approx \left|\mathcal{K}_p^\tau\right| = F_p$. Therefore, if we can accurately estimate $\left|\mathcal{K}_p^\sigma\right|$, we can also well-estimate $F_p$.

The question now becomes how to estimate $\left|\mathcal{K}_p^\sigma\right|$ in one pass. A natural idea is to draw uniform random $p$-tuples from $[m]^p$ and then count how many of them belong to $\mathcal{K}_p^\sigma$. However, this approach has a problem: $\left|\mathcal{K}_p^\sigma\right|$ may be as small as $m$, resulting in a low probability that a sample

[4]For example, for $p = 3$ and $G^\tau = (V, E)$ where $V = \{1, 2\}$ and $E = \{(1, 1), (1, 2), (2, 2)\}$, then $\mathcal{K}_p^\tau$ contains $(1, 1, 1)$, $(1, 1, 2)$, $(1, 2, 1)$, $(2, 1, 1)$, $(1, 2, 2)$, $(2, 1, 2)$, $(2, 2, 1)$ and $(2, 2, 2)$.

belongs to $\mathcal{K}_p^\sigma$. On the other hand, directly counting $\left|\mathcal{K}_p^\sigma\right|$ is challenging in the data stream model, since the nodes within a tuple are ordered, and this ordering typically does not align with the order in which the nodes appear in the stream. We thus focus on *increasingly ordered* $p$-cliques, where the order of nodes in the tuple matches their order in the data stream.

We design a sampling-and-counting procedure, denoted by $\mathcal{A}$, that leverages the observation that each increasingly ordered $j$-clique can be incrementally constructed from an increasingly ordered $(j-1)$-clique. Conceptually, the generative process can be represented as a tree, where the root corresponds to a 0-clique and its children are 1-cliques by adding one more node to the 0-clique, continuing in this manner for $p$ levels. The leaves of the tree precisely enumerate the increasingly ordered $p$-cliques. We then perform a random walk from the root to a leaf to sample a path, and produces an estimate $X$ based on the probability of selecting this path and an appropriate multinomial coefficient. We can show that $X$ satisfies $\mathbf{E}[X] = |\mathcal{K}_p^\sigma|$.

The critical step in the above approach is to determine the number of unbiased estimates needed for the final approximation. Let $X_1, \ldots, X_t$ be estimates returned by $t$ independent runs of $\mathcal{A}$, where for every $j \in [t]$, $\mathbf{E}[X_j] = \left|\mathcal{K}_p^\sigma\right|$. Let $\bar{X} = \frac{1}{t}\sum_{j\in[t]} X_j$. If $\sigma$ is noiseless, it is not difficult to show that $t = O\left(\frac{1}{\epsilon^2} m^{1-1/p}\right)$ estimates are sufficient to ensure that $\bar{X}$ tightly concentrates around its expectation. However, in the noisy setting, $\bar{X}$ may exhibit high instability in worst case scenarios. For example, consider the case when $p = 2$ and $G^\sigma$ is a star graph. In this case, $\mathcal{K}_2^\sigma$ consists of all ordered pairs of similar nodes, with approximately two-thirds of the pairs containing the central node. If the central node arrives first in the stream, then with probability $\left(1 - \frac{1}{m}\right)$ it is not sampled. As a result, roughly two-thirds of the pairs in $\mathcal{K}_2^\sigma$ are not counted, which prevent us to obtain a 1.01-approximation with probability 0.99 when $t = o(m)$.

What we manage to show is that when $\eta_p$ is small, we can estimate $\left|\mathcal{K}_p^\sigma\right|$ accurately using a sublinear number of samples. First, it is not difficult to establish an upper bound of $\bar{X}$ by a Markov inequality. In fact, as $\bar{X}$ is the sum of independent random variables, we can apply an inequality from [18] to achieve a better bound. The challenge lies in establishing the lower bound of $\bar{X}$. The standard method is to bound the variance of $\bar{X}$ (or, the variance of each $X_j$ ($j \in [t]$)) and then apply Chebyshev's inequality. However, if we analyze the variance of $X_j$ directly without taking into account the mismatch ambiguity $\eta_p$, the variance could be very high. Consider again the star graph, which has a high $\eta_p$ with respect to any ground truth. Easy calculation gives $\mathbf{E}[X_j] = \Theta(m)$ and $\mathbf{Var}[X_j] = \Omega(m^p)$. Consequently, we have to use $t = \Omega\left(m^{p-2}\right)$ copies of $X_j$ to bring down the variance of $\bar{X}$ to $O(1)$. Therefore, we must account for $\eta_p$ in our analysis. The difficulty is that the construction of $\bar{X}$ cannot incorporate $\eta_p$, as it is an unknown quantity to the algorithm.

We circumvent the above issue by relating each random variable $X_j$ ($j \in [t]$) to another random variable $Y_j \triangleq \min\left\{X_j, m(p!)\left(d_{I_1}^\tau\right)^{p-1}\right\}$, where $I_1$ is distributed uniformly at random in $\{1, \ldots, m\}$. The intuition of creating such a random variable $Y_j$ is the following: recall that if we apply $\mathcal{A}$ to the ground truth $\tau$, whose graph representation $G^\tau$ is a cluster graph, we can obtain a tighter upper bound on the variance of $X_j$, and consequently only need $t = o(m)$ runs of $\mathcal{A}$. Now, if $\eta_p$ is small, we would expect that $\mathcal{K}_p^\sigma$ and $\mathcal{K}_p^\tau$ overlap significantly, and consequently, $\left|\mathcal{K}_p^\sigma\right|$ and $\left|\mathcal{K}_p^\sigma \cap \mathcal{K}_p^\tau\right|$ are close. We can intuitively interpret $Y_j$ with $\mathbf{E}[Y_j] \approx \left|\mathcal{K}_p^\sigma \cap \mathcal{K}_p^\tau\right|$ as being close to $X_j$ with $\mathbf{E}[X_j] = \left|\mathcal{K}_p^\sigma\right|$ but with a smaller variance. Therefore, we can first establish a lower bound for $Y_j$, which also provides a lower bound for $X_j$.

**Algorithm 1:** Streaming-Robust-$F_p$-One-Sample

**Input:** a noisy data stream $\sigma$
**Output:** an estimate of $F_p$

1 $r_1, \ldots, r_p, i_1, \ldots, i_p \leftarrow 0$, $a_1, \ldots, a_p \leftarrow$ null
2 **foreach** *incoming* $\sigma_j$ **do**
3 $\quad r_1 \leftarrow r_1 + 1$
4 $\quad$ with probability $\frac{1}{r_1}$, $(a_1, i_1, r_2) \leftarrow (\sigma_j, j, 0)$
5 $\quad$ **if** $a_1 \sim \sigma_j$ **then**
6 $\quad\quad r_2 \leftarrow r_2 + 1$
7 $\quad\quad$ with probability $\frac{1}{r_2}$, $(a_2, i_2, r_3) \leftarrow (\sigma_j, j, 0)$
8 $\quad\quad$ **if** $a_2 \sim \sigma_j$ **then**
9 $\quad\quad\quad r_3 \leftarrow r_3 + 1$
10 $\quad\quad\quad$ with probability $\frac{1}{r_3}$, $(a_3, i_3, r_4) \leftarrow (\sigma_j, j, 0)$
11 $\quad\quad\quad \ldots$
12 $\quad\quad\quad$ **if** $a_{p-1} \sim \sigma_j$ **then**
13 $\quad\quad\quad\quad r_p \leftarrow r_p + 1$
14 $\quad\quad\quad\quad$ with probability $\frac{1}{r_p}$, $(a_p, i_p) \leftarrow (\sigma_j, j)$
15 **return** $\text{mult}(i_1, \ldots, i_p) \prod_{k \in [p]} r_k$.

**Algorithm 2:** Streaming-Robust-$F_p$

**Input:** a noisy data stream $\sigma$ of size $m$, parameter $\epsilon$
**Output:** an $((\epsilon + O(\eta_p)), 0.01)$-approximation of $F_p$

1 $\ell \leftarrow 100$, $t \leftarrow \frac{10^6 (p!) m^{1-1/p}}{\epsilon^2}$
2 obtain $\ell t$ estimates $\{X_{i,j}\}_{i \in [\ell], j \in [t]}$ by running $\ell t$ instances of Algorithm 1 in parallel
3 for $i \in [\ell]$, let $\bar{X}_i \leftarrow \frac{1}{t} \sum_{j \in [t]} X_{i,j}$
4 **return** $\min\{\bar{X}_1, \ldots, \bar{X}_\ell\}$.

**Algorithm and Analysis.** The algorithm is described in Algorithm 2, which uses Algorithm 1 as a subroutine. At a high level, Algorithm 1 aims to obtain an unbiased estimator of the number of ordered $p$-cliques in $G^\sigma$, which serves as a low-bias estimate of $F_p$. This is achieved by sampling and counting increasingly ordered $p$-cliques in its main loop (Lines 2–14), followed by scaling the result with appropriate multinomial coefficients (Line 15). By executing Algorithm 1 multiple times in parallel, Algorithm 2 applies a min-average aggregation (Lines 3–4) to produce the final approximation for $F_p$.

We have the following result regarding Algorithm 2.

Theorem 2.1. *For any constant $p \in \mathbb{Z}^+$, given a noisy input data stream of length $m$ with $\eta_p \le \frac{1}{3(p!)}$, Algorithm 2 computes an $((\epsilon + O(\eta_p), 0.01)$-approximation of $F_p$, using a single pass and $O\left(\frac{1}{\epsilon^2} m^{1-1/p}\right)$ words of space.*

In the rest of this section, we prove Theorem 2.1. We will make use of the following inequality.

Lemma 2.2 ([18]). *Let $Z_1, \ldots, Z_t$ be arbitrary non-negative independent random variables, with expectations $\mu_1, \ldots, \mu_t$ where $\mu_i \le 1$ for all $i$. Then for any $\delta > 0$,*

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

We note that when $p = 1$, the output of Algorithm 2 is always $m$, which is exactly $F_1$. We thus focus on integers $p \ge 2$.

We start by introducing the following concept.

Definition 2.3 (Ordered $p$-Clique). *For a data stream $\sigma$ of length $m$, define the set of ordered $p$-cliques in $G^\sigma$ as $\mathcal{K}_p^\sigma = \left\{(i_1, \ldots, i_p) \in [m]^p \mid \forall (j,k) \in [p]^2, \sigma_{i_j} \sim \sigma_{i_k}\right\}$.*

Fact 2.4. *For a noiseless data stream $\tau$, $\left|\mathcal{K}_p^\tau\right| = F_p$.*

Proof. Let $n_\tau$ be the number of distinct elements in $\tau$, and $\{V_1, \ldots, V_{n_\tau}\}$ be the set of cliques in $G^\tau$. It is easy to see that for any $K_p \in \mathcal{K}_p^\tau$, all $p$ nodes in $K_p$ must belong to the same clique $V_k$ for some $k \in [n_\tau]$. Therefore, the total number of ordered $p$-cliques is $\sum_{k\in[n_\tau]} |V_k|^p = F_p$. □

The following lemma shows that for a noisy data stream $\sigma$ and the corresponding ground truth $\tau$, if $\eta_p$ is small, then $\mathcal{K}_p^\sigma$ and $\mathcal{K}_p^\tau$ overlap significantly.

Lemma 2.5. *For graphs $G^\sigma = ([m], E^\sigma)$ and $G^\tau = ([m], E^\tau)$, representing a noisy data stream $\sigma$ and the corresponding ground truth $\tau$ respectively, we have*

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

Proof. We first bound the size of $\mathcal{K}_p^\sigma \setminus \mathcal{K}_p^\tau$. Consider an ordered $p$-clique $K_p = (i_1, \ldots, i_p)$ in $\mathcal{K}_p^\sigma \setminus \mathcal{K}_p^\tau$ with $i_1 = i$. There must exist a $k \in [p]$ such that $i_k \in B_i^\sigma \setminus B_i^\tau$. Therefore, the total number of ordered $p$-cliques in $\mathcal{K}_p^\sigma \setminus \mathcal{K}_p^\tau$ with $i_1 = i$ is at most $\left|B_i^\sigma\right|^{p-1} - \left|B_i^\sigma \cap B_i^\tau\right|^{p-1} \le \left|B_i^\sigma \cup B_i^\tau\right|^{p-1} - \left|B_i^\sigma \cap B_i^\tau\right|^{p-1}$. Summing over $i \in [m]$, we have

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

We next bound the size of $\mathcal{K}_p^\tau \setminus \mathcal{K}_p^\sigma$. Let $n_i$ be the number of ordered $p$-cliques in $\mathcal{K}_p^\tau$ that contains $i$ and at least one node from $B_i^\tau \setminus B_i^\sigma$. We have

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

Note that for each $K_p = (i_1, \ldots, i_p) \in \mathcal{K}_p^\tau \setminus \mathcal{K}_p^\sigma$, there exists $x, y \in [p]$ such that $i_x \ne i_y$ and $(i_x, i_y) \notin E^\sigma$, which implies that $K_p$ is counted in both $n_{i_x}$ and $n_{i_y}$. Summing over $i \in [m]$, we have

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

□

Fact 2.4 and Lemma 2.5 imply that we can estimate $F_p$ via estimating $\left|\mathcal{K}_p^\sigma\right|$. However, it seems difficult to estimate this quantity directly in the data stream model under sublinear memory space. We therefore turn our attention to a closely related set.

Definition 2.6 (Increasingly Ordered $p$-Clique). *For a data stream $\sigma$, define the set of increasingly ordered $p$-cliques in $G^\sigma$ as $\vec{\mathcal{K}}_p^\sigma = \{(i_1, \ldots, i_p) \in \mathcal{K}_p^\sigma \mid i_1 \le \ldots \le i_p\}$.*

Note that for each increasingly ordered $p$-clique $\vec{K}_p = (i_1, \ldots, i_p) \in \vec{\mathcal{K}}_p^\sigma$, there are exactly $\text{mult}(i_1, \ldots, i_p)$ ordered $p$-cliques in $\mathcal{K}_p^\sigma$ that consist of the same multiset of nodes.

To facilitate the subsequent analysis, we introduce the following concept.

Definition 2.7 (Tail Set). *For a data stream $\sigma$ and $j \in [p]$, define the tail set of an increasingly ordered $j$-clique $(i_1, \ldots, i_j)$ in $G^\sigma$ as $\text{tail}(i_1, \ldots, i_j) = \{i \in [m] \mid i \geq i_j \wedge (\forall k \in [j], \sigma_i \sim \sigma_{i_k})\}$.*

In words, $\text{tail}(i_1, \ldots, i_j)$ contains all nodes $i \in [m]$ appearing after $i_j$ (including $i_j$) in the stream that, together with the $j$-clique $(i_1, \ldots, i_j)$, form an increasingly ordered $(j+1)$-clique.

The following lemma shows that Algorithm 1 outputs an unbiased estimator of $\left|\mathcal{K}_p^\sigma\right|$. Let $I_1, \ldots, I_p$ correspond to $i_1, \ldots, i_p$ in Algorithm 1 be the random nodes, and let $R_1, \ldots, R_p$ correspond to $r_1, \ldots, r_p$ in Algorithm 1 be the tail counts. Note that $R_1 = m$ is actually a fixed value, whereas others are true random variables.

Lemma 2.8. *Let $X = \text{mult}(I_1, \ldots, I_p) \prod_{k \in [p]} R_k$ be the output of Algorithm 1. It holds that $\mathbf{E}[X] = \left|\mathcal{K}_p^\sigma\right|$.*

Proof. It is easy to see that in Algorithm 1, $I_1$ is uniformly sampled from $[m]$ and $I_k$ is uniformly sampled from $\text{tail}(I_1, \ldots, I_{k-1})$ for $k = 2, \ldots, p$. Let $\vec{K}_p = (i_1, \ldots, i_p) \in \vec{\mathcal{K}}_p^\sigma$ be an increasingly ordered $p$-clique. We have

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

Note that $R_1 = m$, and $R_k = |\text{tail}(I_1, \ldots, I_{k-1})|$ for $k = 2, \ldots, p$, and $(I_1, \ldots, I_p)$ is always an increasingly ordered $p$-clique. It follows that

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

□

Let $X_1, \ldots, X_t$ denote $t$ independent outputs of Algorithm 1, and let $\bar{X} = \frac{1}{t} \sum_{i \in [t]} X_i$. The next two lemmas show that when the mismatch ambiguity $\eta_p$ is small, $\bar{X}$ is tightly concentrated around $F_p$ with good probability.

Lemma 2.9 ($\bar{X}$ is not too large). $\mathbf{Pr}\left[\bar{X} < (1 + \epsilon + \eta_p) \cdot F_p\right] > \frac{1}{13}$.

Proof. Lemma 2.8 gives $\mathbf{E}[\bar{X}] = \left|\mathcal{K}_p^\sigma\right|$. Applying Lemma 2.2 (setting $Z_i = X_i / |\mathcal{K}_p^\sigma|$ and $\delta = 1$), we have

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

By Fact 2.4 and Lemma 2.5,

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

The lemma follows from (3), (4), and the fact that $\frac{1}{t} \ll \frac{\epsilon}{2}$. □

Lemma 2.10 ($\bar{X}$ is not too small). $\mathbf{Pr}\left[\bar{X} \ge \left(1-\epsilon-2(p!)\eta_p\right)\cdot F_p\right] > 1 - 2\cdot 10^{-5}$.

Proof. For every $j \in [t]$, let $Y_j = \min\left\{X_j, m(p!)\left|B_{I_1}^\tau\right|^{p-1}\right\}$. We have the following claims, whose proofs will be given shortly. The first claim states that the expectation of each $Y_j$ is not far away from $F_p$ given that $\eta_p$ is small, and the second bounds the variance of each $Y_j$.

Claim 2.11. *For any $j \in [t]$, $(1-2(p!)\eta_p)F_p \le \mathbf{E}[Y_j] \le (1+\eta_p)F_p$.*

Claim 2.12. *For any $j \in [t]$, $\frac{\mathbf{Var}[Y_j]}{(\mathbf{E}[Y_j])^2} \le \frac{p!}{1-2(p!)\eta_p}\cdot m^{1-1/p}$.*

We now have

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

where from (5) to (6), we use the fact $X_j \ge Y_j$ and the first inequality of Claim 2.11. From (7) to (8), we apply the second inequality of Claim 2.11. From (8) to (9), we apply Claim 2.12. From (9) to (10), we use our assumption $\eta_p < \frac{1}{3(p!)}$. The last inequality is due to our choice of $t$ (Line 1 of Algorithm 2). □

Now we are ready to prove Theorem 2.1. Recall that we have set $\ell = 100$ in Algorithm 2. Let $X_{\min} = \min\left\{\bar{X}_1, \ldots, \bar{X}_\ell\right\}$ be the output of Algorithm 2. By Lemma 2.9 and 2.10, we have

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

The theorem follows from a union bound.

Finally, we give the proofs for the two claims used in Lemma 2.10.

Proof of Claim 2.11. We prove for every $j \in [t]$, and thus drop the subscript $j$ in $X_j$ and $Y_j$ and write them as $X$ and $Y$ for convenience. We first look at the difference between $X$ and $Y$,

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

Conditioned on $I_1 = i$, we have

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

where from (11) to (12), we use the fact that $\mathrm{tail}(i, I_2, \ldots, I_{p-1}) \subseteq \mathrm{tail}(i, I_2, \ldots, I_{p-2}) \subseteq \cdots \subseteq \mathrm{tail}(i)$. The step from (12) to (13) follows from $\mathrm{tail}(i) \subseteq B_i^\sigma$.

We thus have

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

Noting that $\mathbf{Pr}[I_1 = i] = \frac{1}{m}$, we have

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

Consequently,

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

On the other hand, by (4) and Lemma 2.8, we have $\mathbf{E}[Y] \le \mathbf{E}[X] = \left|\mathcal{K}_p^\sigma\right| \le (1 + \eta_p) F_p$. The claim follows. □

Proof of Claim 2.12. Let $n_\tau$ be the number of distinct elements in the ground truth $\tau$, and $\{V_1, \ldots, V_{n_\tau}\}$ be the set of cliques in $G^\tau$. We prove for every $j \in [t]$, and thus drop the subscript $j$ in $Y_j$ for convenience.

By the definition of $Y$, we have $Y \le m(p!) \cdot \max_{i \in [m]} \left|B_i^\tau\right|^{p-1} = m(p!) \cdot \max_{k \in [n_\tau]} |V_k|^{p-1}$, which implies

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

Note that

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

And by Hölder's inequality,

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

Combining (15), (16), (17) and the first inequality of Claim 2.11, we have

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

□

## 2.2 The Distributed Algorithm

We note that our streaming algorithm cannot be directly translated into a round-efficient distributed algorithm, as sampling from the ordered $p$-cliques needs at least $p$ rounds. The key to designing a distributed algorithm that is efficient in both communication and rounds is the fact that, for the ground-truth dataset $\tau$, the quantity $F_p(\tau)$ coincides with the $(p-1)$-th *degree moment* of $G^\tau$, defined as $M_{p-1}^\tau \triangleq \sum_{i\in[m]} \left|B_i^\tau\right|^{p-1}$. When $\eta_p(\sigma,\tau)$ is small, the $(p-1)$-th degree moment $M_{p-1}^\sigma \triangleq \sum_{i\in[m]} \left|B_i^\sigma\right|^{p-1}$ of $G^\sigma$ is expected to be close to $M_{p-1}^\tau$. Consequently, an accurate estimate of $M_{p-1}^\sigma$ yields a good approximation of $F_p(\tau)$.

In the coordinator model, obtaining an unbiased estimate of $M_{p-1}^\sigma$ is relatively straightforward: we can uniformly sample nodes and query each site for their degrees. However, this vanilla sampling strategy suffers from instability, similar to the streaming setting. Fortunately, the analysis techniques developed for the data stream model can be applied in this context as well.

Our two-round algorithm for approximating $F_p$ in the coordinator model is described in Algorithm 3. In the first round, the sites and coordinator sample a random set of nodes. At Line 11, the coordinator tries to convert the sample from being "without replacement" to "with replacement". In the second round, the players compute the $(p-1)$-th degree moment for each sampled node. Finally, the coordinator use the min-average aggregation to output the final approximation.

We have the following guarantees regarding Algorithm 3.

Theorem 2.13. *For any constant $p \ge 1$, given an input dataset of size $m$ with $\eta_p \le 0.4$ partitioned among $k$ sites in the coordinator model, Algorithm 3 computes a $((\epsilon + O(\eta_p)), 0.01)$-approximation of $F_p$, using two rounds and $O\left(\frac{1}{\epsilon^2} k m^{1-1/p}\right)$ words of communication.*

While Algorithm 3 differs considerably from Algorithm 2, their analyses share similarity, especially regarding the handling of random variable instability.

Due to the space constraints, we defer the proof of Theorem 2.13 to the full version of this paper [32].

## 3 The Lower Bounds

We now explore the lower bounds, beginning with an overview of key ideas and proof techniques.

**Ideas and Technical Overview.** We utilize the classical multiparty communication problem $\mathrm{DISJ}_{n,k}$ for proving the lower bounds. In $\mathrm{DISJ}_{n,k}$, we have $k$ players. Each player $\kappa$ ($\kappa \in [k]$) is given a vector $X^{(\kappa)} = \left(X_1^{(\kappa)}, \ldots, X_n^{(\kappa)}\right) \in \{0,1\}^n$, representing a subset of $[n]$. We use the vector and subset representation interchangeably. The input $\{X^{(1)}, \ldots, X^{(k)}\}$ of $\mathrm{DISJ}_{n,k}$ satisfies one of the following two cases:

**Algorithm 3:** Distributed-Robust-$F_p$

**Input:** a noisy dataset $\sigma$ of size $m$ partitioned among $k$ sites, where the $\kappa$-th site holds $\sigma^{(\kappa)}$; a parameter $\epsilon$

**Output:** an $((\epsilon + O(\eta_p)), 0.01)$-approximation of $F_p$

$t \leftarrow \frac{10^6 m^{1-1/p}}{\epsilon^2}$, $\ell \leftarrow 100$

// First Round:

Coordinator: send $m$ to all sites.

**for** *site* $\kappa \leftarrow 1, \ldots, k$ **do**

  $S^{(\kappa)} \leftarrow \emptyset$ /* set of sampled nodes at the $\kappa$-th site */

  **foreach** $q \in V^{(\kappa)}$ **do**

    with probability $\frac{2t\ell}{m}$, $S^{(\kappa)} \leftarrow S^{(\kappa)} \cup \{q\}$

  send $S^{(\kappa)}$ to the coordinator.

// Second Round:

/* convert "sampled without replacement" to "sampled with replacement" */

Coordinator: $S' \leftarrow \bigcup_{\kappa \in [k]} S^{(\kappa)}$. If $|S'| < t\ell$, the coordinator outputs fail. Otherwise, it computes a sample $S$ from $S'$ such that (1) $|S| = t\ell$, and (2) each element in $S$ is uniformly sampled from $[m]$ with replacement. This can be done using a method described in [14, Section 3.2]. Send $S$ to all sites.

**for** *site* $\kappa \leftarrow 1, \ldots, k$ **do**

  **foreach** $i \in S$ **do**

    $d_i^{(\kappa)} \leftarrow \left|V^{(\kappa)} \cap B_i^{\sigma}\right|$ /* local degree of node $i$ at the $\kappa$-th site */

  send $\{d_i^{(\kappa)}\}_{i \in S}$ to the coordinator.

// Output:

Coordinator: calculate $d_i \leftarrow \sum_{\kappa \in [k]} d_i^{(\kappa)}$ for each $i \in S$. Divide $S$ into $\ell$ groups $\{S_1, \ldots, S_\ell\}$, each containing $t$ samples. For each $j \in [\ell]$, calculate $\bar{X}_j \leftarrow \frac{m}{t} \sum_{i \in S_j} (d_i)^{p-1}$.

**return** $\min\{\bar{X}_1, \ldots, \bar{X}_\ell\}$.

(1) *NO instance*: all $k$ subsets are pairwise disjoint. That is, for any $\kappa \neq \kappa'$, $X^{(\kappa)} \cap X^{(\kappa')} = \emptyset$.
(2) *YES instance*: they have a unique common element but are otherwise disjoint. That is, $\left|\bigcap_{\kappa \in [k]} X^{(\kappa)}\right| = 1$ and for any $\kappa \neq \kappa'$, $X^{(\kappa)} \cap X^{(\kappa')} = \cap_{\kappa \in [k]} X^{(\kappa)}$.

The problem is for the $k$ players to find out via communication whether the input is a YES instance or a NO instance. $\mathrm{DISJ}_{n,k}$ is originally studied in the blackboard model, where whenever a player sends a message, all other $(k-1)$ players can see it. It is easy to see that any lower bound proved in the blackboard model also holds in the coordinator model, where the $k$ sites serve as the $k$ players, who communicate with each other through the coordinator. The following result is known for $\mathrm{DISJ}_{n,k}$. We say an algorithm $\delta$-error if the algorithm errors with probability at most $\delta$.

Theorem 3.1 ([9]). *Any 0.49-error algorithm for $\mathrm{DISJ}_{n,k}$ needs $\Omega\left(\frac{n}{k \log k}\right)$ bits of communication in the blackboard model.*

The $\mathrm{DISJ}_{n,k}$ problem described above was used to prove an $\tilde{\Omega}\left(m^{1-2/p}\right)$ space lower bound for the $F_p$ problem in the noiseless data stream model [9]. The standard reduction works as follows: given an $(\epsilon, \delta)$-approximation streaming algorithm $\mathcal{A}$ for $F_p$, we can solve $\mathrm{DISJ}_{n,k}$ with parameters $n = \Theta(m)$ and $k = ((1+3\epsilon)n)^{1/p}$ in the following way: Each player $\kappa \in [k]$ constructs a set of items $\sigma^{(\kappa)}$ by including every $i \in X^{(\kappa)}$. The final stream $\sigma$ is the concatenation of all $\sigma^{(\kappa)}$, that

is, $\sigma = \sigma^{(1)} \circ \ldots \circ \sigma^{(k)}$. The $k$ players can simulate the streaming algorithm $\mathcal{A}$ to solve $\mathrm{DISJ}_{n,k}$ as follows: the first player runs $\mathcal{A}$ on $\sigma^{(1)}$ and write the end memory configuration $M_1$ to the blackboard. The second player continue runs $\mathcal{A}$ on $\sigma^{(2)}$ starting from $M_1$, and then write the end memory configuration $M_2$ to the blackboard, and so on. At the end, the last player outputs YES if the streaming algorithm outputs an $F_p$-estimation larger than $(1+\epsilon)n$, and NO otherwise. If multiple passes are allowed, then the last player can write its end memory configuration to the blackboard and restart the simulation from the first player. If the streaming algorithm $\mathcal{A}$ uses $O(1)$ passes and $s$ bits of space, then the $k$ players solve $\mathrm{DISJ}_{n,k}$ problem using at most $O(sk)$ bits of communication. By Theorem 3.1, we have $sk = \Omega\left(\frac{n}{k \log k}\right)$, and consequently, $s = \Omega\left(\frac{n}{k^2 \log k}\right) = \Omega\left(\frac{m^{1-2/p}}{\log m}\right)$.

However, this proof strategy is *not* generally applicable in the coordinator model, as the reduction requires $k = \Omega\left(m^{1/p}\right)$. In fact, there exist $(\epsilon, 0.01)$-approximation algorithms for $F_p$ in the coordinator model with $k$ players using $\tilde{O}\left(\frac{k^{p-1}}{\epsilon^2}\right)$ bits of communication [17, 23]. Our key observation is that in the noisy data setting, a similar reduction can be performed using only a *constant* number of players. Moreover, the proof only requires that the dataset has small $F_p$-mismatch-ambiguity, $\eta_p = O(\epsilon)$, under which an $(\epsilon + O(\eta_p), \delta)$-approximation is equivalent to an $(O(\epsilon), \delta)$-approximation.

More precisely, each player $\kappa \in [k]$, for each element $i \in [n]$, adds $t/k$ pairwise *dissimilar* items (for an appropriate parameter $t$ to be specified later); let $Q_i^{(\kappa)}$ denote these items. We further require that if two players $\kappa$ and $\kappa'$ share an element $i$ in the $\mathrm{DISJ}_{n,k}$ instance, then for any pair of items $u \in Q_i^{(\kappa)}$ and $v \in Q_i^{(\kappa')}$, it holds that $u \sim v$. Conversely, if they do not share any element, then all items they create are pairwise dissimilar. It is important to notice that this requirement cannot be satisfied in noiseless settings due to the transitivity constraint: If all items in $Q_i^{(\kappa)}$ were similar to those in $Q_i^{(\kappa')}$, then transitivity would imply that all items within $Q_i^{(\kappa)}$ to be mutually similar. In contrast, we will show that in the noisy data setting, it is possible to construct an explicit dataset and an appropriate similarity function that jointly satisfy both requirements with probability 0.99.

We present two concrete constructions. The first is more natural—its similarity oracle is induced by a natural distance function. More precisely, let $s = \Theta\left(\frac{t^2 \log t}{k^2}\right)$ be a parameter. For each player $\kappa \in [k]$ and each $i \in [n]$, we consider two cases (in the $\mathrm{DISJ}_{n,k}$ instance):

(1) If $X_i^{(\kappa)} = 1$, the player randomly partition the set $\{(i, k+1, 1), \ldots, (i, k+1, s)\}$ into $t/k$ subsets of equal size, and let item $q_{i,j}^{(\kappa)}$ be the $j$-th partition. The randomness used by different players are independent.
(2) If $X_i^{(\kappa)} = 0$, the player arbitrarily partition the set $\{(i, \kappa, 1), \ldots, (i, \kappa, s)\}$ into $t/k$ subsets of equal size, and let item $q_{i,j}^{(\kappa)}$ be the $j$-th partition.

We then let $Q_i^{(\kappa)} = \left\{q_{i,1}^{(\kappa)}, \ldots, q_{i,\frac{t}{k}}^{(\kappa)}\right\}$.

This construction satisfies all the properties that we need. The only issue is that each item requires $O\left(m^{1/p} \log m\right)$ bits to describe, which renders a non-trivial word versus bit gap between the upper and lower bounds. We therefore propose a second, somewhat artificial construction that still satisfies all the required properties, while allowing each item to be described using only a logarithmic number of bits.

Now, let us look at the YES instance of $\mathrm{DISJ}_{n,k}$. Let $i^*$ be the element shared among all players. It is easy to see that the set of items in $\sigma_Y$ created by the $k$ players from $i^*$ form a complete $k$-partitie graph, which is not far away from a clique. According to our definition of $F_p$-mismatch-ambiguity $\eta_p$, there is a noiseless dataset $\tau_Y$ such that, when used as the ground truth, the noisy dataset we construct for $F_p$ from the YES instance of $\mathrm{DISJ}_{n,k}$ has $\eta_p(\sigma_Y, \tau_Y) \le \frac{10\epsilon p}{k}$, which is at most $\frac{\epsilon}{C+1}$ by choosing $k = 10p(C+1)$ (a constant!).

On the other hand, for a NO instance of $\text{DISJ}_{n,k}$, the graph representation of the resulting dataset $\sigma_N$ for $F_p$ consists solely of singleton nodes, corresponding to a noiseless dataset $\tau_N$ with $\eta_p(\sigma_N, \tau_N) = 0$. By choosing the parameter $t = (10\epsilon m)^{1/p}$ and $n = \frac{m}{t}$, we can show that $(1 - \epsilon - C\eta_p(\sigma_Y, \tau_Y))F_p(\tau_Y) > (1 + \epsilon + C\eta_p(\sigma_N, \tau_N))F_p(\tau_N)$. Therefore, any $((\epsilon + C\eta_p), \delta)$-approximation streaming algorithm for $F_p$ that can be used to solve the $\text{DISJ}_{n,k}$ problem with error at most $(\delta+0.01)$, where the term 0.01 counts the error probability of the input reduction. Consequently, Theorem 3.1 implies an $\Omega\left(\frac{1}{\epsilon^{1/p}} m^{1-1/p}\right)$ space lower bound for $((\epsilon + C\eta), 0.48)$-approximating $F_p$ in the noisy data stream.

Using the same approach, we can prove a communication lower bound of $\Omega\left(\frac{1}{\epsilon^{1/p}} m^{1-1/p}\right)$ bits in the coordinator model.

**The Lower Bounds.** We prove the following multiparty communication lower bound for $F_p$-estimation on noisy datasets.

Theorem 3.2. *For any constants $p > 1$, $C \geq 0$ and $\epsilon \in (0, \frac{1}{3})$, any $((\epsilon + C\eta_p), 0.48)$-approximation algorithm for computing $F_p$ on any noisy dataset of up to $m$ items distributed over $k \geq 10p(C + 1)$ players needs $\Omega\left(\frac{1}{\epsilon^{1/p}} m^{1-1/p}\right)$ bits of communication in the blackboard model.*

This theorem yields two immediate corollaries, one for the data stream model and one for the coordinator model. In the standard connection between the data stream model and $k$-party communication (see the technical overview), we set the number of players $k = 10p(C + 1) = \Theta(1)$.

Corollary 3.3. *For any constants $p > 1$, $C \geq 0$ and $\epsilon \in (0, \frac{1}{3})$, any $O(1)$-pass $((\epsilon + C\eta_p), 0.48)$-approximation algorithm for computing $F_p$ on any noisy data stream of length up to $m$ needs $\Omega\left(\frac{1}{\epsilon^{1/p}} m^{1-1/p}\right)$ bits of space.*

Corollary 3.4. *For any constants $p > 1$, $C \geq 0$ and $\epsilon \in (0, \frac{1}{3})$, any $((\epsilon+C\eta_p), 0.48)$-approximation algorithm for computing $F_p$ in the coordinator model on any noisy dataset of up to $m$ items distributed over $k \geq 10p(C + 1)$ sites needs $\Omega\left(\frac{1}{\epsilon^{1/p}} m^{1-1/p}\right)$ bits of communication, regardless the number of rounds.*

Due to the space constraints, we defer the proof of Theorem 3.2 to the full version of this paper [32].

## 4 An Improved Distributed Algorithm for Low-Ambiguity Regime

As noted in the introduction, the communication cost of Algorithm 3 presented in Section 3 is asymptotically optimal in terms of $m$. However, we find that when the mismatch ambiguity $\eta_p$ falls below a certain threshold, it is possible to design algorithms with communication cost *independent* of $m$. In this section, we present a distributed algorithm in the coordinator model for datasets with small $F_p$-mismatch ambiguity, achieving a communication cost of poly $\left(k, \frac{1}{\epsilon}\right)$ words.

**Ideas and Technical Overview.** For ease of reference, we summarize in Table 1 a set of notations that will be used throughout the remainder of this section.

Same as that in the general case (Section 2.2), our goal is to estimate the $(p-1)$-th degree moment $M^{\sigma}_{p-1}$, which is close to $F_p$ when $\eta_p$ is small. Intuitively, we can view $\eta_p$ as a budget available to an adversary, who perturbs the ground truth graph $G^{\tau}$ by adding or removing edges. Consider, for example, a graph consisting of $(m - m^{1/p})$ isolated nodes and a clique of size $m^{1/p}$, whose nodes are evenly distributed across $k$ sites. When $\eta_p = \Theta(\epsilon)$, the adversary can use their budget to convert the clique into a $k$-partite graph by deleting all edges with both end nodes located at the same

| Notation | Description |
|---|---|
| $V^{(\kappa)}$ | subset of $[m]$ held by site $\kappa \in [k]$. |
| $d_i^{(\kappa)} \triangleq \left\lvert V^{(\kappa)} \cap B_i^{\sigma} \right\rvert$ | number of nodes similar to $\sigma_i$ held by site $\kappa$. |
| $d_i \triangleq \sum_{\kappa \in [k]} d_i^{(\kappa)} = \left\lvert B_i^{\sigma} \right\rvert$ | global degree of $\sigma_i$. |
| $d_{i,\text{local}} \triangleq d_i^{(\kappa(i))}$ | local degree of $\sigma_i$, where $\kappa(i) \in [k]$ is the site that holds $\sigma_i$. |
| $D_{\text{local}}^{(\kappa)} \triangleq \sum_{i \in V^{(\kappa)}} (d_{i,\text{local}})^{p-1}$ | sum of local degree moments of $V^{(\kappa)}$. |
| $D_{\text{local}} \triangleq \sum_{\kappa \in [k]} D_{\text{local}}^{(\kappa)} = \sum_{i \in [m]} (d_{i,\text{local}})^{p-1}$ | sum of local degree moments of $[m]$. |
| $d_{i,\text{local}}^{\tau} \triangleq \left\lvert B_i^{\tau} \cap V^{(\kappa(i))} \right\rvert$ | local degree of $\tau_i$. |
| $d_i^{\tau} \triangleq \left\lvert B_i^{\tau} \right\rvert$ | global degree of $\tau_i$. |

Table 1. Summary of notations.

site. As a result, each site cannot locally determine whether a node is part of a clique or merely an isolated node—yet this distinction is critical due to the substantial contribution of clique nodes to the quantity $M_{p-1}^{\sigma}$.

Our main idea for datasets with small $\eta_p$ is that when the adversary does not have sufficient budget to eliminate all local structural information, then we can exploit the remaining local structures to identify high-degree nodes with a small amount of communication. Note that the contribution of high-degree nodes dominate the value of $M_{p-1}^{\sigma}$.

We first consider the noiseless case to see how the local information helps. Let $\sigma$ be a noiseless dataset, and let $V$ denote the largest clique of $G^{\sigma}$. For each $\kappa \in [k]$, let $V^{(\kappa)}$ be the nodes in $V$ that is held by site $\kappa$. Let $\alpha = \frac{k}{\epsilon}$ be a threshold. If $\frac{|V|}{\left|V^{(\kappa)}\right|} > \alpha$, then $V^{(\kappa)}$ is small and the site $\kappa$ may not be able to identify $V^{(\kappa)}$ as part of a large clique $V$ without communication. But this is not a big issue, since the contribution of $V^{(\kappa)}$ is at most $\frac{1}{\alpha}$-fraction of that of $V$ and can thus be ignored. On the other hand, if $\frac{|V|}{\left|V^{(\kappa)}\right|} \le \alpha$, then $\left|V^{(\kappa)}\right|$ is large given $|V|$ is large. In this case, site $\kappa$ can identify $V^{(\kappa)}$ as part of a large clique without any communication. Therefore, to estimate $M_{p-1}^{\sigma}$, it suffices to estimate the contribution of large $V^{(\kappa)}$s.

To implement this idea, we sample nodes according to the probabilities $\left\{\frac{(d_{i,\text{local}})^{p-1}}{D_{\text{local}}}\right\}_{i \in [m]}$ (i.e., proportional to the $(p-1)$-th power of their local degrees; definitions for $d_{i,\text{local}}$, $D_{\text{local}}$, and other newly introduced notations in this overview are provided in Table 1). For each sampled node $I$, we query all $k$ sites to obtain its global degree $d_I$. It is easy to verify that $Y = D_{\text{local}} \cdot \left(\frac{d_I}{d_{I,\text{local}}}\right)^{p-1}$ serves as an unbiased estimator for $M_{p-1}^{\sigma}$. However, this estimator is unstable due to its high variance. To resolve this issue, we introduce a truncated estimator $X = Y \cdot \mathbf{1}\left\{\frac{d_I}{d_{I,\text{local}}} \le \alpha\right\}$. $X$ is an biased estimator of $M_{p-1}^{\sigma}$, where the bias comes from the nodes whose global-to-local degree ratio exceeds $\alpha$ (i.e., the nodes from small $V^{(\kappa)}$s), which counts for at most a $\frac{k}{\alpha}(= \epsilon)$ fraction of $M_{p-1}^{\sigma}$. We can further show that $O\left(\frac{\alpha^{p-1}}{\epsilon^2}\right) = O\left(\frac{k^{p-1}}{\epsilon^{p+1}}\right)$ samples are enough for an accurate approximation of $\mathbf{E}[X]$.

Now consider the case that the dataset has a small ambiguity. For our method to fail, the adversary must select some nodes from large $V^{(\kappa)}$s and change their global-to-local degree ratios by adding or removing edges, in order to push them above the threshold $\alpha$. To this end, we show that the

adversary's most effective strategy is to remove local edges. As a simple example, consider the case when $p = 2$ and a node $i$ such that $d_i^\tau / d_{i,\text{local}}^\tau = \alpha/2$. To increase $d_i^\tau / d_{i,\text{local}}^\tau$ to $\alpha$, the adversary needs to remove roughly $d_{i,\text{local}}^\tau/2 = d_i^\tau/\alpha$ local edges incident to node $i$, which eliminates node $i$ from the set of low global-to-local ratio nodes. Since the contribution of node $i$ to $F_2$ is $d_i \le d_i^\tau$, this operation will only increase the bias of $X$ by at most $d_i^\tau$. Therefore, given a budget of $\eta_2 F_2$ edge insertions/deletions, the adversary can only increase the bias of the input by at most $\alpha\eta_2 F_2$, which is at most $\epsilon F_2$ when $\eta_2 \le \frac{\epsilon}{\alpha} = O\left(\frac{\epsilon^2}{k}\right)$. We can extend this argument to the general setting and any $p \ge 2$, showing that the bias of $X$ is at most $\epsilon F_p$ when $\eta_p \le O\left(\frac{\epsilon^p}{k^{p-1}}\right)$.

**Algorithm and Analysis.** The algorithm is described in Algorithm 4. Let us briefly describe it in words. In the first round, each site sends the sum of the local $(p-1)$-th degree moments of its nodes to the coordinator, who aggregates them to obtain $D_{\text{local}}$, the total $(p-1)$-th degree moments over all $m$ nodes. In the second round, the sites and the coordinator jointly sample each node with probability proportional to its local $(p-1)$-th degree moment. In the third round, the exact global degree of each sampled node is computed. Finally, using $D_{\text{local}}$ together with the local and global $(p-1)$-th degree moments of the sampled nodes, the coordinator constructs the final truncated estimator.

We have the following result regarding Algorithm 4.

Theorem 4.1. *For any constant $p \ge 2$ and a noisy size-$m$ dataset $\sigma$ with $\eta_p \le \frac{\epsilon^p}{4^{p+1}\cdot k^{p-1}}$, Algorithm 4 computes an $((\epsilon + O(\eta_p)), 0.01)$-approximation of $F_p$ in the coordinator model with $k$ sites, using three rounds and $O\left(\frac{k^p}{\epsilon^{p+1}}\right)$ words of communication.*

In the rest of this section, we prove Theorem 4.1. The round complexity is clear from Algorithm 4. The communication cost is dominated by communicating $S$ between sites and the coordinator, which uses $O(kt) = O(k^p/\epsilon^{p+1})$ words. In what follows, we prove the correctness of the algorithm.

We start with the following simple fact.

Fact 4.2. *For a noiseless dataset $\tau$, $F_p = \sum_{i\in[m]} (d_i^\tau)^{p-1}$.*

Proof. Let $n_\tau$ be the number of distinct elements in $\tau$, and $\{V_1, \ldots, V_{n_\tau}\}$ be the set of cliques in $G^\tau$. Then we have

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

□

The following lemma shows that each summand at Line 20 of Algorithm 4 does not deviate from $F_p$ by much.

Lemma 4.3. *For a dataset $\sigma$, let $I \in S$ be a sampled node from Algorithm 4 and $X = D_{local} \cdot \left(\frac{d_I}{d_{I,local}}\right)^{p-1} \cdot \mathbf{1}\left\{\left(\frac{d_I}{d_{I,local}}\right)^{p-1} \le \theta\right\}$. We have $\left(1 - 2\eta_p - \frac{\epsilon}{2}\right) F_p \le \mathbf{E}[X] \le (1+\eta_p)F_p$.*

Proof. We start with the second inequality. Let $p_i = \mathbf{Pr}[I = i]$. We have

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

**Algorithm 4:** Distributed-Robust-$F_p$-Low-Noise

**Input:** a noisy dataset $\sigma$ of size $m$ partitioned among $k$ sites, where the $\kappa$-th site holds $\sigma^{(\kappa)}$; a parameter $\epsilon$

**Output:** an $((\epsilon + O(\eta_p)), 0.01)$-approximation of $F_p$

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

// First Round:

Coordinator: send "start" to all sites.

**for** *site* $\kappa \leftarrow 1, \ldots, k$ **do**

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

  send $D_{\text{local}}^{(\kappa)}$ to the coordinator

// Second Round:

Coordinator: calculate $D_{\text{local}} \leftarrow \sum_{\kappa \in [k]} D_{\text{local}}^{(\kappa)}$. Then, pick $t$ i.i.d. samples from $[k]$ according to probabilities $\left\{\frac{D_{\text{local}}^{(\kappa)}}{D_{\text{local}}}\right\}_{\kappa \in [k]}$. Let $s_\kappa$ be the number of times $\kappa$ is sampled. Send $s_\kappa$ to site $\kappa$ for each $\kappa \in [k]$.

**for** *site* $\kappa \leftarrow 1, \ldots, k$ **do**

  pick $s_\kappa$ i.i.d. samples from $V^{(\kappa)}$ according to probabilities $\left\{\frac{(d_{i,\text{local}})^{p-1}}{D_{\text{local}}^{(\kappa)}}\right\}_{i \in V^{(\kappa)}}$

  send all the sampled nodes to the coordinator

// Third Round:

Coordinator: let $S$ be the set of sampled nodes. Send $S$ to all sites.

**for** *site* $\kappa \leftarrow 1, \ldots, k$ **do**

  **foreach** $i \in S$ **do**

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

  send $\{d_i^{(\kappa)}\}_{i \in S}$ to the coordinator

// Output:

Coordinator: calculate $d_i \leftarrow \sum_{\kappa \in [k]} d_i^{(\kappa)}$ for each $i \in S$.

**return** $\frac{1}{t} \sum_{i \in S} \left( D_{\text{local}} \cdot \left(\frac{d_i}{d_{i,\text{local}}}\right)^{p-1} \cdot \mathbf{1}\left\{\left(\frac{d_i}{d_{i,\text{local}}}\right)^{p-1} \le \theta\right\}\right)$.

Let $Y = D_{\text{local}} \cdot \left(\frac{d_I}{d_{I,\text{local}}}\right)^{p-1}$. By the definition of $X$, we have

$$\mathbf{E}[X] \le \mathbf{E}[Y] = \sum_{i \in [m]} p_i \cdot D_{\text{local}} \cdot \left(\frac{d_i}{d_{i,\text{local}}}\right)^{p-1} = \sum_{i \in [m]} (d_i)^{p-1}. \tag{18}$$

Note that

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

By (19) and Fact 4.2, we have

$$(1 - \eta_p) F_p \le \sum_{i \in [m]} (d_i)^{p-1} \le (1 + \eta_p) F_p. \tag{20}$$

Combining (18) and (20), it follows that $\mathbf{E}[X] \le (1+\eta_p)F_p$.

We now prove the first inequality. We start by bounding the difference between $X$ and $Y$.

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

For the first part,

$$① \le \sum_{i\in[m]} \left(\left|B_i^\sigma \cup B_i^\tau\right|^{p-1} - \left|B_i^\sigma \cap B_i^\tau\right|^{p-1}\right) = \eta_p F_p. \tag{21}$$

For the second part, we divide $[m]$ into two sets using a threshold $\alpha = \left(\frac{\theta}{2}\right)^{\frac{1}{p-1}}$:

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

We utilize the following two claims to bound the contributions of items in $H$ and $L$, respectively. Their proofs will be given shortly.

Claim 4.4.

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

Claim 4.5.

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

By Claim 4.4 and Claim 4.5, we have

$$\begin{aligned}
② &\le \sum_{i\in H} (d_i^\tau)^{p-1} + \sum_{i\in L} (d_i^\tau)^{p-1} \cdot \mathbf{1}\left\{\left(\frac{d_i}{d_{i,\text{local}}}\right)^{p-1} > \theta\right\} \\
&\le \frac{kF_p}{\alpha} + \theta\eta_p F_p \le \frac{\epsilon F_p}{2},
\end{aligned} \tag{22}$$

where in the last inequality, we have used $\theta = 2\left(\frac{4k}{\epsilon}\right)^{p-1}$, $\alpha = \left(\frac{\theta}{2}\right)^{\frac{1}{p-1}} = \frac{4k}{\epsilon}$, and our assumption $\eta_p \le \frac{\epsilon^p}{4^{p+1}\cdot k^{p-1}}$. Combining (21) and (22), we have

$$\mathbf{E}[Y-X] \le \eta_p F_p + \frac{\epsilon F_p}{2}. \tag{23}$$

By (18), (20) and (23), we have

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

□

Finally, we prove the two claims.

Proof of Claim 4.4. Let $n_\tau$ be the number of distinct elements in $\tau$, and $\{V_1, \ldots, V_{n_\tau}\}$ be the set of cliques in $G^\tau$. Note that

$$\begin{aligned}
\sum_{i \in H} (d_i^\tau)^{p-1} &= \sum_{\kappa \in [k]} \sum_{j \in [n_\tau]} \sum_{i \in V^{(\kappa)} \cap V_j} (d_i^\tau)^{p-1} \cdot \mathbf{1}\left\{ \frac{d_i^\tau}{d_{i,\text{local}}^\tau} > \alpha \right\} && (24)\\
&= \sum_{\kappa \in [k]} \sum_{j \in [n_\tau]} \sum_{i \in V^{(\kappa)} \cap V_j} \left|V_j\right|^{p-1} \cdot \mathbf{1}\left\{ \frac{\left|V_j\right|}{\left|V^{(\kappa)} \cap V_j\right|} > \alpha \right\} && (25)\\
&= \sum_{\kappa \in [k]} \sum_{j \in [n_\tau]} \left|V_j\right|^{p-1} \cdot \left|V^{(\kappa)} \cap V_j\right| \cdot \mathbf{1}\left\{ \frac{\left|V_j\right|}{\left|V^{(\kappa)} \cap V_j\right|} > \alpha \right\}\\
&< \sum_{\kappa \in [k]} \sum_{j \in [n_\tau]} \frac{\left|V_j\right|^p}{\alpha} = \sum_{\kappa \in [k]} \frac{F_p}{\alpha} = \frac{k F_p}{\alpha},
\end{aligned}$$

where from (24) to (25), we use the fact that if $i \in V_j \cap V^{(\kappa)}$, then $d_i^\tau = \left|B_i^\tau\right| = \left|V_j\right|$ and $d_{i,\text{local}}^\tau = \left|B_i^\tau \cap V^{(\kappa)}\right| = \left|V_j \cap V^{(\kappa)}\right|$. □

Proof of Claim 4.5. For any $i \in L$ such that $\left(\frac{d_i}{d_{i,\text{local}}}\right)^{p-1} > \theta$, we have

$$\begin{aligned}
\theta &< \frac{(d_i)^{p-1}}{(d_{i,\text{local}})^{p-1}} && (26)\\
&\leq \frac{\left|B_i^\sigma \cup B_i^\tau\right|^{p-1}}{\left|B_i^\sigma \cap B_i^\tau \cap V^{(\kappa(i))}\right|^{p-1}} && (27)\\
&= \frac{(d_i^\tau)^{p-1} + \left(\left|B_i^\sigma \cup B_i^\tau\right|^{p-1} - (d_i^\tau)^{p-1}\right)}{(d_{i,\text{local}}^\tau)^{p-1} - \left((d_{i,\text{local}}^\tau)^{p-1} - \left|B_i^\sigma \cap B_i^\tau \cap V^{(\kappa(i))}\right|^{p-1}\right)}, && (28)
\end{aligned}$$

where from (26) to (27) we use $d_i = \left|B_i^\sigma\right| \leq \left|B_i^\sigma \cup B_i^\tau\right|$ and $d_{i,\text{local}} = \left|B_i^\sigma \cap V^{(\kappa(i))}\right| \geq \left|B_i^\sigma \cap B_i^\tau \cap V^{(\kappa(i))}\right|$.

Rearranging the terms in (28) gives

$$\begin{aligned}
\theta(d_{i,\text{local}}^\tau)^{p-1} - (d_i^\tau)^{p-1} &\leq \left(\left|B_i^\sigma \cup B_i^\tau\right|^{p-1} - (d_i^\tau)^{p-1}\right) + \theta\left((d_{i,\text{local}}^\tau)^{p-1} - \left|B_i^\sigma \cap B_i^\tau \cap V^{(\kappa(i))}\right|^{p-1}\right)\\
&\leq \left(\left|B_i^\sigma \cup B_i^\tau\right|^{p-1} - (d_i^\tau)^{p-1}\right) + \theta\left((d_i^\tau)^{p-1} - \left|B_i^\sigma \cap B_i^\tau\right|^{p-1}\right)\\
&\leq \theta\left(\left|B_i^\sigma \cup B_i^\tau\right|^{p-1} - \left|B_i^\sigma \cap B_i^\tau\right|^{p-1}\right), && (29)
\end{aligned}$$

where in the second inequality, we use the fact that for any $p \geq 2$ and any three sets $A, B, C$ such that $B \subseteq A$, we always have $|A|^{p-1} - |B|^{p-1} \geq |A \cap C|^{p-1} - |B \cap C|^{p-1}$. The last inequality holds since $\theta \geq 1$.

On the other hand, since $i \in L$ and $\alpha = \left(\frac{\theta}{2}\right)^{\frac{1}{p-1}}$, we have

$$\theta(d^\tau_{i,\text{local}})^{p-1} - (d^\tau_i)^{p-1} \geq \theta\left(\frac{d^\tau_i}{\alpha}\right)^{p-1} - (d^\tau_i)^{p-1} = (d^\tau_i)^{p-1}. \tag{30}$$

Combining (29) and (30), we have for any $i \in L$ such that $\left(\frac{d_i}{d_{i,\text{local}}}\right)^{p-1} > \theta$,

$$(d^\tau_i)^{p-1} \leq \theta\left(\left|B^\sigma_i \cup B^\tau_i\right|^{p-1} - \left|B^\sigma_i \cap B^\tau_i\right|^{p-1}\right).$$

It follows that

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

□

Lemma 4.6. *Let $\bar{X} = \frac{1}{t}\sum_{i\in[t]} X_i$, where $X_i$ is defined as in Lemma 4.3 for the i-th sample in $S$. Then with probability at least $\frac{2}{3}$, $(1-2\eta_p-\epsilon)F_p \leq \bar{X} \leq (1+\eta_p+\epsilon)F_p$.*

Proof. We first bound the variance of $X_i$.

$$\begin{aligned}
\mathbf{Var}[X_i] &\leq \mathbf{E}[X_i^2] \\
&\leq \theta \cdot D_{\text{local}} \cdot \mathbf{E}[X_i] \quad (\text{since } X_i \leq \theta \cdot D_{\text{local}}) \\
&= \theta \cdot \left(\sum_{i\in[m]} (d_{i,\text{local}})^{p-1}\right) \cdot \mathbf{E}[X_i] \leq \theta \cdot \left(\sum_{i\in[m]} (d_i)^{p-1}\right) \cdot \mathbf{E}[X_i] \\
&\leq \theta \cdot (1+\eta_p)F_p \cdot \mathbf{E}[X_i] \quad (\text{by (20)}) \\
&\leq \theta \cdot \frac{1+\eta_p}{1-2\eta_p-\frac{\epsilon}{2}} \cdot \mathbf{E}[X_i]^2 \quad (\text{by Lemma 4.3}) \\
&\leq 6\theta \cdot \mathbf{E}[X_i]^2. \quad (\text{by assumption } \eta_p \leq \epsilon^p/(4^{p+1}\cdot k^{p-1}))
\end{aligned}$$

By Chebyshev's inequality,

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

By Lemma 4.3, we have with probability $\frac{2}{3}$,

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

which completes the proof. □

Finally, Theorem 4.1 follows directly from Lemma 4.6. Note that we can always apply the median trick to reduce the error probability of Algorithm 4 to 0.01 without changing the asymptotic communication complexity and the round complexity.

Remark 4.7. *We note that the lower bound of $\Omega\left(\frac{k^{p-1}}{\epsilon^2}\right)$ bits for $F_p$ in the coordinator model with noiseless data [36] also holds in the presence of noise. Consequently, in terms of $k$, the communication cost in Theorem 4.1 is only a factor of $k$ from optimal.*

Remark 4.8. *In our lower bound construction (see the full version of this paper [32]), the input of $F_p$, obtained from a YES instance of $DISJ_{n,k}$, has mismatch ambiguity at most $\frac{c_2\epsilon}{k}$ for a constant $c_2 = 10p$. For $p = 2$, the ambiguity threshold in Theorem 4.1 is up to $\frac{c_1\epsilon^2}{k}$ for a constant $c_1 = \frac{1}{4^{p+1}}$. Thus, the communication cost phase transition for $F_2$ happens in the range $[\frac{c_1\epsilon^2}{k}, \frac{c_2\epsilon}{k}]$.*

## 5 Concluding Remarks

In this work, we introduce the parameter *$F_p$-mismatch-ambiguity* and leverage it to design sublinear algorithms for the $F_p$ problem on noisy data in the data stream model and the coordinator model. Several directions remain open following this work. First, there are still gaps in $\epsilon$ between our upper and lower bounds that need to be closed.

Second, in the coordinator model, it would be interesting to obtain a complete characterization of the tradeoff between mismatch ambiguity and communication cost. In particular, is the communication cost phase transition mentioned in Remark 4.8 smooth or abrupt?

Third, it is worth exploring a setting where the similarity oracle $f$, given two items $\sigma_i$ and $\sigma_j$, outputs a real value in $[0, 1]$ rather than a Boolean value in $\{0, 1\}$ that indicates whether they are similar or dissimilar. By convention, we set $f(i, i) = 1$ for all $i \in [m]$. Define $g : [m] \times [m] \to \{0, 1\}$ such that $g(i, j) = 1$ if $\sigma_i$ and $\sigma_j$ correspond to the same ground truth element, and $g(i, j) = 0$ otherwise. We can then generalize the definition of $F_p$-mismatch-ambiguity as follows:

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

We leave open the question of whether efficient streaming and distributed algorithms can be developed to estimate $F_p$ when using $\eta_p^{\mathrm{wt}}$ as a parameter in the approximation ratio.

Finally, it would be interesting to study other problems under the mismatch ambiguity framework.

## References


[1] Noga Alon, Yossi Matias, and Mario Szegedy. 1999. The Space Complexity of Approximating the Frequency Moments. *J. Comput. Syst. Sci.* 58, 1 (1999), 137–147.

[2] Alexandr Andoni. 2017. High frequency moments via max-stability. In *ICASSP*. 6364–6368.

[3] Alexandr Andoni, Robert Krauthgamer, and Krzysztof Onak. 2011. Streaming Algorithms via Precision Sampling. In *FOCS*. 363–372.

[4] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. 2004. An information statistics approach to data stream and communication complexity. *J. Comput. Syst. Sci.* 68, 4 (2004), 702–732.

[5] Lakshminath Bhuvanagiri, Sumit Ganguly, Deepanjan Kesh, and Chandan Saha. 2006. Simpler algorithm for estimating frequency moments of data streams. In *SODA*. 708–713.

[6] Mark Braverman and Or Zamir. 2025. Optimality of Frequency Moment Estimation. In *STOC*. 360–370.

[7] Vladimir Braverman and Rafail Ostrovsky. 2010. Recursive Sketching For Frequency Moments. *CoRR* abs/1011.2571 (2010).

[8] Vladimir Braverman and Rafail Ostrovsky. 2013. Generalizing the Layering Method of Indyk and Woodruff: Recursive Sketches for Frequency-Based Vectors on Streams. In *APPROX-RANDOM (Lecture Notes in Computer Science, Vol. 8096)*, Prasad Raghavendra, Sofya Raskhodnikova, Klaus Jansen, and José D. P. Rolim (Eds.). Springer, 58–70.

[9] Amit Chakrabarti, Subhash Khot, and Xiaodong Sun. 2003. Near-Optimal Lower Bounds on the Multi-Party Communication Complexity of Set Disjointness. In *CCC*. IEEE Computer Society, 107–117.

[10] Di Chen and Qin Zhang. 2016. Streaming Algorithms for Robust Distinct Elements. In *SIGMOD*, Fatma Özcan, Georgia Koutrika, and Sam Madden (Eds.). ACM, 1433–1447.

[11] Jiecao Chen and Qin Zhang. 2018. Distinct Sampling on Streaming Data with Near-Duplicates. In *PODS*, Jan Van den Bussche and Marcelo Arenas (Eds.). ACM, 369–382.

[12] Graham Cormode and Minos N. Garofalakis. 2007. Sketching probabilistic data streams. In *SIGMOD*, Chee Yong Chan, Beng Chin Ooi, and Aoying Zhou (Eds.). ACM, 281–292.

[13] Graham Cormode, S. Muthukrishnan, and Ke Yi. 2008. Algorithms for distributed functional monitoring. In *SODA*. 1076–1085.

[14] Graham Cormode, S. Muthukrishnan, Ke Yi, and Qin Zhang. 2012. Continuous sampling from distributed streams. *J. ACM* 59, 2 (2012), 10:1–10:25.
[15] Xin Luna Dong and Felix Naumann. 2009. Data fusion: resolving data conflicts for integration. *Proceedings of the VLDB Endowment* 2, 2 (2009), 1654–1655.
[16] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. 2007. Duplicate Record Detection: A Survey. *IEEE Trans. Knowl. Data Eng.* 19, 1 (2007), 1–16.
[17] Hossein Esfandiari, Praneeth Kacham, Vahab Mirrokni, David P. Woodruff, and Peilin Zhong. 2024. Optimal Communication Bounds for Classic Functions in the Coordinator Model and Beyond. In *STOC*, Bojan Mohar, Igor Shinkar, and Ryan O'Donnell (Eds.). ACM, 1911–1922.
[18] Uriel Feige. 2004. On sums of independent random variables with unbounded variance, and estimating the average degree in a graph. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, USA, June 13-16, 2004*, László Babai (Ed.). ACM, 594–603.
[19] Sumit Ganguly. 2011. Polynomial Estimators for High Frequency Moments. *CoRR* abs/1104.4552 (2011).
[20] Sumit Ganguly. 2012. A Lower Bound for Estimating High Moments of a Data Stream. *CoRR* abs/1201.0253 (2012).
[21] Sumit Ganguly and David P. Woodruff. 2018. High Probability Frequency Moment Sketches. In *ICALP (LIPIcs, Vol. 107)*, Ioannis Chatzigiannakis, Christos Kaklamanis, Dániel Marx, and Donald Sannella (Eds.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 58:1–58:15.
[22] Thomas N Herzog, Fritz J Scheuren, and William E Winkler. 2007. *Data quality and record linkage techniques*. Vol. 1. Springer.
[23] Zengfeng Huang, Zhongzheng Xiong, Xiaoyi Zhu, and Zhewei Wei. 2025. Simple and Optimal Algorithms for Heavy Hitters and Frequency Moments in Distributed Models. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing, STOC 2025, Prague, Czechia, June 23-27, 2025*, Michal Koucký and Nikhil Bansal (Eds.). ACM, 371–382.
[24] Piotr Indyk and David P. Woodruff. 2005. Optimal approximations of the frequency moments of data streams. In *STOC*. 202–208.
[25] Rajesh Jayaram and David P. Woodruff. 2023. Towards Optimal Moment Estimation in Streaming and Distributed Models. *ACM Trans. Algorithms* 19, 3 (2023), 27:1–27:35.
[26] T. S. Jayram, Satyen Kale, and Erik Vee. 2007. Efficient aggregation algorithms for probabilistic data. In *SODA*, Nikhil Bansal, Kirk Pruhs, and Clifford Stein (Eds.). SIAM, 346–355.
[27] T. S. Jayram, Andrew McGregor, S. Muthukrishnan, and Erik Vee. 2007. Estimating statistical aggregates on probabilistic data streams. In *PODS*, Leonid Libkin (Ed.). ACM, 243–252.
[28] Cheqing Jin, Ke Yi, Lei Chen, Jeffrey Xu Yu, and Xuemin Lin. 2008. Sliding-window top-k queries on uncertain streams. *Proc. VLDB Endow.* 1, 1 (2008), 301–312.
[29] Ravi Kannan, Santosh S. Vempala, and David P. Woodruff. 2014. Principal Component Analysis and Higher Correlations for Distributed Data. In *COLT (JMLR Workshop and Conference Proceedings, Vol. 35)*. 1040–1057.
[30] Nick Koudas, Sunita Sarawagi, and Divesh Srivastava. 2006. Record linkage: similarity measures and algorithms. In *SIGMOD*. ACM, 802–803.
[31] Yi Li and David P. Woodruff. 2013. A Tight Lower Bound for High Frequency Moment Estimation with Small Error. In *APPROX-RANDOM (Lecture Notes in Computer Science, Vol. 8096)*, Prasad Raghavendra, Sofya Raskhodnikova, Klaus Jansen, and José D. P. Rolim (Eds.). Springer, 623–638.
[32] Kaiwen Liu and Qin Zhang. 2026. Frequency Moments in Noisy Streaming and Distributed Data under Mismatch Ambiguity. arXiv:2603.11216 [cs.DS] https://arxiv.org/abs/2603.11216
[33] Morteza Monemizadeh and David P. Woodruff. 2010. 1-Pass Relative-Error $L_p$-Sampling with Applications. In *SODA*. 1143–1160.
[34] Jeff M. Phillips, Elad Verbin, and Qin Zhang. 2012. Lower bounds for number-in-hand multiparty communication complexity, made easy. In *SODA*, Yuval Rabani (Ed.). SIAM, 486–501.
[35] David P. Woodruff. 2004. Optimal space lower bounds for all frequency moments. In *SODA*. 167–175.
[36] David P. Woodruff and Qin Zhang. 2012. Tight bounds for distributed functional monitoring. In *STOC*, Howard J. Karloff and Toniann Pitassi (Eds.). ACM, 941–960.
[37] Qin Zhang. 2015. Communication-Efficient Computation on Distributed Noisy Datasets. In *SPAA*. 313–322.
[38] Qin Zhang. 2025. Robust Statistical Analysis on Streaming Data with Near-Duplicates in General Metric Spaces. *Proc. ACM Manag. Data* 3, 2 (2025), 111:1–111:25.
[39] Qin Zhang and Mohsen Heidari. 2025. Quantum Data Sketches. In *ICDT (LIPIcs, Vol. 328)*, Sudeepa Roy and Ahmet Kara (Eds.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 16:1–16:19.
[40] Qin Zhang, Feifei Li, and Ke Yi. 2008. Finding frequent items in probabilistic data. In *SIGMOD*, Jason Tsong-Li Wang (Ed.). ACM, 819–832.

Received December 2025; accepted February 2026