Tight Bounds for Distributed Functional Monitoring

David P. Woodruff
IBM Almaden
dpwoodru@us.ibm.com

Qin Zhang
MADALGO*, Aarhus University
qinzhang@cs.au.dk

ABSTRACT

We resolve several fundamental questions in the area of distributed functional monitoring, initiated by Cormode, Muthukrishnan, and Yi (SODA, 2008), and receiving recent attention. In this model there are k sites each tracking their input streams and communicating with a central coordinator. The coordinator's task is to continuously maintain an approximate output to a function computed over the union of the k streams. The goal is to minimize the number of bits communicated.

Let the p-th frequency moment be defined as $F_p = \sum_i f_i^p$, where f_i is the frequency of element i. We show the randomized communication complexity of estimating the number of distinct elements (that is, F_0) up to a $1+\varepsilon$ factor is $\Omega(k/\varepsilon^2)$, improving upon the previous $\Omega(k+1/\varepsilon^2)$ bound and matching known upper bounds. For F_p, $p>1$, we improve the previous $\Omega(k+1/\varepsilon^2)$ communication bound to $\tilde{\Omega}(k^{p-1}/\varepsilon^2)$. We obtain similar improvements for heavy hitters, empirical entropy, and other problems. Our lower bounds are the first of any kind in distributed functional monitoring to depend on the product of k and $1/\varepsilon^2$. Moreover, the lower bounds are for the static version of the distributed functional monitoring model where the coordinator only needs to compute the function at the time when all k input streams end; surprisingly they almost match what is achievable in the (dynamic version of) distributed functional monitoring model where the coordinator needs to keep track of the function continuously at any time step. We also show that we can estimate F_p, for any $p>1$, using $\tilde{O}(k^{p-1}\mathrm{poly}(\varepsilon^{-1}))$ communication. This drastically improves upon the previous $\tilde{O}(k^{2p+1}N^{1-2/p}\mathrm{poly}(\varepsilon^{-1}))$ bound of Cormode, Muthukrishnan, and Yi for general p, and their $\tilde{O}(k^2/\varepsilon + k^{1.5}/\varepsilon^3)$ bound for $p=2$. For $p=2$, our bound resolves their main open question.

Our lower bounds are based on new direct sum theorems for approximate majority, and yield improvements to classical problems in the standard data stream model. First, we improve the known lower bound for estimating $F_p, p>2$, in t passes from $\tilde{\Omega}(n^{1-2/p}/(\varepsilon^{2/p}t))$ to $\tilde{\Omega}(n^{1-2/p}/(\varepsilon^{4/p}t))$, giving the first bound that matches what we expect when $p=2$ for any constant number of passes. Second, we give the first lower bound for estimating F_0 in t passes with $\Omega(1/(\varepsilon^2 t))$ bits of space that does not use the hardness of the gap-hamming problem.

Categories and Subject Descriptors

F.2.0 [**ANALYSIS OF ALGORITHMS AND PROBLEM COMPLEXITY**]: General

General Terms

Algorithms, theory

Keywords

Distributed functional monitoring, data streams, frequency moments, heavy hitters, quantiles, entropy

1. INTRODUCTION

Recent applications in sensor networks and distributed systems have motivated the *distributed functional monitoring* model, initiated by Cormode, Muthukrishnan, and Yi [20]. In this model there are k *sites* and a single central *coordinator*. Each site S_i ($i \in [k]$) receives a stream of data $A_i(t)$ for timesteps $t=1,2,\ldots$, and the coordinator wants to keep track of a function f that is defined over the multiset union of the k data streams at each time t. For example, the function f could be the number of distinct elements in the union of the k streams. We assume that there is a two-way communication channel between each site and the coordinator so that the sites can communicate with the coordinator. The goal is to minimize the total amount of communication between the sites and the coordinator so that the coordinator can approximately maintain $f(A_1(t),\ldots,A_k(t))$ at any time t. Minimizing the total communication is motivated by power constraints in sensor networks, since communication typically uses a power-hungry radio [25]; and also by network bandwidth constraints in distributed systems. There is a large body of work on monitoring problems in this model, including maintaining a random sample [21, 48], estimating frequency moments [18, 20], finding the heavy hitters [6, 40, 43, 53], approximating the quantiles [19, 33, 53], and estimating the entropy [5].

We can think of the distributed functional monitoring model as follows. Each of the k sites holds an N-dimentional vector where N is the size of the universe. An update to a coordinate j on site S_i causes v_j^i to increase by 1. The goal is to estimate a statistic of $v=\sum_{i=1}^k v^i$, such as the p-th frequency moment $F_p=\|v\|_p^p$, the number of distinct elements $F_0=|\mathrm{support}(v)|$, and the empirical entropy $H=\sum_i \frac{v_i}{\|v\|_1}\log\frac{\|v\|_1}{v_i}$. This is the standard *insertion-only* model. For many of these problems, with the exception of

*MADALGO is the Center for Massive Data Algorithmics - a Center of the Danish National Research Foundation.

STOC'12, May 19–22, 2012, New York, New York, USA.

the empirical entropy, there are strong lower bounds (e.g., $\Omega(N)$) if allowing updates to coordinates that cause v_j^i to decrease [5]. The latter is called the *update* model. Thus, except for entropy, we follow previous work and consider the insertion-only model.

To prove lower bounds, we consider the static version of the distributed functional monitoring model, where the coordinator only needs to compute the function at the time when all k input streams end. It is clear that a lower bound for the static case is also a lower bound for the dynamic case in which the coordinator has to keep track of the function at any point in time. The static version of the distributed functional monitoring model is closely related to the multiparty *number-in-hand* communication model, where we again have k sites each holding an N-dimensional vector v^i, and they want to jointly compute a function defined on the k input vectors. It is easy to see that these two models are essentially the same since in the former, if site S_i would like to send a message to S_j, it can always send the message first to the coordinator and then the coordinator can forward the message to S_j. Doing this will only increase the total amount of communication by a factor of two. Therefore, we do not distinguish between these two models in this paper.

There are two variants of the multiparty number-in-hand communication model we will consider: the *blackboard model*, in which each message a site sends is received by all other sites, i.e., it is broadcast, and the *message-passing* model, in which each message is between the coordinator and a specific site.

Despite the large body of work in the distributed functional monitoring model, the complexity of basic problems is not well understood. For example, for estimating F_0 up to a $(1+\varepsilon)$-factor, the best upper bound is $\tilde{O}(k/\varepsilon^2)$ [20][1], while the only known lower bound is $\Omega(k+1/\varepsilon^2)$. The dependence on ε in the lower bound is not very insightful, as the $\Omega(1/\varepsilon^2)$ bound follows just by considering two sites [5, 16]. The real question is whether the k and $1/\varepsilon^2$ factors should multiply. Even more embarrassingly, for the frequency moments F_p, $p>2$, the known algorithms use communication $\tilde{O}(k^{2p+1}N^{1-2/p}\mathrm{poly}(1/\varepsilon))$, while the only known lower bound is $\Omega(k+1/\varepsilon^2)$ [5, 16]. Even for $p=2$, the best known upper bound is $\tilde{O}(k^2/\varepsilon+k^{1.5}/\varepsilon^3)$ [20], and the authors' main open question in their paper is "It remains to close the gap in the F_2 case: can a better lower bound than $\Omega(k)$ be shown, or do there exist $\tilde{O}(k\cdot\mathrm{poly}(1/\varepsilon))$ solutions?"

Our Results: We significantly improve the previous communication bounds for approximating the frequency moments, entropy, heavy hitters, and quantiles in the distributed functional monitoring model. In many cases our bounds are optimal. Our results are summarized in Table 1, where they are compared with previous bounds. We have three main results, each introducing a new technique:

1. We show that for estimating F_0 in the message-passing model, $\Omega(k/\varepsilon^2)$ communication is required, matching an upper bound of [20] up to a polylogarithmic factor. Our lower bound holds in the static model in which the k sites just need to approximate F_0 once on their inputs.

2. We show that we can estimate F_p, for any $p>1$, using $\tilde{O}(k^{p-1}\mathrm{poly}(\varepsilon^{-1}))$ communication in the message-passing model[2]. This drastically improves upon the previous bound $\tilde{O}(k^{2p+1}N^{1-2/p}\mathrm{poly}(\varepsilon^{-1}))$ of [20]. In particular, setting $p=2$, we resolve the main open question of [20].

3. We show $\tilde{\Omega}(k^{p-1}/\varepsilon^2)$ communication is necessary for approximating F_p $(p>1)$ in the blackboard model, significantly improving the prior $\Omega(k+1/\varepsilon^2)$ bound. As with our lower bound for F_0, these are the first lower bounds which depend on the product of k and $1/\varepsilon$. As with F_0, our lower bound holds in the static model in which the sites just approximate F_p once.

[1]We use $\tilde{O}(f)$ to denote a function of the form $f\cdot\log^{O(1)}(Nk/\varepsilon)$.
[2]We assume the total number of updates is $\mathrm{poly}(N)$.

Our other results in Table 1 are explained in the body of the paper, and use similar techniques.

Our Techniques: *Lower Bound for F_0:* Our $\Omega(k/\varepsilon^2)$ bound for F_0 is based on the following primitive problem k-GAP-MAJ. For illustration, suppose $k=1/\varepsilon^2$. There are $1/\varepsilon^2$ sites each holding a random independent bit. Their task is to decide if at least $1/(2\varepsilon^2)+1/\varepsilon$ of the bits are 1, or at most $1/(2\varepsilon^2)-1/\varepsilon$ of the bits are 1. We show any correct protocol must reveal $\Omega(1/\varepsilon^2)$ bits of information about the sites' inputs. We "compose" this with 2-party disjointness (2-DISJ) [46], in which each party has a bitstring of length $1/\varepsilon^2$ and either the strings have disjoint support (the solution is 0) or there is a single coordinate which is 1 in both strings (the solution is 1). Let τ be the hard distribution for 2-DISJ, shown to require $\Omega(1/\varepsilon^2)$ communication to solve [46]. Suppose the coordinator and each site share an instance of 2-DISJ in which the solution to 2-DISJ is a random bit, which is the site's effective input to k-GAP-MAJ. The coordinator has the same input for each of the $1/\varepsilon^2$ instances, while the sites have an independent input drawn from τ conditioned on the coordinator's input and output bit determined by k-GAP-MAJ. The inputs are chosen so that if the output of 2-DISJ is 1, then F_0 increases by 1, otherwise it remains the same. This is not entirely accurate, but it illustrates the main idea. Now, the key is that by the rectangle property of k-party communication protocols, the $1/\varepsilon^2$ different output bits are independent conditioned on the transcript. Thus if a protocol does not reveal $\Omega(1/\varepsilon^2)$ bits of information about these output bits, by an anti-concentration theorem we can show that the protocol cannot succeed with large probability. Finally, since a $(1+\varepsilon)$-approximation to F_0 can decide k-GAP-MAJ, and since any correct protocol for k-GAP-MAJ must reveal $\Omega(1/\varepsilon^2)$ information, the protocol must solve $\Omega(1/\varepsilon^2)$ instances of 2-DISJ, each requiring $\Omega(1/\varepsilon^2)$ communication (otherwise the coordinator could simulate $k-1$ of the sites and obtain an $o(1/\varepsilon^2)$- communication protocol for 2-DISJ with the remaining site, contradicting the communication lower bound for 2-DISJ on this distribution). We obtain an $\Omega(k/\varepsilon^2)$ bound for $k\geq 1/\varepsilon^2$ by using similar arguments. One cannot show this in the blackboard model since there is an $\tilde{O}(k+1/\varepsilon^2)$ bound for F_0 [3].

Lower Bound for F_p: Our $\tilde{\Omega}(k^{p-1}/\varepsilon^2)$ bound for F_p cannot use the above reduction since we do not know how to turn a protocol for approximating F_p into a protocol for solving the composition of k-GAP-MAJ and 2-DISJ. Instead, our starting point is a recent $\Omega(1/\varepsilon^2)$ lower bound for the 2-party gap-hamming distance problem GHD [16]. The parties have a length-$1/\varepsilon^2$ bitstring, x and y, respectively, and they must decide if the Hamming distance $\Delta(x,y)>1/(2\varepsilon^2)+1/\varepsilon$ or $\Delta(x,y)<1/(2\varepsilon^2)-1/\varepsilon$. A simplification by Sherstov [47] shows a related problem called 2-GAP-ORT also has $\Omega(1/\varepsilon^2)$ communication. Here there are two parties, each with $1/\varepsilon^2$-length bitstrings x and y, and they must decide if $|\Delta(x,y)-1/(2\varepsilon^2)|>2/\varepsilon$ or $|\Delta(x,y)-1/(2\varepsilon^2)|<1/\varepsilon$. We observe that Sherstov proves that 2-GAP-ORT is hard when

[3]The idea is to first obtain a 2-approximation. Then, sub-sample so that there are $\Theta(1/\varepsilon^2)$ distinct elements. Then the first party broadcasts his distinct elements, the second party broadcasts the distinct elements he has that the first party does not, etc.

	Previous work	This paper	Previous work	This paper
Problem	LB	LB (all static)	UB	UB
F_0	$\Omega(k)$ [20]	$\mathbf{\Omega(k/\varepsilon^2)}$	$\tilde{O}(k/\varepsilon^2)$ [20]	–
F_2	$\Omega(k)$ [20]	$\mathbf{\tilde{\Omega}(k/\varepsilon^2)}$ (BB)	$\tilde{O}(k^2/\varepsilon + k^{1.5}/\varepsilon^3)$ [20]	$\mathbf{\tilde{O}(\frac{k}{\mathrm{poly}(\varepsilon)})}$
F_p $(p > 1)$	$\Omega(k + 1/\varepsilon^2)$ [5, 16]	$\mathbf{\tilde{\Omega}(k^{p-1}/\varepsilon^2)}$ (BB)	$\tilde{O}(\frac{p}{\varepsilon^{1+2/p}} k^{2p+1} N^{1-2/p})$ [20]	$\mathbf{\tilde{O}(\frac{k^{p-1}}{\mathrm{poly}(\varepsilon)})}$
All-quantile	$\tilde{\Omega}(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})$ [33]	$\mathbf{\Omega}(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})$ (BB)	$\tilde{O}(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})$ [33]	–
Heavy Hitters	$\tilde{\Omega}(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})$ [33]	$\mathbf{\Omega}(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})$ (BB)	$\tilde{O}(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})$ [33]	–
Entropy	$\tilde{\Omega}(1/\sqrt{\varepsilon})$ [5]	$\mathbf{\tilde{\Omega}(k/\varepsilon^2)}$ (BB)	$\tilde{O}(\frac{k}{\varepsilon^3})$ [5], $\tilde{O}(\frac{k}{\varepsilon^2})$ (static) [32]	–
ℓ_p $(p \in (0, 2])$	–	$\mathbf{\tilde{\Omega}(k/\varepsilon^2)}$ (BB)	$\tilde{O}(k/\varepsilon^2)$ (static) [38]	–

Table 1: UB denotes upper bound; LB denotes lower bound; BB denotes blackboard model. N denotes the universe size. All bounds are for randomized algorithms. We assume all bounds hold in the dynamic setting by default, and will state explicitly if they hold in the static setting. For lower bounds we assume the message-passing model by default, and state explicitly if they also hold in the blackboard model.

x and y are drawn from a product uniform distribution [4]. Therefore, by a simulation result of Barak et al. [9], this implies that any correct protocol for 2-GAP-ORT must reveal $\tilde{\Omega}(1/\varepsilon^2)$ [5] information about (x, y). By independence and the chain rule, this means for $\tilde{\Omega}(1/\varepsilon^2)$ indices i, $\tilde{\Omega}(1)$ information is revealed about (x_i, y_i) conditioned on values (x_j, y_j) for $j < i$. We now "embed" an independent copy of a variant of k-party-disjointness, the k-XOR problem, on each of the $1/\varepsilon^2$ coordinates of 2-GAP-ORT. In this variant, there are k parties each holding a bitstring of length k^p. On all but one "special" randomly chosen coordinate, there is a single site assigned to the coordinate and that site uses private randomness to choose whether the value on the coordinate is 0 or 1 (with equal probability), and the remaining $k-1$ sites have 0 on this coordinate. On the special coordinate, with probability $1/4$ all sites have a 0 on this coordinate (a "00" instance), with probability $1/4$ the first $k/2$ parties have a 1 on this coordinate and the remaining $k/2$ parties have a 0 (a "10" instance), with probability $1/4$ the second $k/2$ parties have a 1 on this coordinate and the remaining $k/2$ parties have a 0 (a "01" instance), and with the remaining probability $1/4$ all k parties have a 1 on this coordinate (a "11" instance). We show, via a direct sum for distributional communication complexity, that any deterministic protocol that decides which case the special coordinate is in with probability $1/4+\tilde{\Omega}(1)$ has conditional information cost $\tilde{\Omega}(k^{p-1})$. This implies that any protocol that can decide whether the output is in the set $\{10, 01\}$ (the "XOR" of the output bits) with probability $1/2 + \tilde{\Omega}(1)$ has conditional information cost $\tilde{\Omega}(k^{p-1})$. We do the direct sum argument by conditioning the mutual information on low-entropy random variables which allow us to fill in inputs on remaining coordinates without any communication between the parties and without asymptotically affecting our $\tilde{\Omega}(k^{p-1})$ lower bound. We design a reduction so that on the i-th coordinate of 2-GAP-ORT, the input of the first $k/2$-players of k-XOR is determined by the public coin (which we condition on) and the first party's input bit to 2-GAP-ORT, and the input of the second $k/2$-players of k-XOR is determined by the public coin and the second party's input bit to 2-GAP-ORT . We show that any protocol that solves the composition of 2-GAP-ORT with $1/\varepsilon^2$ copies of k-XOR , a problem that we call k-BTX , must reveal $\tilde{\Omega}(1)$ bits of information about the two output bits of an $\tilde{\Omega}(1)$ fraction of the $1/\varepsilon^2$ copies, and from our $\tilde{\Omega}(k^{p-1})$ information cost lower bound for a single copy, we can obtain an overall $\tilde{\Omega}(k^{p-1}/\varepsilon^2)$ bound. Finally, one can show that a $(1+\varepsilon)$-approximation algorithm for F_p can be used to solve k-BTX .

Upper Bound for F_p: We illustrate the algorithm for $p = 2$ and constant ε. Unlike [20], we do not use AMS sketches [4]. A nice property of our protocol is that it is the first 1-way protocol (the protocol of [20] is not), in the sense that only the sites send messages to the coordinator (the coordinator does not send any messages). Moreover, all messages are simple: if a site receives an update to the j-th coordinate, provided the frequency of coordinate j in its stream exceeds a threshold, it decides with a certain probability to send j to the coordinator. Unfortunately, one can show that this probability cannot be the same for all coordinates j, as otherwise the communication would be too large.

To determine the threshold and probability to send an update to a coordinate j, the sites use the public coin to randomly group all coordinates j into buckets S_ℓ, where S_ℓ contains a $1/2^\ell$ fraction of the input coordinates. For $j \in S_\ell$, the threshold and probability are only a function of ℓ. Inspired by work on sub-sampling [34], we try to estimate the number of coordinates j of magnitude in the range $[2^h, 2^{h+1})$, for each h. Call this class of coordinates C_h. If the contribution to F_2 from C_h is significant, then $|C_h| \approx 2^{-2h} \cdot F_2$, and to estimate $|C_h|$ we only consider those $j \in C_h$ that are in S_ℓ for a value ℓ which satisfies $|C_h| \cdot 2^{-\ell} \approx 2^{-2h} \cdot F_2 \cdot 2^{-\ell} \approx 1$. We do not know F_2 and so we also do not know ℓ, but we can make a logarithmic number of guesses. We note that the work [34] was available to the authors of [20] for several years, but adapting it to the distributed framework here is tricky in the sense that the "heavy hitters" algorithm used in [34] for finding elements in different C_h needs to be implemented in a k-party communication-efficient way.

When choosing the threshold and probability we have two competing constraints; on the one hand these values must be chosen so that we can accurately estimate the values $|C_h|$ from the samples. On the other hand, these values need to be chosen so that the communication is not excessive. Balancing these two constraints forces us to use a threshold instead of just the same probability for all coordinates in S_ℓ. By choosing the thresholds and probabilities to be appropriate functions of ℓ, we can satisfy both constraints. Other minor issues in the analysis arise from the fact that different classes contribute at different times, and that the coordinator must be correct at all times. These issues can be resolved by conditioning on a quantity related to the protocol's correctness being accurate at a small number of selected times in the stream, and then arguing that the quantity is non-decreasing and that this implies that it is correct at all times.

[4]We note that the hardness under the product uniform distribution may also follow from ideas in [16].

[5]We assume that the communication cost of all protocols in the paper is at most $\mathrm{poly}(N)$, where N is the number of coordinates in the vector inputs to the parties, since otherwise the lower bound can be proved directly (will be discussed in more detail in Section 4.1). In this case, applying Theorem 1.3 of [9], we have that the external information cost of the protocol is at least $\tilde{\Omega}(1/\varepsilon^2)$.

Implications for the Data Stream Model: In 2003, Indyk and Woodruff introduced the GHD problem [35], where a 1-round lower bound shortly followed [50]. Ever since, it seemed the space complexity of estimating F_0 in a data stream with $t > 1$ passes hinged on whether GHD required $\Omega(1/\varepsilon^2)$ communication for t rounds, see, e.g., Question 10 in [2]. A flurry [10, 11, 16, 47, 49] of recent work finally resolved the complexity of GHD. What our lower bound shows for F_0 is that this is not the only way to prove the $\Omega(1/\varepsilon^2)$ space bound for multiple passes for F_0. Indeed, we just needed to look at $1/\varepsilon^2$ parties instead of 2 parties. Since we have an $\Omega(1/\varepsilon^4)$ communication lower bound for F_0 with $1/\varepsilon^2$ parties, this implies an $\Omega((1/\varepsilon^4)/(t/\varepsilon^2)) = \Omega(1/(t\varepsilon^2))$ bound for t-pass algorithms for approximating F_0. Arguably our proof is simpler than the recent GHD lower bounds.

Our $\tilde{\Omega}(k^{p-1}/\varepsilon^2)$ bound for F_p also improves a long line of work on the space complexity of estimating F_p for $p > 2$ in a data stream. The current best upper bound is $\tilde{O}(N^{1-2/p}\varepsilon^{-2})$ bits of space [28]. See Figure 1 of [28] for a list of papers which make progress on the ε and logarithmic factors. The previous best lower bound is $\tilde{\Omega}(N^{1-2/p}\varepsilon^{-2/p}/t)$ for t passes [8]. By setting $k^p = \varepsilon^2 N$, we obtain that the total communication is at least $\tilde{\Omega}(\varepsilon^{2-2/p}N^{1-1/p}/\varepsilon^2)$, and so the implied space lower bound for t-pass algorithms for F_p in a data stream is $\tilde{\Omega}(\varepsilon^{-2/p}N^{1-1/p}/(tk)) = \tilde{\Omega}(N^{1-2/p}/(\varepsilon^{4/p}t))$. This gives the first bound that agrees with the tight $\tilde{\Theta}(1/\varepsilon^2)$ bound when $p = 2$ for any constant t. After our work, Ganguly [29] improved this for the special case $t = 1$. That is, for 1-pass algorithms for estimating $F_p, p > 2$, he shows a space lower bound of $\Omega(N^{1-2/p}/(\varepsilon^2 \log n))$.

As mentioned, we observe that 2-GAP-ORT has information cost $\tilde{\Omega}(1/\varepsilon^2)$ under the product uniform distribution or the protocol must have super-polynomial (in N) communication. Since 2-GAP-ORT can be written as the AND of two GHD instances on $\Theta(1/\varepsilon^2)$ bits (see the Corollary after the Main Theorem in [47]), this implies a useful distribution for which either the communication cost of GHD is super-polynomial or the external information cost is at least $\tilde{\Omega}(1/\varepsilon^2)$, partly answering Question 25 in the Open Problems in Data Streams list from the Bertinoro and IITK workshops [3]. Using standard direct sum theorems, this implies solving r independent instances of F_0 or F_2, say, in a data stream requires $\tilde{\Omega}(r/\varepsilon^2)$ bits of space, which was unknown.

Other Related Work: There are quite a few papers on multiparty number-in-hand communication complexity, though they are not directly relevant for the problems studied in this paper. Alon et al. [4] and Bar-Yossef et al. [8] studied lower bounds for multiparty set-disjointness, which has applications to p-th frequency moment estimation for $p > 2$ in the streaming model. Their results were further improved in [15, 30, 36]. Chakrabarti et al. [13] studied random-partition communication lower bounds for multiparty set-disjointness and pointer jumping, which have a number of applications in the random-order data stream model. Other work includes Chakrabarti et al. [14] for median selection, Magniez et al. [42] and Chakrabarti et al. [12] for streaming language recognition. Very few studies have been conducted in the message-passing model. Duris and Rolim [23] proved several lower bounds in the message-passing model, but only for some simple boolean functions. Three related but more restrictive private-message models were studied by Gal and Gopalan [27], Ergün and Jowhari [24], and Guha and Huang [31]. The first two only investigated deterministic protocols and the third was tailored for the random-order data stream model.

Recently Phillips et al. [45] introduced a technique called symmetrization for the number-in-hand communication model. The idea is to try to find a symmetric hard distribution for the k players. Then one reduces the k-player problem to a 2-player problem by assigning Alice the input of a random player and Bob the inputs of the remaining $k-1$ players. The answer to the k-player problem gives the answer to the 2-player problem. By symmetrization one can argue that if the communication lower bound for the resulting 2-player problem is L, then the lower bound for the k-player problem is $\Omega(kL)$. While symmetrization can be used to solve some problems for which other techniques are not known, such as bitwise AND and OR, it has several serious limitations. First, symmetrization requires a symmetric hard distribution, and for many problems this is not known or unlikely to exist; this is true of all of the problems (except for the auxiliary problem k-GAP-MAJ) considered in this paper. Second, for many problems (e.g., the k-GAP-MAJ), when Bob knows the inputs of $k-1$ players, he can determine the answer without any communication, and so no embedding into a k-player protocol of the form studied in [45] is possible. Also, it does not give information cost bounds, and so it is difficult to use when composing problems as is done in this paper.

Paper Outline: In Section 3 and Section 4 we prove our lower bounds for F_0 and $F_p, p > 1$. The lower bounds apply to functional monitoring, but hold even in the static model. In Section 5 we show improved upper bounds for $F_p, p > 1$, for functional monitoring. Finally, in Section 6 we prove lower bounds for all-quantile, heavy hitters, entropy and ℓ_p for any $p \geq 1$ in the blackboard model.

2. PRELIMINARIES

In this section we review some basics on communication complexity and information theory.

Information Theory We refer the reader to [22] for a comprehensive introduction to information theory. Here we review a few concepts and notation.

Let $H(X)$ denote the Shannon entropy of the random variable X, and let $H_b(p)$ denote the binary entropy function when $p \in [0, 1]$. Let $H(X \mid Y)$ denote conditional entropy of X given Y. Let $I(X;Y)$ denote the mutual information between two random variables X, Y. Let $I(X;Y \mid Z)$ denote the mutual information between two random variables X, Y conditioned on Z. The following is a summarization of the basic properties of entropy and mutual information that we need.

PROPOSITION 1. *Let X, Y, Z be random variables.*

1. *If X takes value in $\{1, 2, \ldots, m\}$, then $H(X) \in [0, \log m]$.*
2. *$H(X) \geq H(X \mid Y)$ and $I(X;Y) = H(X) - H(X \mid Y) \geq 0$.*
3. *If X and Z are independent, then we have $I(X;Y \mid Z) \geq I(X;Y)$.*
4. *(Chain rule of mutual information)*

   $I(X, Y; Z) = I(X;Z) + I(Y;Z \mid X)$.

   And in general, for any random variables $X_1, X_2, \ldots, X_n, Y$,

   $I(X_1, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y \mid X_1, \ldots, X_{i-1})$.
5. *(Data processing inequality) If X and Z are conditionally independent given Y, then $I(X;Y \mid Z) \leq I(X;Y)$.*

6. *(Fano's inequality) Let X be a random variable chosen from domain $\mathcal{X}$ according to distribution μ_X, and Y be a random variable chosen from domain $\mathcal{Y}$ according to distribution μ_Y. For any reconstruction function $g : \mathcal{Y} \to \mathcal{X}$ with error δ_g,*

$$H_b(\delta_g) + \delta_g \log(|\mathcal{X}| - 1) \geq H(X \mid Y).$$

7. *(The Maximum Likelihood Estimation principle) With the notation as in Fano's inequality, if the reconstruction function is $g(y) = x$ for the x that maximizes the conditional probability $\mu_X(x \mid Y = y)$, then*

$$\delta_g \leq 1 - \frac{1}{2^{H(X \mid Y)}}.$$

Communication complexity In the two-party randomized communication complexity model (see e.g., [41]), we have two players Alice and Bob. Alice is given $x \in \mathcal{X}$ and Bob is given $y \in \mathcal{Y}$, and they want to jointly compute a function $f(x, y)$ by exchanging messages according to a protocol Π. Let $\Pi(x, y)$ denote the message transcript when Alice and Bob run protocol Π on input pair (x, y). We sometimes abuse notation by identifying the protocol and the corresponding random transcript, as long as there is no confusion.

The *communication complexity* of a protocol is defined as the maximum number of bits exchanged among all pairs of inputs. We say a protocol Π computes f with error probability δ ($0 \leq \delta \leq 1$) if there exists a function g such that for all input pairs (x, y), $\mathsf{Pr}[g(\Pi(x, y)) \neq f(x, y)] \leq \delta$. The δ-error randomized communication complexity, denoted by $R_\delta(f)$, is the cost of the minimum-communication randomized protocol that computes f with error probability δ. The (μ, δ)-distributional communication complexity of f, denoted by $D^\delta_\mu(f)$, is the cost of the minimum-communication deterministic protocol that gives the correct answer for f on at least a $1 - \delta$ fraction of all input pairs, weighted by distribution μ. Yao [52] showed that $R^\delta(f) \geq \max_\mu D^\delta_\mu(f)$. Thus, one way to prove a lower bound for randomized protocols is to find a hard distribution μ and lower bound $D^\delta_\mu(f)$. This is called Yao's Minimax Principle.

The definitions for two-party protocols can be easily extended to the multiparty setting, where we have k players and the i-th player is given an input $x_i \in \mathcal{X}_i$. Again the k players want to jointly compute a function $f(x_1, x_2, \ldots, x_k)$ by exchanging messages according to a protocol Π.

Information complexity Information complexity was introduced in a series of papers including [8, 17]. We refer the reader to Bar-Yossef's Thesis [7]; see Chapter 6 for a detailed introduction. Here we briefly review the concepts of information cost and conditional information cost for k-player communication problems. All of them are defined in the blackboard number-in-hand model.

Let μ be an input distribution on $\mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_k$ and let X be a random input chosen from μ. Let Π be a randomized protocol running on inputs in $\mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_k$. The *information cost* of Π with respect to μ is $I(X; \Pi)$ [6]. The *information complexity* of a problem f with respect to a distribution μ and error parameter δ ($0 \leq \delta \leq 1$), denoted $\mathrm{IC}_{\mu,\delta}(f)$, is the minimum information cost of a δ-error protocol for f with respect to μ. We will work in the public coin model, in which all parties also share a common source of randomness.

[6]In some of the literature this is called the *external* information cost, in contrast with the *internal* information cost. In this paper we only need the former.

We say a distribution λ partitions μ if conditioned on λ, μ is a product distribution. Let X be a random input chosen from μ and D be a random variable chosen from λ. For a randomized protocol Π on $\mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_k$, the *conditional information cost* of Π with respect to the distribution μ on $\mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_k$ and a distribution λ partitioning μ is defined as $I(X; \Pi \mid D)$. The *conditional information complexity* of a problem f with respect to a distribution μ, a distribution λ partitioning μ, and error parameter δ ($0 \leq \delta \leq 1$), denoted $\mathrm{IC}_{\mu,\delta}(f \mid \lambda)$, is the minimum information cost of a δ-error protocol for f with respect to μ and λ. The following proposition can be found in [8].

PROPOSITION 2. *For any distribution μ, distribution λ partitioning μ, and error parameter δ ($0 \leq \delta \leq 1$),*

$$R_\delta(f) \geq \mathrm{IC}_{\mu,\delta}(f) \geq \mathrm{IC}_{\mu,\delta}(f \mid \lambda).$$

Statistical distance measures Given two probability distributions μ and ν over the same space $\mathcal{X}$, the following statistical distance measures will be used in this paper:

1. Total variation distance: $V(\mu, \nu) \stackrel{def}{=} \max_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)|$.
2. Hellinger distance: $h(\mu, \nu) \stackrel{def}{=} \sqrt{\frac{1}{2} \sum_{x \in \mathcal{X}} \left(\sqrt{\mu(x)} - \sqrt{\nu(x)}\right)^2}$

We have the following relation between total variation distance and Hellinger distance (cf. [7], Chapter 2).

PROPOSITION 3. $h^2(\mu, \nu) \leq V(\mu, \nu) \leq h(\mu, \nu)\sqrt{2 - h^2(\mu, \nu)}$.

Conventions In the rest of the paper we call a player a *site*, as to be consistent with the distributed functional monitoring model. We denote $[n] = \{1, \ldots, n\}$. Let $\oplus$ be the XOR function. All logarithms are base-2 unless noted otherwise. We say $\tilde{W}$ is a $(1 + \varepsilon)$-approximation of W, $0 < \varepsilon < 1$, if $W \leq \tilde{W} \leq (1 + \varepsilon)W$.

3. A LOWER BOUND FOR F_0

We introduce the problem k-GAP-MAJ, and then compose it with 2-DISJ to prove a lower bound for F_0.

3.1 The k-GAP-MAJ Problem

In the k-GAP-MAJ problem we have k sites $S_1, S_2, \ldots, S_k$, and each site has a bit z_i ($1 \leq i \leq k$). The sites want to compute the following function in the blackboard model:

$$k\text{-GAP-MAJ}(z_1, \ldots, z_k) = \begin{cases} 0, & \text{if } \sum_{i \in [k]} z_i \leq \beta k - \sqrt{\beta k}, \\ 1, & \text{if } \sum_{i \in [k]} z_i \geq \beta k + \sqrt{\beta k}, \\ *, & \text{otherwise,} \end{cases}$$

where β ($\omega(1/k) \leq \beta \leq 1/2$) is a parameter, and "$*$" means that the answer can be arbitrary. We define the input distribution μ as follows. For each $i \in [k]$, let $z_i = 1$ with probability β and $z_i = 0$ with probability $(1 - \beta)$.

Let $Z = \{Z_1, Z_2, \ldots, Z_k\}$ be a random input chosen according to distribution μ. Let Π be the transcript of any protocol for k-GAP-MAJ on the random input vector Z. Let $\tilde{\mu}$ be the probability distribution of the random transcript Π.

DEFINITION 1. *We say a transcript π is* weak *if for at least $0.5k$ of Z_i ($i \in [k]$), it holds that $H(Z_i \mid \Pi = \pi) \geq H_b(0.01\beta)$, otherwise we say it is* strong.

In this section we will prove the following main theorem for k-GAP-MAJ. Intuitively, it says that in order to correctly compute k-GAP-MAJ with a good probability, we have to learn $\Omega(k)$ Z_i's well.

THEOREM 1. *If a protocol correctly computes k-GAP-MAJ on input distribution μ with error probability δ for some sufficiently small constant δ, then $\mathsf{Pr}_{\Pi\sim\tilde{\mu}}[\Pi$ is strong$] = \Omega(1)$.*

We have the following immediate corollary, which will be used to prove the lower bound for the quantile problem in Section 6.1.

COROLLARY 1. *Suppose that $\beta = \Theta(1)$, then $I(Z;\Pi) = \Omega(k)$ for any protocol that computes k-GAP-MAJ on input distribution μ with error probability δ for some sufficiently small constant δ.*

PROOF. By the chain rule and independence, we have

$$\begin{aligned} & I(Z;\Pi) \geq \sum_{i\in[k]} I(Z_i;\Pi) \\ \geq & \sum_{\pi:\pi \text{ is strong}} \left(\mathsf{Pr}_{\Pi\sim\tilde{\mu}}[\Pi=\pi] \sum_{i\in[k]} (H(Z_i) - H(Z_i \mid \Pi=\pi)) \right) \\ \geq & \ \Omega(1)\cdot 0.5k \cdot (H_b(\beta) - H_b(0.01\beta)) \\ \geq & \ \Omega(k) \qquad (\text{for } \beta = \Theta(1)). \end{aligned}$$

□

Now we prove Theorem 1. The following observation, which easily follows from the rectangle property of communication protocols, is crucial in our proof.

OBSERVATION 1. *Conditioned on Π, we have that the random variables $Z_1, Z_2, \ldots, Z_k$ are independent.*

Let c_1 be a constant chosen later. We introduce the following definition.

DEFINITION 2. *(Goodness of a transcript) We say a transcript π is* bad$^+$ *if* $\mathsf{E}\left[\sum_{i\in[k]} Z_i \mid \Pi=\pi\right] \geq \beta k + c_1\sqrt{\beta k}$ *and* bad$^-$ *if* $\mathsf{E}\left[\sum_{i\in[k]} Z_i \mid \Pi=\pi\right] \leq \beta k - c_1\sqrt{\beta k}$. *In both cases we say π is* bad. *Otherwise we say it is* good.

We first show that a transcript is bad only with a small probability.

LEMMA 1. $\mathsf{Pr}_{\Pi\sim\tilde{\mu}}[\Pi$ *is bad*$] \leq 2e^{-(c_1-1)^2/3}/(1-e^{-1/3})$.

PROOF. Set $c_2 = c_1 - 1$. We say $Z = \{Z_1, Z_2, \ldots, Z_k\}$ is a *joker*$^+$ if $\sum_{i\in[k]} Z_i \geq \beta k + c_2\sqrt{\beta k}$, and a *joker*$^-$ if $\sum_{i\in[k]} Z_i \leq \beta k - c_2\sqrt{\beta k}$. In both cases we say Z is a *joker*.

First, we can apply a Chernoff bound on random variables Z_i for $i = 1, \ldots, k$, and so we have that

$$\mathsf{Pr}[Z \text{ is a joker}^+] = \mathsf{Pr}\left[\sum_{i\in[k]} Z_i \geq \beta k + c_2\sqrt{\beta k}\right] \leq e^{-c_2^2/3}.$$

Second, by Observation 1, we can apply a Chernoff bound on random variables Z_i for $i = 1, \ldots, k$ conditioned on Π being bad,

$$\begin{aligned} & \mathsf{Pr}[Z \text{ is a joker}^+ \mid \Pi \text{ is bad}^+] \\ \geq & \sum_{\pi} \mathsf{Pr}\left[\Pi=\pi \mid \pi \text{ is bad}^+\right] \mathsf{Pr}\left[Z \text{ is a joker}^+ \mid \Pi=\pi, \pi \text{ is bad}^+\right] \\ = & \sum_{\pi} \mathsf{Pr}\left[\Pi=\pi \mid \pi \text{ is bad}^+\right] \mathsf{Pr}\left[\sum_{i\in[k]} Z_i \geq \beta k + c_2\sqrt{\beta k} \,\middle|\, \mathsf{E}\left[\sum_{i\in[k]} Z_i \mid \Pi=\pi\right] \geq \beta k + c_1\sqrt{\beta k}, \Pi=\pi\right] \\ \geq & \sum_{\pi} \mathsf{Pr}\left[\Pi=\pi \mid \pi \text{ is bad}^+\right] \left(1 - e^{-(c_1-c_2)^2/3}\right) \\ = & \left(1 - e^{-(c_1-c_2)^2/3}\right). \end{aligned}$$

Finally by Bayes' theorem, we have that

$$\begin{aligned} \mathsf{Pr}[\Pi \text{ is bad}^+] &= \frac{\mathsf{Pr}[Z \text{ is a joker}^+] \cdot \mathsf{Pr}[\Pi \text{ is bad}^+ \mid Z \text{ is a joker}^+]}{\mathsf{Pr}[Z \text{ is a joker}^+ \mid \Pi \text{ is bad}^+]} \\ &\leq \frac{e^{-c_2^2/3}}{1 - e^{-(c_1-c_2)^2/3}}. \end{aligned}$$

Similarly, we can also show that

$$\mathsf{Pr}[\Pi \text{ is bad}^-] \leq e^{-c_2^2/3}/(1 - e^{-(c_1-c_2)^2/3}).$$

Therefore $\mathsf{Pr}[\Pi \text{ is bad}] \leq 2e^{-(c_1-1)^2/3}/(1-e^{-1/3})$ (recall that we set $c_2 = c_1 - 1$). □

Our next lemma indicates that if a transcript π is good and weak, then the sum of Z_i's will deviate from its mean considerably with a significant probability. Let c_3 be a constant chosen later.

LEMMA 2. *For a good and weak transcript π, there exists a universal constant $\tilde{c}$ such that*

$$\begin{aligned} & \mathsf{Pr}\left[\textstyle\sum_{i\in[k]} Z_i \leq \beta k - (c_3 - c_1)\sqrt{\beta k} \;\middle|\; \Pi=\pi\right] \\ \geq & \ \tilde{c}\cdot e^{-100(c_3+1)^2}, \\ \text{and} \quad & \mathsf{Pr}\left[\textstyle\sum_{i\in[k]} Z_i \geq \beta k + (c_3 - c_1)\sqrt{\beta k} \;\middle|\; \Pi=\pi\right] \\ \geq & \ \tilde{c}\cdot e^{-100(c_3+1)^2}. \end{aligned}$$

PROOF. We only need to prove the first inequality. The proof for the second inequality is the same.

Since π is weak, we can find a set $T \in [n]$ with $|T| = 0.5k$, such that for any $i \in T$ we have $H(Z_i \mid \Pi=\pi) \geq H_b(0.01\beta)$. Let $N_1 = \sum_{i\in T} Z_i$ and $N_2 = \sum_{i\in[k]\setminus T} Z_i$. Let c_4 and c_5 with $c_5 - c_4 = c_3$ be constants chosen later. The idea of the proof is to show that conditioned on $\Pi = \pi$, N_2 will concentrate around $\mathsf{E}[N_2 \mid \Pi=\pi]$ within $c_4\sqrt{\beta k}$ with a good probability, while N_1 will deviate from $\mathsf{E}[N_1 \mid \Pi=\pi]$ by at least $c_5\sqrt{\beta k}$ with a good probability, therefore $\sum_{i\in[k]} Z_i = N_1 + N_2$ will deviate from its mean by at least $(c_5 - c_4)\sqrt{\beta k} = c_3\sqrt{\beta k}$ with a good probability. Here we use the fact that N_1 and N_2 are independent random variables conditioned on $\Pi = \pi$.

To show that N_2 will concentrate around its mean, we use a Chernoff bound. Since π is good, we have by the definition of the goodness of a transcript that $\mathsf{E}[N_2 \mid \Pi=\pi] \leq \mathsf{E}[Z \mid \Pi=\pi] \leq \beta k + c_1\sqrt{\beta k} \leq 2\beta k$. Thus by a Chernoff bound,

$$\begin{aligned} & \mathsf{Pr}\left[N_2 - \mathsf{E}[N_2 \mid \Pi=\pi] \leq -c_4\sqrt{\beta k} \;\middle|\; \Pi=\pi\right] \\ & \qquad \leq e^{-\frac{c_4^2\beta k}{3\cdot 2\beta k}} = e^{-c_4^2/6}. \end{aligned} \quad (1)$$

To show that N_1 will deviate from its mean considerably, we prove an anti-concentration property of the distribution of N_1 conditioned on $\Pi = \pi$. We need the following result which is an easy consequence of Feller [26] (cf. [44]).

LEMMA 3. *([44]) Let Y be a sum of independent random variables, each attaining values in $[0,1]$, and let $\sigma = \sqrt{\mathsf{Var}[Y]} \geq 200$. Then for all $t \in [0, \sigma^2/100]$, we have*

$$\mathsf{Pr}[Y \geq \mathsf{E}[Y] + t] \geq c\cdot e^{-t^2/(3\sigma^2)}$$

for a universal constant $c > 0$.

Since for each $i \in T$ it holds that $H(Z_i \mid \Pi=\pi) \geq H_b(0.01\beta)$, we have $\mathsf{Var}(Z_i \mid \Pi=\pi) \geq 0.01\beta(1-0.01\beta) \geq 0.009\beta$. Since conditioned on $\Pi = \pi$, the Z_i's are independent, we have

$\mathsf{Var}(N_1 \mid \Pi = \pi) \geq 0.009\beta \cdot 0.5k \geq 0.004\beta k$. By Lemma 3 we have for some universal constant c,

$$\mathsf{Pr}\left[N_1 \geq \mathsf{E}[N_1 \mid \Pi = \pi] + c_5\sqrt{\beta k} \;\middle|\; \Pi = \pi\right] \geq c \cdot e^{-\frac{(c_5\sqrt{\beta k})^2}{3 \cdot 0.004\beta k}} \geq c \cdot e^{-100c_5^2}. \quad (2)$$

Set $c_4 = 1$ and $c_5 = c_3 + 1$. By (1) and (2) and the fact that π is good and weak, we obtain

$$\begin{aligned} & \mathsf{Pr}\left[\textstyle\sum_{i\in[k]} Z_i \geq \beta k + (c_3 - c_1)\sqrt{\beta k} \;\middle|\; \Pi = \pi\right] \\ \geq\ & \mathsf{Pr}\left[\textstyle\sum_{i\in[k]} Z_i - \mathsf{E}[\sum_{i\in[k]} Z_i \mid \Pi = \pi] \geq c_3\sqrt{\beta k} \;\middle|\; \Pi = \pi\right] \\ \geq\ & (1 - e^{-c_4^2/6}) \cdot c \cdot e^{-100c_5^2} \\ =\ & c \cdot (1 - e^{-1/6}) \cdot e^{-100(c_3+1)^2} \\ =\ & \tilde{c} \cdot e^{-100(c_3+1)^2}, \end{aligned}$$

where $\tilde{c}$ is a universal constant. □

Now we prove our main theorem for k-GAP-MAJ.

PROOF. (of Theorem 1) First, by Lemma 1 we know that with probability $\left(1 - 2e^{-(c_1-1)^2/3}/(1 - e^{-1/3})\right)$ a transcript π sampled according to $\tilde{\mu}$ is good. Second, conditioned on π being good, it cannot be weak with probability more than $1/2$. We show this by contradiction. Suppose that π is weak with probability at least $1/2$ conditioned on it being good. Set $c_3 - c_1 = 1$, $c_1 = 5$ and constant δ sufficiently small. By Lemma 2, we have that the error probability of the protocol will be at least

$$\left(1 - 2e^{-(c_1-1)^2/3}/(1 - e^{-1/3})\right) \cdot 1/2 \cdot \tilde{c} \cdot e^{-100(c_1+2)^2} > \delta,$$

violating the success guarantee of Theorem 1.

Therefore with probability at least

$$1/2 \cdot \left(1 - 2e^{-(c_1-1)^2/3}/(1 - e^{-1/3})\right) \geq \Omega(1),$$

π is both good and strong (thus strong). We are done. □

3.2 The 2-DISJ Problem

In 2-DISJ Alice and Bob each have an n-bit vector. If we view vectors as sets, then each of them has a subset of $[n]$ corresponding to the 1 bits. Let x be the set of Alice and y be the set of Bob. The goal is to return 1 if $x \cap y \neq \emptyset$, and 0 otherwise.

We define the input distribution τ_t as follows. Let $\ell = (n+1)/4$. With probability $1/t$, x and y are random subsets of $[n]$ such that $|x| = |y| = \ell$ and $|x \cap y| = 1$. And with probability $1 - 1/t$, x and y are random subsets of $[n]$ such that $|x| = |y| = \ell$ and $x \cap y = \emptyset$. Razborov [46] (see also [37]) proved that for $t = 4$, $D^{1/(400)}_{\tau_4}(\text{2-DISJ}) = \Omega(n)$. It is easy to extend this result to general t by the following claim.

CLAIM 1. *If a protocol $\mathcal{P}$ solves the problem for general t with error $1/(100t)$ and communication cost $o(n)$, then it also solves the problem when $t = 4$ with error $1/400$ and communication cost $o(n)$.*

PROOF. Under input distribution τ_t, let p be the probability that $\mathcal{P}$ succeeds conditioned on x and y intersecting, and q be the probability that $\mathcal{P}$ succeeds conditioned on x and y being disjoint. Then $p/t + q(1 - 1/t) \geq 1 - 1/(100t)$ by definition of τ_t. Notice that conditioned on x and y intersecting, or conditioned on x and y being disjoint, τ_t and τ_4 are equal as distributions. Hence, the success probability of the same protocol $\mathcal{P}$ on distribution τ_4 is $p/4 + 3q/4$. Substituting $p/t \geq 1 - 1/(100t) - q(1 - 1/t)$ into this, the success probability of $\mathcal{P}$ on τ_4 is at least $t/4 - 1/400 - tq/4 + q/4 + 3q/4 = t(1 - q)/4 - 1/400 + q$, and since $t \geq 4$, this is at least $1 - q - 1/400 + q = 399/400$, as desired. □

By Razborov's lower bound for τ_4, $D^{1/(100t)}_{\tau_t}(\text{2-DISJ}) = \Omega(n)$. In the rest of the paper we omit the subscript t in τ_t when there is no confusion.

3.3 The Complexity of F_0

We choose the input distribution ζ for the $(1 + \varepsilon)$-approximate F_0 problem as follows. Set $n = A/\varepsilon^2$ where $A = 20000/\delta$ is a constant, $\beta = 1/(k\varepsilon^2)$ and $t = 1/\beta$. We start with a set Y with cardinality $\ell = (n + 1)/4$ chosen uniformly at random from $[n]$, and then choose $X_1, X_2, \ldots, X_k$ according to the marginal distribution $\tau \mid Y$ independently, where τ is the hard input distribution for 2-DISJ. We assign $X_1, X_2, \ldots, X_k$ to the k sites, respectively.

Let $T_i = X_i \cap Y$ if $|X_i \cap Y| \neq \emptyset$ and NULL otherwise. Let $N = |\{i \in [k] \mid T_i \neq \text{NULL}\}|$. Let $R = F_0(T_1, T_2, \ldots, T_k)$. The following lemma shows that R will concentrate around its expectation $\mathsf{E}[R]$, which can be calculated exactly.

LEMMA 4. *With probability at least $(1 - 6500/A)$, we have $|R - \mathsf{E}[R]| \leq 1/(10\varepsilon)$, where $\mathsf{E}[R] = (1 - \lambda)N$ for some fixed constant $0 \leq \lambda \leq 4/A$.*

PROOF. We can think of our problem as a bin-ball game: those T_i ($i \in [k]$)'s that are not NULL are balls (thus we have N balls), and elements in the set Y are bins (thus we have ℓ bins). We throw each of the N balls into one of the ℓ bins uniformly at random. Our goal is to estimate the number of non-empty bins at the end of the process.

By a Chernoff bound we have that with probability at least $\left(1 - e^{-\Omega(\beta k)}\right) = 1 - o(1)$, it holds that $N \leq 2\beta k = 2/\varepsilon^2$. By Fact 1 and Lemma 1 in [39] we have $\mathsf{E}[R] = \ell\left(1 - (1 - 1/\ell)^N\right)$ and $\mathsf{Var}[R] < 4N^2/\ell$. Thus by Chebyshev's inequality we have

$$\mathsf{Pr}[|R - \mathsf{E}[R]| > 1/(10\varepsilon)] \leq \frac{\mathsf{Var}[R]}{1/(100\varepsilon^2)} \leq \frac{6400}{A}.$$

Let $\theta = N/\ell \leq 8/A$. We can write

$$\mathsf{E}[R] = \ell\left(1 - e^{-\theta}\right) + O(1) = \theta\ell\left(1 - \frac{\theta}{2!} + \frac{\theta^2}{3!} - \frac{\theta^3}{4!} + \right) + O(1).$$

This series converges and thus we can write $\mathsf{E}[R] = (1 - \lambda)\theta\ell = (1 - \lambda)N$ for some fixed constant $0 \leq \lambda \leq \theta/2 \leq 4/A$. □

The next lemma shows that we can use a protocol for F_0 to solve k-GAP-MAJ with good properties.

LEMMA 5. *If there exists a protocol $\mathcal{P}'$ that computes a $(1 + \alpha\varepsilon)$-approximation to F_0 (for some sufficiently small constant α) on input distribution ζ with error probability $\delta/2$, then there exists a protocol $\mathcal{P}$ that computes the k-GAP-MAJ problem on input distribution μ with error probability δ.*

PROOF. We first describe the construction of $\mathcal{P}$ using $\mathcal{P}'$ and then show its correctness.

Protocol $\mathcal{P}$. Given a random input $Z = \{Z_1, Z_2, \ldots, Z_k\}$ of k-GAP-MAJ chosen from distribution μ, we construct an input $(X_1, X_2, \ldots, X_k)$ of F_0 as follows: We first choose Y to be a subset of $[n]$ of size ℓ uniformly at random. Let $I_1^\ell, I_2^\ell, \ldots, I_k^\ell$ be random subsets of size ℓ from $[n] \setminus Y$, and $I_1^{\ell-1}, I_2^{\ell-1}, \ldots, I_k^{\ell-1}$

be random subsets of size $(\ell-1)$ from $[n]\setminus Y$. Let $I_1^1, I_2^1, \ldots, I_k^1$ be random elements from Y. We next choose

$$X_j\ (j=1,\ldots,k) = \begin{cases} I_j^\ell & \text{if } Z_j = 0, \\ I_j^{\ell-1} \cup I_j^1 & \text{if } Z_j = 1. \end{cases}$$

It is easy to see that $(X_1, X_2, \ldots, X_k, Y)$ is chosen from distribution ζ.

Protocol $\mathcal{P}$ first uses $\mathcal{P}'$ to compute $\tilde{W}$ which is a $(1+\alpha\varepsilon)$-approximation of $F_0(X_1, X_2, \ldots, X_k)$, and then determines the answer to k-GAP-MAJ as follows.

$$k\text{-GAP-MAJ}(Z_1,\ldots,Z_k) = \begin{cases} 1, & \text{if } \frac{\tilde{W}-(n-\ell)}{1-\lambda} > 1/\varepsilon^2 (= \beta k), \\ 0, & \text{otherwise.} \end{cases}$$

Recall that we set $n = A/\varepsilon^2, \ell = (n+1)/4$ and $0 \le \lambda \le 4/A$ is some fixed constant.

Correctness. Given a random input $(X_1, X_2, \ldots, X_k, Y)$ chosen from distribution ζ, the exact value of $W = F_0(X_1, X_2, \ldots, X_k)$ can be written as the sum of two components.

$$W = Q + R, \tag{3}$$

where Q is a random variable that counts $F_0(\cup_{i\in[k]} X_i \setminus Y)$, and R is a random variable that counts $F_0(\cup_{i\in[k]} X_i \bigcap Y)$. First, from our construction it is easy to see by Chernoff bounds and the union bound that with probability $\left(1 - 1/\varepsilon^2 \cdot e^{-\Omega(k)}\right) = 1 - o(1)$, we have $Q = |\{[n] - Y\}| = n - \ell$, since each element in $\{[n] - Y\}$ will be chosen by every X_i $(i = 1, 2, \ldots, k)$ with probability more than $1/4$. Second, by Lemma 4 we know that with probability $(1 - 6500/A)$, R is within $1/(10\varepsilon)$ from its mean $(1-\lambda)N$ for some fixed constant $0 \le \lambda \le 4/A$. Thus with probability $(1 - 6600/A)$, we can write Equation (3) as

$$W = (n-\ell) + (1-\lambda)N + \kappa_1, \tag{4}$$

for a value $|\kappa_1| \le 1/(10\varepsilon)$.

Since $F_0(X_1, X_2, \ldots, X_k)$ computes a value $\tilde{W}$ which is a $(1+\alpha\varepsilon)$-approximation of W, we can substitute W with $\tilde{W}$ in Equation (4), resulting in the following.

$$\tilde{W} = (n-\ell) + (1-\lambda)N + \kappa_1 + \kappa_2, \tag{5}$$

where $\kappa_2 \le \alpha\varepsilon \cdot F_0(X_1, X_2, \ldots, X_k) \le \alpha A/\varepsilon$. We can choose $\alpha = 1/(10A)$ to make $\kappa_2 \le 1/(10\varepsilon)$. Now we have

$$\begin{aligned} N &= (\tilde{W} - (n-\ell) - \kappa_1 - \kappa_2)/(1-\lambda) \\ &= (\tilde{W} - (n-\ell))/(1-\lambda) + \kappa_3, \end{aligned}$$

where $|\kappa_3| \le (1/(10\varepsilon) + 1/(10\varepsilon))/(1-4/A) \le 1/(4\varepsilon)$. Therefore $(\tilde{W} - (n-\ell))/(1-\lambda)$ estimates $N = \sum_{i\in[k]} Z_i$ correctly up to an additive error $1/(4\varepsilon) < \sqrt{\beta k} = 1/\varepsilon$, thus computes k-GAP-MAJ correctly. The total error probability of this simulation is at most $(\delta/2 + 6600/A)$, where the first term counts the error probability of $\mathcal{P}'$ and the second term counts the error probability introduced by the reduction. This is less than δ if we choose $A = 20000/\delta$. $\square$

From Theorem 1 we know that if a protocol computes k-GAP-MAJ$(Z_1, Z_2, \ldots, Z_k)$ correctly with error probability δ, then with probability $\Omega(1)$, for at least $0.5k$ Z_i's we have $H(Z_i \mid \Pi = \pi) \le H_b(0.01\beta)$. This is equivalent to the following: With probability $\Omega(1)$, the protocol has to solve at least $0.5k$ copies of 2-DISJ(X_i, Y) $(i \in [k])$ on input distribution τ each with error probability at most $0.01\beta = 1/(100t)$. By the lower bound for 2-DISJ on input distribution τ, solving each copy of 2-DISJ requires $\Omega(1/\varepsilon^2)$ bits of communication (recall that we set $n = A/\varepsilon^2$ for a constant A), thus in total we need $\Omega(k/\varepsilon^2)$ bits of communication.

THEOREM 2. *Any protocol that computes a $(1+\varepsilon)$-approximation to F_0 on input distribution ζ with error probability δ for some sufficiently small constant δ has communication complexity $\Omega(k/\varepsilon^2)$.*

4. A LOWER BOUND FOR F_P $(P > 1)$

We first introduce a problem called k-XOR which can be considered to some extent as a combination of two k-DISJ (introduced in [4, 8]) instances, and then compose it with 2-GAP-ORT (introduced in [47]) to create another problem that we call the k-BLOCK-THRESH-XOR (k-BTX) problem. We prove that the communication complexity of k-BTX is large. Finally, we prove a communication complexity lower bound for F_p by performing a reduction from k-BTX.

4.1 The 2-GAP-ORT Problem

In the 2-GAP-ORT problem we have two players Alice and Bob. Alice has a vector $x = \{x_1, x_2, \ldots, x_{1/\varepsilon^2}\} \in \{0,1\}^{1/\varepsilon^2}$ and Bob has a vector $y = \{y_1, y_2, \ldots, y_{1/\varepsilon^2}\} \in \{0,1\}^{1/\varepsilon^2}$. They want to compute

$$\text{2-GAP-ORT}(x,y) = \begin{cases} 1, & \left| \sum_{i\in[1/\varepsilon^2]} \text{XOR}(x_i, y_i) - \frac{1}{2\varepsilon^2} \right| \ge \frac{2}{\varepsilon}, \\ 0, & \left| \sum_{i\in[1/\varepsilon^2]} \text{XOR}(x_i, y_i) - \frac{1}{2\varepsilon^2} \right| \le \frac{1}{\varepsilon}, \\ *, & \text{otherwise.} \end{cases}$$

Let ϕ be the uniform distribution on $\{0,1\}^{1/\varepsilon^2} \times \{0,1\}^{1/\varepsilon^2}$ and let (X, Y) be a random input chosen from distribution ϕ.

We assume that the communication cost of all protocols in the paper is at most $\text{poly}(N)$, where N is the number of coordinates in the vector inputs to the parties. This assumption is fine for our purposes because we will show in Section 4.4 that a k-party protocol $\mathcal{P}$ for F_2 implies a 2-party protocol $\mathcal{P}'$ for 2-GAP-ORT with asymptotically the same communication. Thus if $\mathcal{P}'$ has communication cost larger than $\text{poly}(N)$, then we obtain a $\text{poly}(N)$ lower bound for the communication cost of F_2 immediately (for any $\text{poly}(N)$).

THEOREM 3. *Let Π be the transcript of any protocol for 2-GAP-ORT on input distribution ϕ with error probability ι, for a sufficiently small constant $\iota > 0$, and assume Π uses at most $\text{poly}(N)$ communication. Then, $I(X, Y; \Pi) \ge \tilde{\Omega}(1/\varepsilon^2)$.*

PROOF. Sherstov [47] proved that under the product uniform distribution ϕ, any protocol that computes 2-GAP-ORT correctly with error probability ι for some sufficiently small constant $\iota > 0$ has communication complexity $\Omega(1/\varepsilon^2)$. By Theorem 1.3 of Barak et al. [9] which says that under a product distribution, if the communication complexity of a two-player problem is at most $\text{poly}(t)$, then the information cost of the two-player game is at least the communication complexity of the two-player game up to a factor of $\text{poly}\log(t)$. That is, we have $I(X, Y; \Pi) \ge \tilde{\Omega}(1/\varepsilon^2)$. $\square$

4.2 The k-XOR Problem

In the k-XOR problem we have k sites $S_1, S_2, \ldots, S_k$. Each site S_i $(i = 1, 2, \ldots, k)$ holds a block $b_i = \{b_{i,1}, b_{i,2}, \ldots, b_{i,n}\}$ of n $(n \ge k^{1+\Omega(1)})$ bits. Let $b = (b_1, b_2, \ldots, b_k)$ be the list of the inputs of k sites. We assume $k \ge 4$ is a power of 2. The k sites want to compute the following function in the blackboard model.

$$k\text{-XOR}(b_1,\ldots,b_k) = \begin{cases} 1, & \text{if } \exists\, j \in [n] \text{ such that } b_{i,j} = 1 \\ & \quad \text{for exactly } k/2 \text{ } i\text{'s}, \\ 0, & \text{otherwise.} \end{cases}$$

We define the input distribution φ_n for the k-XOR problem as follows. For each coordinate ℓ ($\ell \in [n]$) there is a variable D_ℓ chosen uniformly at random from $\{1, 2, \ldots, k\}$. Conditioned on D_ℓ, all but the D_ℓ-th sites set their inputs to 0, whereas the D_ℓ-th site sets its input to 0 or 1 with equal probability. We call the D_ℓ-th site the special site in the ℓ-th coordinate. Let φ_1 denote this input distribution on one coordinate.

Next, we choose a random special coordinate $M \in [n]$ and replace the k sites' inputs on the M-th coordinate as follows: for the first $k/2$ sites, with probability $1/2$ we replace all $k/2$ sites' inputs with 0 and with probability $1/2$ we replace all $k/2$ sites' inputs with 1; and we independently perform the same operation to the second $k/2$ sites. Let ψ_1 denote the distribution on this special coordinate. And let ψ_n denote the input distribution that on the special coordinate M is distributed as ψ_1 and on each of the remaining $n-1$ coordinates is distributed as φ_1.

Let $B, B_i, B_{i,\ell}$ be the corresponding random variables of $b, b_i, b_{i,\ell}$ when the input of k-XOR is chosen according to the distribution ψ_n. Let $D = \{D_1, D_2, \ldots, D_n\}$. Let $X = 1$ if the inputs of the first $k/2$ sites in the special coordinate M are all 1 and $X = 0$ otherwise. Let $Y = 1$ if the inputs of the second $k/2$ sites in the special coordinate M are all 1 and $Y = 0$ otherwise. It is easy to see that under ψ_n we have k-XOR$(B) = X \oplus Y$. We say the instance B is a 00-instance if $X = Y = 0$, a 10-instance if $X = 1$ and $Y = 0$, a 01-instance if $X = 0$ and $Y = 1$, and a 11-instance if $X = Y = 1$. Let $S \in \{00, 01, 10, 11\}$ be the type of the instance.

THEOREM 4. *Let Π be the transcript of any protocol on input distribution ψ_n for which $I(X, Y; \Pi) = \tilde{\Omega}(1)$. Then we have $I(B; \Pi \mid M, D, S) = \tilde{\Omega}(n/k)$, where information is measured* [7] *with respect to the input distribution ψ_n.*

PROOF. Since $I(X, Y; \Pi) = \tilde{\Omega}(1)$, we have

$$\tilde{\Omega}(1) = I(X, Y; \Pi) = H(X, Y) - H(X, Y|\Pi) = 2 - H(X, Y|\Pi),$$

or $H(X, Y|\Pi) = 2 - \tilde{\Omega}(1)$. By the Maximum Likelihood Principle in Proposition 1, there is a reconstruction function g from the transcript of Π for which the error probability δ_g satisfies

$$\delta_g \leq 1 - \frac{1}{2^{H(X,Y|\Pi)}} \leq 1 - \frac{1}{2^{2-\tilde{\Omega}(1)}} = 1 - \frac{2^{\tilde{\Omega}(1)}}{4} = \frac{3}{4} - \tilde{\Omega}(1),$$

and therefore the success probability of the reconstruction function g over inputs X, Y is $\frac{1}{4} + \tilde{\Omega}(1)$. Since g is deterministic given the transcript Π, we abuse notation and say the success probability of Π is $\frac{1}{4} + \tilde{\Omega}(1)$.

For an $\ell \in [n]$, say ℓ is *good* if $\mathsf{Pr}[\Pi(B) = (X, Y)|M = \ell] = 1/4 + \tilde{\Omega}(1)$. By averaging, there are $\tilde{\Omega}(n)$ good ℓ.

By the chain rule, expanding the conditioning, and letting $D^{-\ell}$ denote the random variable D with ℓ-th component missing, and $B_{[k],<\ell}$ and $B_{[k],\ell}$ the inputs to the k sites on the first $\ell - 1$ coordinates and the ℓ-th coordinate, respectively, we have

$$\begin{aligned} I(B; \Pi \mid D, S, M) &= \sum_{\ell=1}^{n} I(B_{[k],\ell}; \Pi \mid D, S, M, B_{[k],<\ell}) \\ &\geq \sum_{\text{good } \ell} I(B_{[k],\ell}; \Pi \mid D, S, M, B_{[k],<\ell}), \end{aligned}$$

which is

$$\mathbf{E}_{b,d}\left[\sum_{\text{good } \ell} I(B_{[k],\ell}; \Pi \mid D_\ell, S, M, D^{-\ell} = d, B_{[k],<\ell} = b)\right].$$

Say a pair (b, d) is good for a good ℓ if

$$\mathsf{Pr}[\Pi(B) = (X, Y)|M = \ell, D^{-\ell} = d, B_{[k],<\ell} = b] = 1/4 + \tilde{\Omega}(1).$$

By a Markov argument,

$$\mathsf{Pr}[(b, d) \text{ is good }] = \tilde{\Omega}(1).$$

We therefore have that $I(B; \Pi \mid D, S, M)$ is at least

$$\tilde{\Omega}(1) \sum_{\text{good } \ell} I(B_{[k],\ell}; \Pi \mid D_\ell, S, M, D^{-\ell} = d, B_{[k],<\ell} = b, (b, d) \text{ is good}).$$

Now define a protocol $\Pi_{\ell,b,d}$ which on input $A_1, \ldots, A_k$ distributed according to ψ_1, attempts to output (U, V), where $U = 1$ if $A_1 = \cdots = A_{k/2} = 1$ and $U = 0$ otherwise, and $V = 1$ if $A_{k/2+1} = \cdots = A_k = 1$ and $V = 0$ otherwise. The protocol $\Pi_{\ell,b,d}$ has ℓ, b and d hardwired into it. It fills in the inputs for coordinates $\ell' > \ell$ by using the value d and the fact that the inputs to the parties are independent conditioned on $D^{-\ell} = d$. It fills in the inputs for coordinates $\ell' < \ell$ using the value b. This can all be done with no communication. Since ℓ is good and (b, d) is good for ℓ, it follows that $\mathsf{Pr}[\Pi_{\ell,b,d}(A_1, \ldots, A_k) = (U, V)] = \frac{1}{4} + \tilde{\Omega}(1)$.

Hence, for a good ℓ,

$$\begin{aligned} &I(B_{[k],\ell}; \Pi \mid D, S, M, B_{[k],<\ell}) \\ &= \tilde{\Omega}(1) \cdot I(A_1, \ldots, A_k; \Pi' \mid R, S, M), \end{aligned}$$

where Π' is a (randomized) protocol which succeeds in outputting (U, V) with probability $1/4 + \tilde{\Omega}(1)$ when $A_1, \ldots, A_k$ are distributed as in ψ_1, and $R \in [k]$ is chosen uniformly at random and independently of $S, M, A_1, \ldots, A_k$, and the private randomness of Π' (here R denotes the random variable D_ℓ in the reduction above). The information is measured with respect to the marginal distribution of ψ_n on a good coordinate ℓ. Observe that

$$\begin{aligned} &I(A_1, \ldots, A_k; \Pi' \mid R, M, S) \\ = \ &\frac{1}{4} \sum_{s \in \{00,01,10,11\}} I(A_1, \ldots, A_k; \Pi' \mid R, M, S = s), \end{aligned}$$

and so

$$\begin{aligned} &I(B_{[k],\ell}; \Pi \mid D, S, M, B_{[k],<\ell}) \\ &= \tilde{\Omega}(1) \cdot I(A_1, \ldots, A_k; \Pi' \mid R, M, S = 00). \end{aligned}$$

Let $\mathcal{E}$ be the event that all sites have the value 0 in the M-th coordinate when the inputs are drawn from φ_n. Observe that $(\varphi_n|\mathcal{E}) = (\psi_n|S = 00)$ as distributions, and so

$$\begin{aligned} &I(B_{[k],\ell}; \Pi \mid D, S, M, B_{[k],<\ell}) \\ &= \tilde{\Omega}(1) \cdot I(A_1, \ldots, A_k; \Pi' \mid R, M, \mathcal{E}), \end{aligned}$$

where the information on the left hand side is measured with respect to inputs B drawn from ψ_n, and the information on the right hand side is measured with respect to inputs $A_1, \ldots, A_k$ drawn from φ_n. Observe that $\mathsf{Pr}[M = \ell] = 1/n$, and so

$$\begin{aligned} &I(B_{[k],\ell}; \Pi \mid D, S, M, B_{[k],<\ell}) \\ \geq \ &\tilde{\Omega}(1) \cdot I(A_1, \ldots, A_k; \Pi' \mid R, M \neq \ell, \mathcal{E}) \cdot \frac{n-1}{n} \end{aligned}$$

[7] When we say that the information is measured with respect to a distribution α we mean that the inputs to the protocol are distributed according to α when computing the mutual information.

$$\geq \quad \tilde{\Omega}(1) \cdot I(A_1, \ldots, A_k; \Pi' \mid R, M \neq \ell, \mathcal{E}).$$

By definition of the mutual information, and using that $A_1, \ldots, A_k$ are independent of $\mathcal{E}$ given $M \neq \ell$,

$$\begin{aligned} & I(A_1, \ldots, A_k; \Pi' \mid R, M \neq \ell, \mathcal{E}) \\ = \quad & H(A_1, \ldots, A_k \mid R, M \neq \ell, \mathcal{E}) \\ & -H(A_1, \ldots, A_k \mid \Pi', R, M \neq \ell, \mathcal{E}) \\ \geq \quad & H(A_1, \ldots, A_k \mid R, M \neq \ell) - H(A_1, \ldots, A_k \mid \Pi', R, M \neq \ell) \\ = \quad & I(A_1, \ldots, A_k ; \Pi' \mid R, M \neq \ell). \end{aligned}$$

Notice that we have that $I(A_1, \ldots, A_k ; \Pi' \mid R, M \neq \ell)$ is equal to $I(A_1, \ldots, A_k; \Pi' \mid R)$ where the information is measured with respect to the input distribution φ_1, and Π' is a protocol which succeeds with probability $1/4 + \tilde{\Omega}(1)$ on ψ_1.

It remains to show that $I(A_1, \ldots, A_k; \Pi' \mid R) = \tilde{\Omega}(1/k)$ where the information is measured with respect to φ_1. Let $\mathbf{0}$ be the all-0 vector, $\mathbf{1}$ be the all-1 vector and $\mathbf{e_i}$ be the standard basis vector with the i-th coordinate being 1. By the relationship between mutual information and Hellinger distance (see Proposition 2.51 and Proposition 2.53 of [7]), we have

$$\begin{aligned} I(A_1, \ldots, A_k; \Pi' \mid R) \quad &= \quad (1/k) \sum_{i \in [k]} I(A_1, \ldots, A_k; \Pi' \mid R = i) \\ &= \quad \Omega(1/k) \textstyle\sum_{i \in [k]} h^2(\Pi'(\mathbf{0}), \Pi'(\mathbf{e_i})), \end{aligned}$$

where $h(\cdot, \cdot)$ is the Hellinger distance (see Section 2 for a definition). Now we assume k and $k/2$ are powers of 2, and we use Theorem 7 of [36], which says that the following three statements hold:

1. $\sum_{i \in [k]} h^2(\Pi'(\mathbf{0}), \Pi'(\mathbf{e_i})) = \Omega(1) \cdot h^2(\Pi'(\mathbf{0}), \Pi'(1^{k/2}0^{k/2}))$
2. $\sum_{i \in [k]} h^2(\Pi'(\mathbf{0}), \Pi'(\mathbf{e_i})) = \Omega(1) \cdot h^2(\Pi'(\mathbf{0}), \Pi'(0^{k/2}1^{k/2}))$
3. $\sum_{i \in [k]} h^2(\Pi'(\mathbf{0}), \Pi'(\mathbf{e_i})) = \Omega(1) \cdot h^2(\Pi'(\mathbf{0}), \Pi'(\mathbf{1}))$

It follows that

$$\begin{aligned} & I(A_1, \ldots, A_k; \Pi' \mid R) \\ = \quad & \Omega(1/k) \cdot \Big(h^2(\Pi'(\mathbf{0}), \Pi'(1^{k/2}0^{k/2})) \\ & + h^2(\Pi'(\mathbf{0}), \Pi'(0^{k/2}1^{k/2})) + h^2(\Pi'(\mathbf{0}), \Pi'(\mathbf{1})) \Big). \end{aligned}$$

By the Cauchy-Schwartz inequality we have,

$$\begin{aligned} & I(A_1, \ldots, A_k; \Pi' \mid R) \\ = \quad & \Omega(1/k) \cdot \Big(h(\Pi'(\mathbf{0}), \Pi'(1^{k/2}0^{k/2})) \\ & + h(\Pi'(\mathbf{0}), \Pi'(0^{k/2}1^{k/2})) + h(\Pi'(\mathbf{0}), \Pi'(\mathbf{1})) \Big)^2. \end{aligned}$$

We can rewrite this as

$$\begin{aligned} & I(A_1, \ldots, A_k; \Pi' \mid R) \\ = \quad & \Omega(1/k) \cdot \Big(3h(\Pi'(\mathbf{0}), \Pi'(1^{k/2}0^{k/2})) \\ & + 3h(\Pi'(\mathbf{0}), \Pi'(0^{k/2}1^{k/2})) + 3h(\Pi'(\mathbf{0}), \Pi'(\mathbf{1})) \Big)^2. \end{aligned}$$

Now by the triangle inequality of Hellinger distance (which is just the Euclidean norm of the so-called transcript wave function, see [36]), we obtain the following,

$$\begin{aligned} & I(A_1, \ldots, A_k; \Pi' \mid R) \\ = \quad & \Omega(1/k) \cdot \Big(\textstyle\sum_{a,b \in \{\mathbf{0},\ \mathbf{1},\ 1^{k/2}0^{k/2},\ 0^{k/2}1^{k/2}\}} h(\Pi'(a), \Pi'(b)) \Big)^2 \end{aligned}$$

The claim is that at least one of $h(\Pi'(a), \Pi'(b))$ in the RHS in Equation (6) is $\tilde{\Omega}(1)$, and this will complete the proof. By Proposition 3, this is true if the total variation distance between $\Pi'(a)$ and $\Pi'(b)$ is $\tilde{\Omega}(1)$ for an $a, b \in \{\mathbf{0},\ \mathbf{1},\ 1^{k/2}0^{k/2},\ 0^{k/2}1^{k/2}\}$, and there must be such a pair (a, b), as otherwise Π' cannot succeed with probability $1/4 + \tilde{\Omega}(1)$ on distribution ψ_1 (since it cannot distinguish different outputs), violating its success probability guarantee. $\square$

4.3 The k-BTX Problem

The input of the k-BTX problem is a concatenation of $1/\varepsilon^2$ copies of inputs of the k-XOR problem. That is, each site S_i ($i = 1, 2, \ldots, k$) holds an input consisting of $1/\varepsilon^2$ blocks each of which is an input for a site in the k-XOR problem. More precisely, each S_i ($i \in [k]$) holds an input $b_i = \{b_i^1, b_i^2, \ldots, b_i^{1/\varepsilon^2}\}$ where $b_i^j = \{b_{i,1}^j, b_{i,2}^j, \ldots, b_{i,n}^j\}$ ($j \in [1/\varepsilon^2]$) is a vector of n ($n > k^{1+\Omega(1)}$) bits. Let $b = \{b_1, b_2, \ldots, b_k\}$ be the union of the inputs of k sites. In the k-BTX problem the k sites want to compute the following.

$$k\text{-BTX}(b_1, \ldots, b_k) = \begin{cases} 1, & \text{if } \left| \sum\limits_{j \in [1/\varepsilon^2]} k\text{-XOR}(b_1^j, \ldots, b_k^j) - \frac{1}{2\varepsilon^2} \right| \\ & \geq 2/\varepsilon, \\ 0, & \text{if } \left| \sum\limits_{j \in [1/\varepsilon^2]} k\text{-XOR}(b_1^j, \ldots, b_k^j) - \frac{1}{2\varepsilon^2} \right| \\ & \leq 1/\varepsilon, \\ *, & \text{otherwise.} \end{cases}$$

We define the input distribution ν for the k-BTX problem as follows: the input of the k sites in each block is chosen independently according to the input distribution ψ_n, which is defined for the k-XOR problem. Let $B, B_i, B_i^j, B_{i,\ell}^j$ be the corresponding random variables of $b, b_i, b_i^j, b_{i,\ell}^j$ when the input of k-BTX is chosen according to the distribution ν. Let $D^j = \{D_1^j, D_2^j, \ldots, D_n^j\}$ where D_ℓ^j ($\ell \in [n], j \in [1/\varepsilon^2]$) is the special site in the ℓ-th coordinate of block j, and let $D = \{D^1, D^2, \ldots, D^{1/\varepsilon^2}\}$. Let $M = \{M^1, M^2, \ldots, M^{1/\varepsilon^2}\}$ where M^j is the special coordinate in block j. Let $S = \{S^1, S^2, \ldots, S^{1/\varepsilon^2}\}$ where $S^j \in \{00, 01, 10, 11\}$ is the type of the k-XOR instance in block j.

For each block j ($j \in [1/\varepsilon^2]$), let $X^j = 1$ if the inputs of the first $k/2$ sites in the special coordinate M^j are all 1 and $X^j = 0$ otherwise; and similarly let $Y^j = 1$ if the inputs of the second $k/2$ sites in the coordinate M^j are all 1 and $Y^j = 0$ otherwise. Let $X = \{X^1, X^2, \ldots, X^{1/\varepsilon^2}\}$ and $Y = \{Y^1, Y^2, \ldots, Y^{1/\varepsilon^2}\}$. We first show the following theorem.

THEOREM 5. *Let Π be the transcript of any protocol for k-BTX on input distribution ν with error probability δ for a sufficiently small constant $\delta > 0$. Then $I(X, Y; \Pi) = \tilde{\Omega}(1/\varepsilon^2)$, where the information is measured with respect to the uniform distribution on X, Y.*

PROOF. Consider the following randomized 2-player protocol Π' for 2-GAP-ORT, where the error probability is over both the coin tosses of Π' and the uniform distribution ϕ on inputs (X, Y). Alice and Bob run Π, with Alice controlling the first $k/2$ players, and Bob controlling the second $k/2$ players. Alice and Bob use the public coin to generate M^j and D^j values for each $j \in [1/\varepsilon^2]$. For each $j \in [1/\varepsilon^2]$, Alice sets the M^j-th coordinate of each of the first $k/2$ players to X_j. Similarly, Bob sets the M^j-th coordinate of each of the last $k/2$ players to Y_j. Alice and Bob then use private randomness and the D^j vectors to fill in the remaining coordinates. Observe that the resulting inputs are distributed according to ν for

k-BTX by definition of ν and the fact that (X, Y) is uniformly distributed.

Alice and Bob run the deterministic protocol Π. Every time a message is sent between any two of the k players in Π, it is appended to the transcript. That is, if the two players are among the first $k/2$, Alice still forwards this message to Bob. If the two players are among the last $k/2$, Bob still forwards this message to Alice. If the message is between a player in the first group and the second group, Alice and Bob exchange a message. The output of Π' is equal to that of Π. Let $rand$ denote the randomness used in Π', which since Π is deterministic, is just the randomness used to help create the inputs to Π. Note that $rand$ consists of the public randomness and private randomness. The only public randomness is that used to define the M^j and D^j values for each $j \in [1/\varepsilon^2]$. Let $\Pi'_{rand}(X, Y)$ denote the induced deterministic protocol we obtain by hardwiring $rand$.

By a Markov argument if Π succeeds with probability at least $1 - \delta$, then for at least a $1/2$ fraction of choices of $rand$,

$$\mathsf{Pr}_{X,Y}[\Pi'_{rand}(X, Y) = \text{2-GAP-ORT}(X, Y)] \geq 1 - 2\delta.$$

By construction of Π',

$$I(X, Y; \Pi(X, Y)) = I(X, Y; \Pi'(X, Y, rand)),$$

where $rand$ is *not* included in the transcript of Π'. By definition of the mutual information,

$$\begin{aligned} & I(X, Y; \Pi'(X, Y, rand)) \\ = \; & \mathbf{E}_{rand}\left[\mathbf{E}_{\Pi'_{rand}(X,Y)}\left[D_{KL}(p(X, Y|\Pi'_{rand}(X, Y)) \,||\, p(X, Y))\right]\right], \end{aligned}$$

where $D_{KL}(p, q)$ is the KL-divergence between distributions p and q, and $p(V)$ for a random variable V denotes its distribution. By a Markov argument, for at least a $2/3$ fraction of random strings $rand$,

$$\begin{aligned} & I(X, Y, ; \Pi'_{rand}(X, Y)) \\ = \; & \mathbf{E}_{\Pi'_{rand}(X,Y)}\left[D_{KL}(p(X, Y|\Pi'_{rand}(X, Y)) \,||\, p(X, Y)\right] \\ \leq \; & 3 \cdot I(X, Y; \Pi(X, Y)). \end{aligned}$$

By a union bound, there exists a setting of $rand$ for which we have

$$\mathsf{Pr}[\Pi'_{rand}(X, Y) = \text{2-GAP-ORT}(X, Y)] \geq 1 - 2\delta, \tag{6}$$

and

$$I(X, Y; \Pi'_{rand}(X, Y)) \leq 3I(X, Y; \Pi(X, Y)). \tag{7}$$

Since Π'_{rand} is deterministic, it follows by (6) and Theorem 3 that $I(X, Y; \Pi'_{rand}(X, Y)) = \tilde{\Omega}(1/\varepsilon^2)$, and hence by (7), we have $I(X, Y; \Pi(X, Y)) = \tilde{\Omega}(1/\varepsilon^2)$, which completes the proof. □

Now we are ready to prove our main theorem for k-BTX.

THEOREM 6. *Let Π be the transcript of any protocol for k-BTX on input distribution ν with error probability δ for a sufficiently small constant $\delta > 0$. We have $I(B; \Pi \mid M, D, S) \geq \tilde{\Omega}(n/(k\varepsilon^2))$ for any $n \geq k^{1+\Omega(1)}$, where the information is measured with respect to the input distribution ν.*

PROOF. By Theorem 5 we have $I(X, Y; \Pi) = \tilde{\Omega}(1/\varepsilon^2)$. Using the chain rule we obtain that

$$I(X^j, Y^j; \Pi \mid X^{<j}, Y^{<j}) = \tilde{\Omega}(1)$$

for at least $\tilde{\Omega}(1/\varepsilon^2)$ j's, where $X^{<j} = \{X^1, X^2, \ldots, X^{j-1}\}$ and similarly for $Y^{<j}$. We say such a j for which this holds is *good*.

Now we consider a good $j \in [1/\varepsilon^2]$. We show that

$$I(B^j; \Pi \mid M, D, S, B^{<j}) = \tilde{\Omega}(n/k)$$

if j is good. Since $B^{<j}$ determines $(X^{<j}, Y^{<j})$ and $B^{<j}$ is independent of B^j, by the third part of Proposition 1, it suffices to prove that $I(B^j; \Pi \mid M, D, S, X^{<j}, Y^{<j}) = \tilde{\Omega}(n/k)$. By expanding the conditioning, we can write $I(B^j; \Pi \mid M, D, S, X^{<j}, Y^{<j})$ as

$$\begin{aligned} \mathsf{E}_{m,d,s,x,y}[I(B^j; \Pi \mid M^j, D^j, S^j, M^{-j} = m, D^{-j} = d, \\ S^{-j} = s, X^{<j} = x, Y^{<j} = y)]. \end{aligned}$$

For each m, d, s, x, y, we define a randomized protocol $\Pi_{m,d,s,x,y}$ for computing X^j, Y^j on distribution ψ_n. Suppose the k sites are given inputs $a_1, a_2, \ldots, a_k$ chosen randomly according to ψ_n. For each $i \in [k]$ the i-th site sets $B^j_i = a_i$. The k sites set the remaining inputs as follows. Independently for each block $j' \neq j$, conditioned on $S^{j'}, M^{j'}$ and $D^{j'}$, the k sites sample the input $B^{j'}$ randomly and independently according to ψ_n, using their private random coins (note that S^{-j} determines $X^{<j}$ and $Y^{<j}$). Finally the k sites run Π on B and define

$$\Pi_{m,d,s,x,y}(a_1, \ldots, a_k) = \Pi(B).$$

By the definition of a good $j \in [1/\varepsilon^2]$, we know by a Markov bound that with probability $\tilde{\Omega}(1)$ over the choice of (x, y) from the uniform distribution, if $(X^{<j}, Y^{<j}) = (x, y)$ then we have

$$I(X^j, Y^j; \Pi \mid X^{<j} = x, Y^{<j} = y) = \tilde{\Omega}(1).$$

Call these (x, y) for which this holds *good*. Now for a good pair (x, y), we say a tuple (m, d, s) is *good* if

$$\begin{aligned} I(X^{<j}, Y^{<j}; \Pi \mid M^{-j} = m, D^{-j} = d, \\ S^{-j} = s, X^{<j} = x, Y^{<j} = y) = \tilde{\Omega}(1). \end{aligned}$$

Since $I(X^{<j}, Y^{<j}; \Pi \mid X^{<j} = x, Y^{<j} = y) = \tilde{\Omega}(1)$ for a good pair (x, y), by another Markov bound we have that

$$\mathsf{Pr}_{m,d,s}[(m, d, s) \text{ is good}] = \tilde{\Omega}(1).$$

Combining the above arguments with Theorem 4, we obtain

$$\begin{aligned} & I(B^j; \Pi \mid M, D, S, B^{<j}) \\ \geq \; & I(B^j; \Pi \mid M, D, S, X^{<j}, Y^{<j}) \\ = \; & \mathsf{E}_{m,d,s,x,y}[I(B^j; \Pi \mid M^j, D^j, S^j, M^{-j} = m, D^{-j} = d, \\ & S^{-j} = s, X^{<j} = x, Y^{<j} = y)] \\ \geq \; & \tilde{\Omega}(1) \cdot \mathsf{E}_{m,d,s,x,y}[I(B^j; \Pi \mid M^j, D^j, S^j, \\ & (M^{-j}, D^{-j}, S^{-j}) = (m, d, s), (X^{<j}, Y^{<j}) = (x, y), \\ & (m, d, s) \text{ is good}, (x, y) \text{ is good})] \\ = \; & \tilde{\Omega}(n/k) \quad \text{(By Theorem 4)}. \end{aligned}$$

By the chain rule, the fact that there are $\tilde{\Omega}(1/\varepsilon^2)$ good $j \in [1/\varepsilon^2]$, and part 3 of Proposition 1,

$$\begin{aligned} I(B; \Pi \mid M, D, S) & \geq \sum_{j \in [1/\varepsilon^2] \wedge \; j \text{ is good}} I(B^j; \Pi \mid M, D, S, B^{<j}) \\ & \geq \sum_{j \in [1/\varepsilon^2] \wedge \; j \text{ is good}} I(B^j; \Pi \mid M, D, S) \\ & \geq \tilde{\Omega}(n/(k\varepsilon^2)). \end{aligned}$$

This completes the proof. □

By Proposition 2 that says that the randomized communication complexity is always at least the conditional information cost, we have the following immediate corollary.

COROLLARY 2. *Any protocol that computes k-BTX on input distribution ν with error probability δ for some sufficient small constant δ has communication complexity $\tilde{\Omega}(n/(k\varepsilon^2))$.*

4.4 The Complexity of F_p $(p > 1)$

The input of ε-approximate F_p $(p > 1)$ is chosen to be the same as k-BTX by setting $n = k^p$. That is, we choose $\{b_1, b_2, \ldots, b_k\}$ randomly according to distribution ν. b_i is the input vector for site S_i consisting of $1/\varepsilon^2$ blocks each having $n = k^p$ coordinates. We prove the lower bound for F_p by performing a reduction from k-BTX.

LEMMA 6. *If there exists a protocol $\mathcal{P}'$ that computes a $(1+\alpha\varepsilon)$-approximate F_p $(p > 1)$ for a sufficiently small constant α on input distribution ν with communication complexity C and error probability at most δ, then there exists a protocol $\mathcal{P}$ for k-BTX on input distribution ν with communication complexity C and error probability at most $3\delta+\sigma$, where σ is an arbitrarily small constant.*

PROOF. We pick a random input $B = \{B_1, B_2, \ldots, B_k\}$ from distribution ν. Each coordinate (column) of B represents an item. Thus we have a total of $1/\varepsilon^2 \cdot k^p = k^p/\varepsilon^2$ possible items. If we view each input vector B_i $(i \in [k])$ as a set, then each site has a subset of $[k^p/\varepsilon^2]$ corresponding to these 1 bits. Let W_0 be the exact value of $F_p(B)$. W_0 can be written as the sum of four components:

$$W_0 = \left(\frac{k^p-1}{2\varepsilon^2}+Q\right)\cdot 1^p + \left(\frac{1}{2\varepsilon^2}+U\right)\cdot (k/2)^p + \left(\frac{1}{4\varepsilon^2}+V\right)\cdot k^p, \quad (8)$$

where Q, U, V are random variables (it will be clear why we write it this way in what follows). The first term of the RHS of Equation (8) is the contribution of non-special coordinates across all blocks in each of which one site has 1. The second term is the contribution of the special coordinates across all blocks in each of which $k/2$ sites have 1. The third term is the contribution of the special coordinates across all blocks in each of which all k sites have 1.

Note that k-BTX$(b_1, b_2, \ldots, b_k)$ is 1 if $|U| \geq 2/\varepsilon$ and 0 if $|U| \leq 1/\varepsilon$. Our goal is to use a protocol $\mathcal{P}'$ for F_p to construct a protocol $\mathcal{P}$ for k-BTX such that we can differentiate the two cases (i.e., $|U| \geq 2/\varepsilon$ or $|U| \leq 1/\varepsilon$) with a very good probability.

Given a random input B, let W_1 be the exact F_p-value on the first $k/2$ sites, and W_2 be the exact F_p-value on the second $k/2$ sites. That is, $W_1 = F_p(B_1, \ldots, B_{k/2})$ and $W_2 = F_p(B_{k/2+1}, \ldots, B_k)$. We have

$$W_1 + W_2 = \left(\frac{k^p-1}{2\varepsilon^2}+Q\right)\cdot 1^p + \left(\frac{1}{2\varepsilon^2}+U\right)\cdot (k/2)^p + \left(\frac{1}{4\varepsilon^2}+V\right)\cdot 2\cdot (k/2)^p. \quad (9)$$

By Equation (8) and (9) we can cancel out V:

$$2^{p-1}(W_1+W_2) - W_0 = (2^{p-1}-1)\left(\left(\frac{k^p-1}{2\varepsilon^2}+Q\right) + \left(\frac{1}{2\varepsilon^2}+U\right)\cdot (k/2)^p\right). \quad (10)$$

Let $\tilde{W}_0$, $\tilde{W}_1$ and $\tilde{W}_2$ be the estimated W_0, W_1 and W_2 obtained by running $\mathcal{P}'$ on the k sites' inputs, the first $k/2$ sites' inputs and the second $k/2$ sites' inputs, respectively. Observe that $W_0 \leq (2^p+1)k^p/\varepsilon^2$ and $W_1, W_2 \leq 2k^p/\varepsilon^2$. By properties of $\mathcal{P}'$ and the discussion above we have that with probability at least $1-3\delta$,

$$2^{p-1}(W_1+W_2) - W_0 = 2^{p-1}(\tilde{W}_1+\tilde{W}_2) - \tilde{W}_0 \pm \beta' k^p/\varepsilon, \quad (11)$$

where $|\beta'| \leq 3(2^p+1)\alpha$.

By a Chernoff bound we have that $|Q| \leq c_1 k^{p/2}/\varepsilon$ with probability at least $1-\sigma$, where σ is an arbitrarily small constant and $c_1 \leq \kappa \log^{1/2}(1/\sigma)$ for some universal constant κ. Combining this fact with Equation (10) and (11) and letting $\tilde{W} = (2^{p-1}(\tilde{W}_1 + \tilde{W}_2) - \tilde{W}_0)/(2^{p-1}-1)$, we have that with probability at least $1-3\delta-\sigma$,

$$U = \frac{2^p\tilde{W}}{k^p} - \frac{2^p+1}{2\varepsilon^2} - \frac{2^p\beta}{(2^{p-1}-1)\varepsilon}, \quad (12)$$

where $|\beta| \leq 3(2^p+1)\alpha + o(1)$.

Protocol $\mathcal{P}$. Given an input B for k-BTX, protocol $\mathcal{P}$ first uses $\mathcal{P}'$ to obtain the value $\tilde{W}$ described above, and then determines the answer to k-BTX as follows:

$$k\text{-BTX}(B) = \begin{cases} 1, & \text{if } \left|2^p\tilde{W}/k^p - (2^p+1)/(2\varepsilon^2)\right| \geq 1.5/\varepsilon, \\ 0, & \text{otherwise.} \end{cases}$$

Correctness. Note that with probability at least $1-3\delta-\sigma$, we have $|\beta| \leq 3(2^p+1)\alpha + o(1)$, where $\alpha > 0$ is a sufficiently small constant, and thus $\left|\frac{2^p\beta}{(2^{p-1}-1)\varepsilon}\right| < 0.5/\varepsilon$. Therefore, in this case protocol $\mathcal{P}$ will always succeed. □

Theorem 6 (set $n = k^p$) and Lemma 6 directly imply the following main theorem for F_p.

THEOREM 7. *Any protocol that computes a $(1+\varepsilon)$-approximate F_p $(p > 1)$ on input distribution ν with error probability δ for some sufficiently small constant δ has communication complexity $\tilde{\Omega}(k^{p-1}/\varepsilon^2)$.*

5. AN UPPER BOUND FOR F_P $(P > 1)$

We describe the following protocol to give a factor $(1+\Theta(\varepsilon))$-approximation to F_p at all points in time in the union of k streams each held by a different site. Each site has a non-negative vector $v^i \in \mathbb{R}^m$, [8] which evolves with time, and at all times the coordinator holds a $(1+\Theta(\varepsilon))$-approximation to $\|\sum_{i=1}^k v^i\|_p^p$. Let n be the length of the union of the k streams. We assume $n = \text{poly}(m)$, and that k is a power of 2.

As observed in [20], up to a factor of $O(\varepsilon^{-1}\log n \log(\varepsilon^{-1}\log n))$ in communication, the problem is equivalent to the threshold problem: given a threshold τ, with probability 2/3: when $\|\sum_{i=1}^k v^i\|_p^p > \tau$, the coordinator outputs 1, when $\|\sum_{i=1}^k v^i\|_p^p < \tau/(1+\varepsilon)$, the coordinator outputs 0, and for $\tau/(1+\varepsilon) \leq \|\sum_{i=1}^k v^i\|_p^p \leq \tau$, the coordinator can output either 0 or 1[9].

We can thus assume we are given a threshold τ in the following algorithm description. For notational convenience, define $\tau_\ell = \tau/2^\ell$ for an integer ℓ. A nice property of the algorithm is that it is one-way, namely, all communication is from the sites to the coordinator. We leave optimization of the $\text{poly}(\varepsilon^{-1}\log n)$ factors in the communication complexity to future work.

5.1 Our Protocol

The protocol consists of four algorithms illustrated in Algorithm 1 to Algorithm 4. Let $v = \sum_{i=1}^k v^i$ at any point in time during the

[8] We use m instead of N for universe size only in this section.

[9] To see the equivalence, by independent repetition, we can assume the success probability of the protocol for the threshold problem is $1-\Theta(\varepsilon/\log n)$. Then we can run a protocol for each $\tau = 1, (1+\varepsilon), (1+\varepsilon)^2, (1+\varepsilon)^3, \ldots, \Theta(n^2)$, and we are correct on all instantiations with probability at least 2/3.

Algorithm 1: Intepretation of the random public coin by sites and the coordinator

$r = \Theta(\log n)$ /* A parameter used by the sites and coordinator */
for $z = 1, 2, \ldots, r$ **do**
  for $\ell = 0, 1, 2, \ldots, \log m$ **do**
    Create a set S_ℓ^z by including each coordinate in $[m]$ independently with probability $2^{-\ell}$.

Algorithm 2: Initialization at Coordinator

$\gamma = \Theta(\varepsilon), B = \text{poly}(\varepsilon^{-1} \log n)$. Choose $\eta \in [0, 1]$ uniformly at random /* Parameters */
for $z = 1, 2, \ldots, r$ **do**
  for $\ell = 0, 1, 2, \ldots, \log m$ **do**
    for $j = 1, 2, \ldots, m$ **do**
      $f_{z,\ell,j} \leftarrow 0$ /* Initialize all frequencies seen to 0 */
$out \leftarrow 0$ /* The coordinator's current output */

Algorithm 3: When Site i receives an update $v^i \leftarrow v^i + e_j$ for standard unit vector e_j

for $z = 1, 2, \ldots, r$ **do**
  for $\ell = 0, 1, 2, \ldots, \log m$ **do**
    if $j \in S_\ell^z$ *and* $v_j^i > \tau_\ell^{1/p}/(kB)$ **then**
      With probability $\min(B/\tau_\ell^{1/p}, 1)$, send (j, z, ℓ) to the coordinator

Algorithm 4: Algorithm at Coordinator if a tuple (j, z, ℓ) arrives

$f_{z,\ell,j} \leftarrow f_{z,\ell,j} + \tau_\ell^{1/p}/B$
for $h = 0, 1, 2, \ldots, O(\gamma^{-1} \log(n/\eta^p))$ **do**
  for $z = 1, 2, \ldots, r$ **do**
    Choose ℓ for which $2^\ell \le \frac{\tau}{\eta^p (1+\gamma)^{ph} B} < 2^{\ell+1}$, or $\ell = 0$ if no such ℓ exists
    Let $F_{z,h} = \{j \in [m] \mid f_{z,\ell,j} \in [\eta(1+\gamma)^h, \eta(1+\gamma)^{h+1})\}$
  $\tilde{c}_h = \text{median}_z \; 2^\ell \cdot |F_{z,h}|$
if $\sum_{h \ge 0} \tilde{c}_h \cdot \eta^p \cdot (1+\gamma)^{ph} > (1-\varepsilon)\tau$ **then**
  $out \leftarrow 1$
  Terminate the protocol

union of the k streams. At times we will make the following assumptions on the algorithm parameters γ, B, and r: we assume $\gamma = \Theta(\varepsilon)$ is sufficiently small, and $B = \text{poly}(\varepsilon^{-1} \log n)$ and $r = \Theta(\log n)$ are sufficiently large.

5.2 Communication Cost

LEMMA 7. *Consider any setting of $v^1, \ldots, v^k$ for which we have $\| \sum_{i=1}^k v^i \|_p^p \le 2^p \cdot \tau$. Then the expected total communication is $k^{p-1} \cdot \text{poly}(\varepsilon^{-1} \log n)$ bits.*

PROOF. Fix any particular $z \in [r]$ and $\ell \in [0, 1, \ldots, \log m]$. Let $v_j^{i,\ell}$ equal v_j^i if $j \in S_\ell$ and equal 0 otherwise. Let $v^{i,\ell}$ be the vector with coordinates $v_j^{i,\ell}$ for $j \in [m]$. Also let $v^\ell = \sum_{i=1}^k v^{i,\ell}$. Observe that $\mathbf{E}[\|v^\ell\|_p^p] \le 2^p \cdot \tau / 2^\ell = 2^p \cdot \tau_\ell$.

Because of non-negativity of the v^i,

$$\sum_{i=1}^k \sum_{j \in S_\ell} (v_j^{i,\ell})^p \le \sum_{i=1}^k \|v^{i,\ell}\|_p^p \le \|v^\ell\|_p^p.$$

Notice that a $j \in S_\ell$ is sent by a site with probability at most $B/\tau_\ell^{1/p}$ and only if $(v_j^i)^p \ge \frac{\tau_\ell}{k^p B^p}$. Hence the expected number of messages sent for this z and ℓ, over all randomness, is

$$\frac{B}{\tau_\ell^{1/p}} \mathbf{E}\left[\sum_{i,j \mid (v_j^i)^p \ge \frac{\tau_\ell}{k^p B^p}} v_j^i\right] \le \frac{B}{\tau_\ell^{1/p}} \cdot \frac{\mathbf{E}[\|v^\ell\|_p^p]}{\tau_\ell/(k^p B^p)} \cdot \frac{\tau_\ell^{1/p}}{kB} \le \frac{2^p \cdot \tau_\ell \cdot k^{p-1} \cdot B^p}{\tau_\ell} = 2^p \cdot k^{p-1} \cdot B^p, \quad (13)$$

where we used that $\sum v_j^i$ is maximized subject to $(v_j^i)^p \ge \frac{\tau_\ell}{k^p B^p}$ and $\sum (v_j^i)^p \le \|v^\ell\|_p^p$ when all the v_j^i are equal to $\tau_\ell^{1/p}/(kB)$. Summing over all z and ℓ, it follows that the expected number of messages sent in total is $O(k^{p-1} B^p \log^2 n)$. Since each message is $O(\log n)$ bits, the expected number of bits is $k^{p-1} \cdot \text{poly}(\varepsilon^{-1} \log n)$. □

5.3 Correctness

We let $C > 0$ be a sufficiently large constant.

5.3.1 Concentration of Individual Frequencies

We shall make use of the following standard multiplicative Chernoff bound.

FACT 1. *Let $X_1, \ldots X_s$ be i.i.d. Bernoulli(q) random variables. Then for all $0 < \beta < 1$,*

$$\Pr\left[|\sum_{i=1}^s X_i - qs| \ge \beta qs\right] \le 2 \cdot e^{-\frac{\beta^2 qs}{3}}.$$

LEMMA 8. *For a sufficiently large constant $C > 0$, with probability $1 - n^{-\Omega(C)}$, for all z, ℓ, $j \in S_\ell$, and all times in the union of the k streams,*

1. *$f_{z,\ell,j} \le 2e \cdot v_j + \frac{C \tau_\ell^{1/p} \log n}{B}$, and*
2. *if $v_j \ge \frac{C(\log^5 n)\tau_\ell^{1/p}}{B\gamma^{10}}$, then*

$$|f_{z,\ell,j} - v_j| \le \frac{\gamma^5}{\log^2 n} \cdot v_j$$

.

PROOF. Fix a particular time snapshot in the stream. Let $g_{z,\ell,j} = f_{z,\ell,j} \cdot B/\tau_\ell^{1/p}$. Then $g_{z,\ell,j}$ is a sum of indicator variables, where the number of indicator variables depends on the values of the v_j^i. The indicator variables are independent, each with expectation $\min(B/\tau_\ell^{1/p}, 1)$.

First part of lemma. The number s of indicator variables is at most v_j, and the expectation of each is at most $B/\tau_\ell^{1/p}$. Hence, the probability that $w = 2e \cdot v_j \cdot B/\tau_\ell^{1/p} + C \log n$ or more of them equal 1 is at most

$$\binom{v_j}{w} \cdot \left(\frac{B}{\tau_\ell^{1/p}}\right)^w \le \left(\frac{ev_j B}{w\tau_\ell^{1/p}}\right)^w \le \left(\frac{1}{2}\right)^{C \log n} = n^{-C}.$$

This part of the lemma now follows by scaling the $g_{z,\ell,j}$ by $\tau_\ell^{1/p}/B$ to obtain a bound on the $f_{z,\ell,j}$.

Second part of lemma. Suppose at this time $v_j \geq \frac{C(\log^5 n)\tau_\ell^{1/p}}{B\gamma^{10}}$. The number s of indicator variables is minimized when there are $k-1$ distinct i for which $v_j^i = \frac{\tau_\ell^{1/p}}{kB}$, and one value of i for which

$$v_j^i = v_j - (k-1)\cdot\frac{\tau_\ell^{1/p}}{kB}.$$

Hence,

$$s \geq v_j - (k-1)\cdot\frac{\tau_\ell^{1/p}}{kB} - \frac{\tau_\ell^{1/p}}{kB} = v_j - \frac{\tau_\ell^{1/p}}{B}.$$

If the expectation is 1, then $f_{z,\ell,j} = v_j - \frac{\tau_\ell^{1/p}}{B}$, and using that $v_j \geq \frac{C(\log^5 n)\tau_\ell^{1/p}}{B\gamma^{10}}$ establishes this part of the lemma. Otherwise, applying Fact 1 with $s \geq v_j - \frac{\tau_\ell^{1/p}}{B} \geq \frac{C(\log^5 n)\tau_\ell^{1/p}}{2B\gamma^{10}}$ and $q = \frac{B}{\tau_\ell^{1/p}}$, and using that $qs \geq \frac{C\log^5 n}{2\gamma^{10}}$, we have

$$\mathsf{Pr}\left[|g_{z,\ell,j} - qs| > \frac{\gamma^5 qs}{2\log^2 n}\right] = n^{-\Omega(C)}.$$

Scaling by $\frac{\tau_\ell^{1/p}}{B} = \frac{1}{q}$, we have

$$\mathsf{Pr}\left[|f_{s,\ell,j} - s| > \frac{\gamma^5 s}{2\log^2 n}\right] = n^{-\Omega(C)},$$

and since $v_j - \frac{\tau_\ell^{1/p}}{B} \leq s \leq v_j$,

$$\mathsf{Pr}\left[|f_{s,\ell,j} - v_j| \geq \frac{\gamma^5 v_j}{2\log^2 n} + \frac{\tau_\ell^{1/p}}{B}\right] = n^{-\Omega(C)},$$

and finally using that $\frac{\tau_\ell^{1/p}}{B} < \frac{\gamma^5 v_j}{2\log^2 n}$, and union-bounding over a stream of length n as well as all choices of z, ℓ, and j, the lemma follows. □

5.3.2 Estimating Class Sizes

Define the classes C_h as follows:

$$C_h = \{j \in [m] \mid \eta(1+\gamma)^h \leq v_j < \eta(1+\gamma)^{h+1}\}.$$

Say that C_h contributes at a point in time in the union of the k streams if

$$|C_h|\cdot\eta^p(1+\gamma)^{ph} \geq \frac{\gamma\|v\|_p^p}{B^{1/2}\log(n/\eta^p)}.$$

Since the number of non-zero $|C_h|$ is $O(\gamma^{-1}\log(n/\eta^p))$, we have

$$\sum_{\text{non-contributing } h} |C_h|\cdot\eta^p(1+\gamma)^{ph+p} = O\left(\frac{\|v\|_p^p}{B^{1/2}}\right). \quad (14)$$

LEMMA 9. *With probability $1 - n^{-\Omega(C)}$, at all points in time in the union of the k streams and for all h and ℓ, for at least a $3/5$ fraction of the $z \in [r]$,*

$$|C_h \cap S_\ell^z| \leq 3\cdot 2^{-\ell}\cdot|C_h|$$

PROOF. The random variable $|C_h \cap S_\ell^z|$ is a sum of $|C_h|$ independent Bernoulli$(2^{-\ell})$ random variables. By a Markov bound, $\mathsf{Pr}[|C_h \cap S_\ell^z| \leq 3\cdot 2^{-\ell}|C_h|] \geq 2/3$. Letting X_z be an indicator variable which is 1 iff $|C_h \cap S_\ell^z| \leq 3\cdot 2^{-\ell}|C_h|$, the lemma follows by applying Fact 1 to the X_z, using that r is large enough, and union-bounding over a stream of length n and all h and ℓ. □

For a given C_h, let $\ell(h)$ be the value of ℓ for which we have $2^\ell \leq \frac{\tau}{\eta^p(1+\gamma)^{ph}B} < 2^{\ell+1}$, or $\ell = 0$ if no such ℓ exists.

LEMMA 10. *With probability $1 - n^{-\Omega(C)}$, at all points in time in the union of the k streams and for all h, for at least a $3/5$ fraction of the $z \in [r]$,*

1. *$2^{\ell(h)}\cdot|C_h \cap S_{\ell(h)}^z| \leq 3|C_h|$, and*
2. *if at this time C_h contributes and $\|v\|_p^p \geq \frac{\tau}{5}$, then $2^{\ell(h)}\cdot|C_h \cap S_{\ell(h)}^z| = (1\pm\gamma)|C_h|$.*

PROOF. We show this statement for a fixed h and at a particular point in time in the union of the k streams. The lemma will follow by a union bound.

The first part of the lemma follows from Lemma 9.

We now prove the second part. In this case $\|v\|_p^p \geq \frac{\tau}{5}$. We can assume that there exists an ℓ for which $2^\ell \leq \frac{\tau}{\eta^p(1+\gamma)^{ph}B} < 2^{\ell+1}$. Indeed, otherwise $\ell(h) = 0$ and $|C_h \cap S_{\ell(h)}^z| = |C_h|$ and the second part of the lemma follows.

Let $q(z) = |C_h \cap S_{\ell(h)}^z|$, which is a sum of independent indicator random variables and so $\mathbf{Var}[q(z)] \leq \mathbf{E}[q(z)]$. Also,

$$\mathbf{E}[q(z)] \quad = \quad 2^{-\ell}|C_h| \geq \frac{\eta^p(1+\gamma)^{ph}B}{\tau}\cdot|C_h|. \quad (15)$$

Since C_h contributes, $|C_h|\cdot\eta^p\cdot(1+\gamma)^{ph} \geq \frac{\gamma\|v\|_p^p}{B^{1/2}\log(n/\eta^p)}$, and combining this with (15),

$$\mathbf{E}[q(z)] \geq \frac{B\gamma\|v\|_p^p}{B^{1/2}\tau\log(n/\eta^p)} \geq \frac{B^{1/2}\gamma}{5\log(n/\eta^p)}.$$

It follows that for B sufficiently large, and assuming $\eta \geq 1/n^C$ which happens with probability $1 - 1/n^C$, we have $\mathbf{E}[q(z)] \geq \frac{3}{\gamma^2}$, and so by Chebyshev's inequality,

$$\mathsf{Pr}\left[|q(z) - \mathbf{E}[q(z)]| \geq \gamma\mathbf{E}[q(z)]\right] \leq \frac{\mathbf{Var}[q(z)]}{\gamma^2\cdot\mathbf{E}^2[q(z)]} \leq \frac{1}{3}.$$

Since $\mathbf{E}[q(z)] = 2^{-\ell}|C_h|$, and $r = \Theta(\log n)$ is large enough, the lemma follows by a Chernoff bound. □

5.3.3 Combining Individual Frequency Estimation and Class Size Estimation

We define the set T to be the set of times in the input stream for which the F_p-value of the union of the k streams first exceeds $(1+\gamma)^i$ for an i satisfying

$$0 \leq i \leq \log_{(1+\gamma)} 2^p\cdot\tau.$$

LEMMA 11. *With probability $1 - O(\gamma)$, for all times in T and all h,*

1. *$\tilde{c}_h \leq 3|C_h| + 3\gamma(2+\gamma)(|C_{h-1}| + |C_{h+1}|)$, and*
2. *if at this time C_h contributes and $\|v\|_p^p \geq \frac{\tau}{5}$, then*

$$(1-4\gamma)|C_h| \leq \tilde{c}_h \leq (1+\gamma)|C_h| + 3\gamma(2+\gamma)(|C_{h-1}| + |C_{h+1}|).$$

PROOF. We assume the events of Lemma 8 and Lemma 10 occur, and we add $n^{-\Omega(C)}$ to the error probability. Let us fix a class C_h, a point in time in T, and a $z \in [r]$ which is among the at least $3r/5$ different z that satisfy Lemma 10 at this point in time.

By Lemma 8, for any $j \in C_h \cap S_{\ell(h)}^z$ for which $v_j \geq \frac{C(\log^5 n)\tau_{\ell(h)}^{1/p}}{B\gamma^{10}}$, if

$$|\min(v_j - \eta(1+\gamma)^h, \eta(1+\gamma)^{h+1} - v_j)| \geq \frac{\gamma^5}{\log^2 n}\cdot v_j, \quad (16)$$

then $j \in F_{z,h}$. Let us first verify that for $j \in C_h$, we have $v_j \geq \frac{C(\log^5 n)\tau_{\ell(h)}^{1/p}}{B\gamma^{10}}$. We have

$$v_j^p \geq \eta^p(1+\gamma)^{ph} \geq \frac{\tau}{2^{\ell(h)+1}B} \geq \frac{\tau_{\ell(h)}}{2B}, \tag{17}$$

and so

$$v_j \geq \left(\frac{\tau_{\ell(h)}}{2B}\right)^{1/p} \geq \frac{C(\log^5 n)\tau_{\ell(h)}^{1/p}}{B\gamma^{10}},$$

where the final inequality follows for large enough $B = \mathrm{poly}(\varepsilon^{-1}\log n)$ and $p > 1$.

It remains to consider the case when (16) does not hold.

Conditioned on all other randomness, $\eta \in [0,1]$ is uniformly random subject to $v_j \in C_h$, or equivalently,

$$\frac{v_j}{(1+\gamma)^{h+1}} < \eta \leq \frac{v_j}{(1+\gamma)^h}.$$

If (16) does not hold, then either

$$\frac{(1-\gamma^5/\log^2 n)v_j}{(1+\gamma)^h} \leq \eta, \text{ or } \eta \leq \frac{(1+\gamma^5/\log^2 n)v_j}{(1+\gamma)^{h+1}}.$$

Hence, the probability over η that inequality (16) holds is at least

$$1 - \frac{\frac{\gamma^5 v_j}{(1+\gamma)^h \log^2 n} + \frac{\gamma^5 v_j}{(1+\gamma)^{h+1}\log^2 n}}{\frac{v_j}{(1+\gamma)^h} - \frac{v_j}{(1+\gamma)^{h+1}}} = 1 - \frac{\gamma^4(2+\gamma)}{\log^2 n}.$$

It follows by a Markov bound that

$$\mathsf{Pr}\left[|C_h \cap S^z_{\ell(h)}| \geq |C_h| \cdot (1-\gamma(2+\gamma))\right] \leq \frac{\gamma^3}{\log^2 n}. \tag{18}$$

Now we must consider the case that there is a $j' \in C_{h'} \cap S^z_{\ell(h)}$ for which $j' \in F_{z,h}$ for an $h' \neq h$. There are two cases, namely, if $v_{j'} < \frac{C(\log^5 n)\tau_{\ell(h)}^{1/p}}{B\gamma^{10}}$ or if $v_{j'} \geq \frac{C(\log^5 n)\tau_{\ell(h)}^{1/p}}{B\gamma^{10}}$. We handle each case in turn.

Case: $v_{j'} < \frac{C(\log^5 n)\tau_{\ell(h)}^{1/p}}{B\gamma^{10}}$**.** Then by Lemma 8,

$$f_{z,\ell(h),j'} \leq 2e \cdot v_{j'} + \frac{C\tau_{\ell(h)}^{1/p}\log n}{B}.$$

Therefore, it suffices to show that

$$2e \cdot \frac{C(\log^5 n)\tau_{\ell(h)}^{1/p}}{B\gamma^{10}} + \frac{C\tau_{\ell(h)}^{1/p}\log n}{B} < \eta(1+\gamma)^h,$$

from which we can conclude that $j' \notin F_{z,h}$. But by (17),

$$\eta(1+\gamma)^h \geq \left(\frac{\tau_{\ell(h)}}{2B}\right)^{1/p} > 2e \cdot \frac{C(\log^5 n)\tau_{\ell(h)}^{1/p}}{B\gamma^{10}} + \frac{C\tau_{\ell(h)}^{1/p}\log n}{B},$$

where the last inequality follows for large enough $B = \mathrm{poly}(\varepsilon^{-1}\log n)$. Hence, $j' \notin F_{z,h}$.

Case: $v_{j'} \geq \frac{C(\log^5 n)\tau_{\ell(h)}^{1/p}}{B\gamma^{10}}$**.** We claim that $h' \in \{h-1, h+1\}$. Indeed, by Lemma 8 we must have

$$\eta(1+\gamma)^h - \frac{\gamma^5}{\log^2 n} \cdot v_{j'} \leq v_{j'} \leq \eta(1+\gamma)^{h+1} + \frac{\gamma^5}{\log^2 n} \cdot v_{j'}.$$

This is equivalent to

$$\frac{\eta(1+\gamma)^h}{1+\gamma^5/\log^2 n} \leq v_{j'} \leq \frac{\eta(1+\gamma)^{h+1}}{1-\gamma^5/\log^2 n},$$

If $j' \in C_{h'}$ for $h' < h-1$, then

$$v_{j'} \leq \eta(1+\gamma)^{h-1} = \frac{\eta(1+\gamma)^h}{1+\gamma} < \frac{\eta(1+\gamma)^h}{1+\gamma^5/\log^2 n},$$

which is impossible. Also, if $j' \in C_{h'}$ for $h' > h+1$, then

$$v_{j'} \geq \eta(1+\gamma)^{h+2} = \eta(1+\gamma)^{h+1} \cdot (1+\gamma) > \frac{\eta(1+\gamma)^{h+1}}{1-\gamma^5/\log^2 n},$$

which is impossible. Hence, $h' \in \{h-1, h+1\}$.

Let $N_{z,h} = F_{z,h} \setminus C_h$. Then

$$\mathbf{E}[|N_{z,h}| \leq \frac{\gamma^4(2+\gamma)}{\log^2 n} \cdot (|C_{h-1} \cap S^z_{\ell(h)}| + |C_{h+1} \cap S^z_{\ell(h)}|). \tag{19}$$

By (18) and applying a Markov bound to (19), together with a union bound, with probability $\geq 1 - \frac{2\gamma^3}{\log^2 n}$,

$$(1-\gamma(2+\gamma)) \cdot |C_h \cap S^z_{\ell(h)}| \leq |F_{z,h}| \tag{20}$$

$$|F_{z,h}| \leq |C_h \cap S^z_{\ell(h)}| + \gamma(2+\gamma) \cdot (|C_{h-1} \cap S^z_{\ell(h)}| + |C_{h+1} \cap S^z_{\ell(h)}|). \tag{21}$$

By Lemma 9,

$$2^{\ell(h)}|C_{h-1} \cap S^z_{\ell(h)}| \leq 3|C_{h-1}| \text{ and } 2^{\ell(h)}|C_{h+1} \cap S^z_{\ell(h)}| \leq 3|C_{h+1}|. \tag{22}$$

First part of lemma. At this point we can prove the first part of this lemma. By the first part of Lemma 10,

$$2^{\ell(h)} \cdot |C_h \cap S^z_{\ell(h)}| \leq 3|C_h|. \tag{23}$$

Combining (21), (22), and (23), we have with probability at least $1 - \frac{2\gamma^3}{\log^2 n} - n^{-\Omega(C)}$,

$$2^{\ell(h)}|F_{z,h}| \leq 3|C_h| + 3\gamma(2+\gamma)(|C_{h-1}| + |C_{h+1}|).$$

Since this holds for at least $3r/5$ different z, it follows that

$$\tilde{c}_h \leq 3|C_h| + 3\gamma(2+\gamma)(|C_{h-1}| + |C_{h+1}|),$$

and the first part of the lemma follows by a union bound. Indeed, the number of h is $O(\gamma^{-1}\log(n/\eta^p))$, which with probability $1 - 1/n$, say, is $O(\gamma^{-1}\log n)$ since with this probability $\eta^p \geq 1/n^p$. Also, $|T| = O(\gamma^{-1}\log n)$. Hence, the probability this holds for all h and all times in T is $1 - O(\gamma)$.

Second part of the lemma. Now we can prove the second part of the lemma. By the second part of Lemma 10, if at this time C_h contributes and $\|v\|_p^p \geq \frac{\tau}{5}$, then

$$2^{\ell(h)} \cdot |C_h \cap S^z_{\ell(h)}| = (1 \pm \gamma)|C_h|. \tag{24}$$

Combining (20), (21), (22), and (24), we have with probability at least $1 - \frac{2\gamma^3}{\log^2 n} - n^{-\Omega(C)}$,

$$(1-\gamma(2+\gamma))(1-\gamma)|C_h| \leq 2^{\ell(h)}|F_{z,h}| \leq (1+\gamma)|C_h| + 3\gamma(2+\gamma)(|C_{h-1}| + |C_{h+1}|).$$

Since this holds for at least $3r/5$ different z, it follows that

$$(1-\gamma(2+\gamma))(1-\gamma)|C_h| \leq \tilde{c}_h \leq (1+\gamma)|C_h| + 3\gamma(2+\gamma)(|C_{h-1}| + |C_{h+1}|).$$

and the second part of the lemma now follows by a union bound over all h and all times in T, exactly in the same way as the first part of the lemma. Note that $1-4\gamma \leq (1-\gamma(2+\gamma))(1-\gamma)$ for small enough $\gamma = \Theta(\varepsilon)$. $\square$

5.3.4 Putting It All Together

LEMMA 12. *With probability at least $5/6$, at all times the coordinator's output is correct.*

PROOF. The coordinator outputs 0 up until the first point in time in the union of the k streams for which $\sum_{h\geq 0} \tilde{c}_h \cdot \eta^p \cdot (1+\gamma)^{ph} > (1-\varepsilon/2)\tau$. It suffices to show that

$$\sum_{h\geq 0} \tilde{c}_h \eta^p (1+\gamma)^{ph} = (1 \pm \varepsilon/2)\|v\|_p^p \tag{25}$$

at all times in the stream. We first show that with probability at least $5/6$, for all times in T,

$$\sum_{h\geq 0} \tilde{c}_h \eta^p (1+\gamma)^{ph} = (1 \pm \varepsilon/4)\|v\|_p^p, \tag{26}$$

and then use the structure of T and the protocol to argue that (25) holds at all times in the stream.

Fix a particular time in T. We condition on the event of Lemma 11, which by setting $\gamma = \Theta(\varepsilon)$ small enough, can assume occurs with probability at least $5/6$.

First, suppose at this point in time we have $\|v\|_p^p < \frac{\tau}{5}$. Then by Lemma 11, for sufficiently small $\gamma = \Theta(\varepsilon)$, we have

$$\begin{aligned}
&\sum_{h\geq 0} \tilde{c}_h \cdot \eta^p (1+\gamma)^{ph} \\
\leq\ & \sum_{h\geq 0} (3|C_h| + 3\gamma(2+\gamma)(|C_{h-1}| + |C_{h+1}|)) \cdot \eta^p (1+\gamma)^{ph} \\
\leq\ & \sum_{h\geq 0} \left(3 \sum_{j\in C_h} v_j^p + 3\gamma(2+\gamma)(1+\gamma)^2 \sum_{j \in C_{h-1}\cup C_{h+1}} v_j^p \right) \\
\leq\ & 4\|v\|_p^p \\
\leq\ & \frac{4\tau}{5},
\end{aligned}$$

and so the coordinator will correctly output 0, provided $\varepsilon < \frac{1}{5}$.

We now handle the case $\|v\|_p^p \geq \frac{\tau}{5}$. Then for all contributing C_h, we have

$$(1-4\gamma)|C_h| \leq \tilde{c}_h \leq (1+\gamma)|C_h| + 3\gamma(2+\gamma)(|C_{h-1}| + |C_{h+1}|),$$

while for all C_h, we have

$$\tilde{c}_h \leq 3|C_h| + 3\gamma(2+\gamma)(|C_{h-1}| + |C_{h+1}|).$$

Hence, using (14),

$$\begin{aligned}
\sum_{h\geq 0} \tilde{c}_h \cdot \eta^p (1+\gamma)^{ph} &\geq \sum_{\text{contributing } C_h} (1-4\gamma)|C_h| \eta^p (1+\gamma)^{ph} \\
&\geq \frac{(1-4\gamma)}{(1+\gamma)^2} \sum_{\text{contributing } C_h} \sum_{j\in C_h} v_j^p \\
&\geq (1-6\gamma)\cdot(1-O(1/B^{1/2}))\cdot \|v\|_p^p.
\end{aligned}$$

For the other direction,

$$\begin{aligned}
&\sum_{h\geq 0} \tilde{c}_h \cdot \eta^p (1+\gamma)^{ph} \\
\leq\ & \sum_{\text{contributing } C_h} (1+\gamma)|C_h| \eta^p (1+\gamma)^{ph} \\
&+ \sum_{\text{non-contributing } C_h} 3|C_h| \eta^p (1+\gamma)^{ph} \\
&+ \sum_{h\geq 0} 3\gamma(2+\gamma)(|C_{h-1}| + |C_{h+1}|)\eta^p (1+\gamma)^{ph} \\
\leq\ & (1+\gamma) \sum_{\text{contributing } C_h} \sum_{j\in C_h} v_j^p + O(1/B^{1/2})\cdot \|v\|_p^p + O(\gamma)\cdot \|v\|_p^p \\
\leq\ & (1 + O(\gamma) + O(1/B^{1/2}))\|v\|_p^p.
\end{aligned}$$

Hence, (26) follows for all times in T provided that $\gamma = \Theta(\varepsilon)$ is small enough and $B = \text{poly}(\varepsilon^{-1}\log n)$ is large enough.

It remains to argue that (25) holds for all points in time in the union of the k streams. Recall that each time in the union of the k streams for which $\|v\|_p^p \geq (1+\gamma)^i$ for an integer i is included in T, provided $\|v\|_p^p \leq 2^p\tau$.

The key observation is that the quantity $\sum_{h\geq 0} \tilde{c}_h \eta^p (1+\gamma)^{ph}$ is non-decreasing, since the values $|F_{z,h}|$ are non-decreasing. Now, the value of $\|v\|_p^p$ at a time t not in T is, by definition of T, within a factor of $(1\pm\gamma)$ of the value of $\|v\|_p^p$ for some time in T. Since (26) holds for all times in T, it follows that the value of $\sum_{h\geq 0} \tilde{c}_h \eta^p (1+\gamma)^{ph}$ at time t satisfies

$$(1-\gamma)(1-\varepsilon/4)\|v\|_p^p \leq \sum_{h\geq 0} \tilde{c}_h \eta^p (1+\gamma)^{ph} \leq (1+\gamma)(1+\varepsilon/4)\|v\|_p^p,$$

which implies for $\gamma = \Theta(\varepsilon)$ small enough that (25) holds for all points in time in the union of the k streams. This completes the proof. $\square$

THEOREM 8. *(MAIN) With probability at least $2/3$, at all times the coordinator's output is correct and the total communication is $k^{p-1}\cdot \text{poly}(\varepsilon^{-1}\log n)$ bits.*

PROOF. Consider the setting of $v^1, \ldots, v^k$ at the first time in the stream for which $\|\sum_{i=1}^k v^i\|_p^p > \tau$. For any non-negative integer vector w and any update e_j, we have $\|w+e_j\|_p^p \leq (\|w\|_p + 1)^p \leq 2^p\|w\|_p^p$. Since $\|\sum_{i=1}^k v^i\|_p^p$ is an integer and $\tau \geq 1$, we therefore have $\|\sum_{i=1}^k v^i\|_p^p \leq 2^p \cdot \tau$. By Lemma 7, the expected communication for these $v^1, \ldots, v^k$ is $k^{p-1}\cdot\text{poly}(\varepsilon^{-1}\log n)$ bits, so with probability at least $5/6$ the communication is $k^{p-1}\cdot\text{poly}(\varepsilon^{-1}\log n)$ bits. By Lemma 12, with probability at least $5/6$, the protocol terminates at or before the time for which the inputs held by the players equal $v^1, \ldots, v^k$. The theorem follows by a union bound. $\square$

6. RELATED PROBLEMS

In this section we show that the techniques we have developed for distributed F_0 and F_p ($p > 1$) can also be used to solve other fundamental problems. In particular, we consider the problems: all-quantile, heavy hitters, empirical entropy and ℓ_p for any $p > 0$. For the first three problems, we are able to show that our lower bounds holds even if we allow some additive error ε. From definitions below one can observe that lower bounds for additive ε-approximations also hold for their multiplicative $(1+\varepsilon)$-approximation counterparts.

6.1 The All-Quantile and Heavy Hitters

We first give the definitions of the problems. Given a set $A = \{a_1, a_2, \ldots, a_m\}$ where each a_i is drawn from the universe $[N]$, let f_i be the frequency of item i in the set A. Thus $\sum_{i\in[N]} f_i = m$.

DEFINITION 3. *(ϕ-heavy hitters) For any $0 \leq \phi \leq 1$, the set of ϕ-heavy hitters of A is $H_\phi(A) = \{x \mid f_x \geq \phi m\}$. If an ε-approximation is allowed, then the returned set of heavy hitters must contain $H_\phi(A)$ and cannot include any x such that $f_x <$*

$(\phi-\varepsilon)m$. If $(\phi-\varepsilon)m \le f_x < \phi m$, then x may or may not be included in $H_\phi(A)$.

DEFINITION 4. *(ϕ-quantile) For any $0 \le \phi \le 1$, the ϕ-quantile of A is some x such that there are at most ϕm items of A that are smaller than x and at most $(1-\phi)m$ items of A that are greater than x. If an ε-approximation is allowed, then when asking for the ϕ-quantile of A we are allowed to return any ϕ'-quantile of A such that $\phi - \varepsilon \le \phi' \le \phi + \varepsilon$.*

DEFINITION 5. *(All-quantile) The ε-approximate all-quantile (QUAN) problem is defined in the coordinator model, where we have k sites and a coordinator. Site S_i ($i \in [k]$) has a set A_i of items. The k sites want to communicate with the coordinator so that at the end of the process the coordinator can construct a data structure from which all ε-approximate ϕ-quantile for any $0 \le \phi \le 1$ can be extracted. The cost is defined as the total number of bits exchanged between the coordinator and the k sites.*

THEOREM 9. *Any protocol that computes ε-approximate QUAN or ε-approximate $\min\{\frac{1}{2}, \frac{\varepsilon\sqrt{k}}{2}\}$-heavy hitters with error probability δ for some sufficiently small constant δ has communication complexity $\Omega(\min\{\sqrt{k}/\varepsilon, 1/\varepsilon^2\})$ bits.*

PROOF. We first prove the theorem for QUAN. In the case that $k \ge 1/\varepsilon^2$, we prove an $\Omega(1/\varepsilon^2)$ lower bound. We prove this by a simple reduction from k-GAP-MAJ. We can assume $k = 1/\varepsilon^2$ since if $k > 1/\varepsilon^2$ then we can just give inputs to the first $1/\varepsilon^2$ sites. Set $\beta = 1/2$. Given a random input $Z_1, Z_2, \ldots, Z_k$ of k-GAP-MAJ chosen from distribution μ, we simply give the site S_i with Z_i for the first $1 \le i \le k$ sites. It is easy to observe that a protocol that computes $\varepsilon/2$-approximate QUAN on $A = \{Z_1, Z_2, \ldots, Z_k\}$ with error probability δ also computes k-GAP-MAJ on input distribution μ with error probability δ, since the answer to k-GAP-MAJ is simply the answer to $\frac{1}{2}$-quantile. The $\Omega(1/\varepsilon^2)$ lower bound follows from Corollary 1.

In the case that $k < 1/\varepsilon^2$, we prove an $\Omega(\sqrt{k}/\varepsilon)$ lower bound. We again perform a reduction from k-GAP-MAJ. Set $\beta = 1/2$. The reduction works as follows. We are given $\ell = 1/(\varepsilon\sqrt{k})$ independent copies of k-GAP-MAJ with $Z^1, Z^2, \ldots, Z^\ell$ being the inputs, where $Z^i = \{Z_1^i, Z_2^i, \ldots, Z_k^i\} \in \{0,1\}^k$ is chosen from distribution μ. We construct an input for QUAN by giving the j-th site the item set $A_j = \{Z_j^1, 2+Z_j^2, 4+Z_j^3, \ldots, 2(\ell-1)+Z_j^\ell\}$. It is not difficult to observe that a protocol that computes $\varepsilon/2$-approximate QUAN on the set $A = \{A_1, A_2, \ldots, A_j\}$ with error probability δ also computes the answer to each copy of k-GAP-MAJ on distribution μ with error probability δ, simply by returning $(X_i - 2(i-1))$ for the i-th copy of k-GAP-MAJ, where X_i is the $\varepsilon/2$-approximate $\frac{i-1/2}{\ell}$-quantile.

On the other hand, any protocol that computes each of the ℓ independent copies of k-GAP-MAJ correctly with error probability δ for a sufficiently small constant δ has communication complexity $\Omega(\sqrt{k}/\varepsilon)$. This is simply because for any transcript Π, by Corollary 1, independence and the chain rule we have that

$$I(Z^1, Z^2, \ldots, Z^\ell; \Pi) \ge \sum_{i \in [\ell]} I(Z^i; \Pi) \ge \Omega(\ell k) \ge \Omega(\sqrt{k}/\varepsilon). \tag{27}$$

By our reduction the theorem follows.

The proof for heavy hitters is done by essentially the same reduction as that for QUAN. In the case that $k = 1/\varepsilon^2$ (or $k \ge 1/\varepsilon^2$ in general), a protocol that computes $\varepsilon/2$-approximate $\frac{1}{2}$-heavy hitters on $A = \{Z_1, Z_2, \ldots, Z_k\}$ with error probability δ also computes k-GAP-MAJ on input distribution μ with error probability δ. In the case that $k < 1/\varepsilon^2$, it also holds that a protocol that computes $\varepsilon/2$-approximate $\frac{\varepsilon\sqrt{k}}{2}$-heavy hitters on the set $A = \{A_1, A_2, \ldots, A_j\}$ where $A_j = \{Z_j^1, 2 + Z_j^2, 4 + Z_j^3, \ldots, 2(\ell-1) + Z_j^\ell\}$ with error probability δ also computes the answer to each copy of k-GAP-MAJ on distribution μ with error probability δ. □

6.2 Entropy Estimation

We are given a set $A = \{(e_1, a_1), (e_2, a_2), \ldots, (e_m, a_m)\}$ where each e_k ($k \in [m]$) is drawn from the universe $[N]$, and $a_k \in \{+1, -1\}$ denotes an insertion or a deletion of item e_k. The entropy estimation problem (ENTROPY) asks for the value $H(A) = \sum_{j \in [N]} (|f_j|/L) \log(L/|f_j|)$ where $f_j = \sum_{k:e_k=j} a_k$ and $L = \sum_{j \in [N]} |f_j|$. In the ε-approximate ENTROPY problem, the items in the set A are distributed among k sites who want to compute a value $\tilde{H}(A)$ for which $\left|\tilde{H}(A) - H(A)\right| \le \varepsilon$. In this section we prove the following theorem.

THEOREM 10. *There exists an input distribution such that any protocol that computes ε-approximate ENTROPY on this distribution correctly with error probability at most δ for some sufficiently small constant δ has communication complexity $\tilde{\Omega}(k/\varepsilon^2)$.*

PROOF. Due to space constraints, we refer the reader to the full version of this paper [51] for the proof. □

6.3 ℓ_p for any constant $p \ge 1$

Consider an n-dimensional vector x with integer entries. It is well-known that for a vector v of n i.i.d. $N(0,1)$ random variables that $\langle v, x\rangle \sim N(0, \|x\|_2^2)$. Hence, for any real $p > 0$, $\mathbf{E}[|\langle v, x\rangle|^p] = \|x\|_2^p G_p$, where $G_p > 0$ is the p-th moment of the standard half-normal distribution (see [1] for a formula for these moments in terms of confluent hypergeometric functions). Let $r = O(\varepsilon^{-2})$, and $v^1, \ldots, v^r$ be independent n-dimensional vectors of i.i.d. $N(0,1)$ random variables. Let $y_j = \langle v^j, x\rangle / G_p^{1/p}$, so that $y = (y_1, \ldots, y_r)$. By Chebyshev's inequality for $r = O(\varepsilon^{-2})$ sufficiently large, $\|y\|_p^p = (1 \pm \varepsilon/3)\|x\|_2^p$ with probability at least $1 - c$ for an arbitrarily small constant $c > 0$.

We thus have the following reduction which shows that estimating ℓ_p up to a $(1+\varepsilon)$-factor requires communication complexity $\tilde{\Omega}(k/\varepsilon^2)$ for any $p > 0$. Let the k parties have respective inputs $x^1, \ldots, x^k$, and let $x = \sum_{i=1}^k x^i$. The parties use the shared randomness to choose shared vectors $v^1, \ldots, v^r$ as described above. For $i = 1, \ldots, k$ and $j = 1, \ldots, r$, let $y_j^i = \langle v^j, x^i\rangle / G_p^{1/p}$, so that $y^i = (y_1^i, \ldots, y_r^i)$. Let $y = \sum_{i=1}^k y^i$. By the above, $\|y\|_p^p = (1 \pm \varepsilon/3)\|x\|_2^p$ with probability at least $1 - c$ for an arbitrarily small constant $c > 0$. We note that the entries of the v^i can be discretized to $O(\log n)$ bits, changing the p-norm of y by only a $(1 \pm O(1/n))$ factor, which we ignore.

Hence, given a randomized protocol for estimating $\|y\|_p^p$ up to a $(1+\varepsilon/3)$ factor with probability $1 - \delta$, and given that the parties have respective inputs $y^1, \ldots, y^k$, this implies a randomized protocol for estimating $\|x\|_2^p$ up to a $(1 \pm \varepsilon/3) \cdot (1 \pm \varepsilon/3) = (1 \pm \varepsilon)$ factor with probability at least $1 - \delta - c$, and hence a protocol for estimating ℓ_2 up to a $(1 \pm \varepsilon)$ factor with this probability. The communication complexity of the protocol for ℓ_2 is the same as that for ℓ_p. By our communication lower bound for estimating ℓ_2 (in fact, for estimating F_2 in which all coordinates of x are non-negative), this implies the following theorem.

THEOREM 11. *The randomized communication complexity of approximating the ℓ_p-norm, $p \ge 1$, up to a factor of $1+\varepsilon$ with constant probability, is $\tilde{\Omega}(k/\varepsilon^2)$.*

Acknowledgements

We would like to thank Elad Verbin for many helpful discussions, in particular, for helping us with the F_0 lower bound, which was discovered in joint conversations with him. We also thank Amit Chakrabarti and Oded Regev for helpful discussions, as well as the anonymous referees for useful comments. Finally, we thank the organizers of the Synergies in Lower Bounds workshop that took place in Aarhus for bringing the authors together.

7. REFERENCES

[1] http://en.wikipedia.org/wiki/Normal_distribution.

[2] Open problems in data streams and related topics. http://www.cse.iitk.ac.in/users/sganguly/data-stream-probs.pdf, 2006.

[3] Open problems in data streams, property testing, and related topics. http://people.csail.mit.edu/konak/download/publications/bertinoro_and_kanpur_open_problems.pdf, 2011.

[4] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *Proc. ACM Symposium on Theory of Computing*, 1996.

[5] C. Arackaparambil, J. Brody, and A. Chakrabarti. Functional monitoring without monotonicity. In *Proc. International Colloquium on Automata, Languages, and Programming*, 2009.

[6] B. Babcock and C. Olston. Distributed top-k monitoring. In *Proc. ACM SIGMOD International Conference on Management of Data*, 2003.

[7] Z. Bar-Yossef. *The complexity of massive data set computations*. PhD thesis, University of California at Berkeley, 2002.

[8] Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *J. Comput. Syst. Sci.*, 68:702–732, June 2004.

[9] B. Barak, M. Braverman, X. Chen, and A. Rao. How to compress interactive communication. In *Proc. ACM Symposium on Theory of Computing*, pages 67–76, 2010.

[10] J. Brody and A. Chakrabarti. A multi-round communication lower bound for gap hamming and some consequences. In *IEEE Conference on Computational Complexity*, pages 358–368, 2009.

[11] J. Brody, A. Chakrabarti, O. Regev, T. Vidick, and R. de Wolf. Better gap-hamming lower bounds via better round elimination. In *APPROX-RANDOM*, pages 476–489, 2010.

[12] A. Chakrabarti, G. Cormode, R. Kondapally, and A. McGregor. Information cost tradeoffs for augmented index and streaming language recognition. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 387–396, 2010.

[13] A. Chakrabarti, G. Cormode, and A. McGregor. Robust lower bounds for communication and stream computation. In *Proc. ACM Symposium on Theory of Computing*, pages 641–650, 2008.

[14] A. Chakrabarti, T. S. Jayram, and M. Patrascu. Tight lower bounds for selection in randomly ordered streams. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, pages 720–729, 2008.

[15] A. Chakrabarti, S. Khot, and X. Sun. Near-optimal lower bounds on the multi-party communication complexity of set disjointness. In *Proc. IEEE Conference on Computational Complexity*, pages 107–117, 2003.

[16] A. Chakrabarti and O. Regev. An optimal lower bound on the communication complexity of gap-hamming-distance. In *Proc. ACM Symposium on Theory of Computing*, 2011.

[17] A. Chakrabarti, Y. Shi, A. Wirth, and A. Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 270–278, 2001.

[18] G. Cormode and M. Garofalakis. Sketching streams through the net: Distributed approximate query tracking. In *Proc. International Conference on Very Large Data Bases*, 2005.

[19] G. Cormode, M. Garofalakis, S. Muthukrishnan, and R. Rastogi. Holistic aggregates in a networked world: Distributed tracking of approximate quantiles. In *Proc. ACM SIGMOD International Conference on Management of Data*, 2005.

[20] G. Cormode, S. Muthukrishnan, and K. Yi. Algorithms for distributed functional monitoring. *ACM Transactions on Algorithms*, 7(2):21, 2011.

[21] G. Cormode, S. Muthukrishnan, K. Yi, and Q. Zhang. Optimal sampling from distributed streams. In *Proc. ACM Symposium on Principles of Database Systems*, 2010. Invited to Journal of the ACM.

[22] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc., 1991.

[23] P. Duris and J. D. P. Rolim. Lower bounds on the multiparty communication complexity. *J. Comput. Syst. Sci.*, 56(1):90–95, 1998.

[24] F. Ergün and H. Jowhari. On distance to monotonicity and longest increasing subsequence of a data stream. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, pages 730–736, 2008.

[25] D. Estrin, R. Govindan, J. S. Heidemann, and S. Kumar. Next century challenges: Scalable coordination in sensor networks. In *MOBICOM*, pages 263–270, 1999.

[26] W. Feller. Generalization of a probability limit theorem of cramer. *Trans. Amer. Math. Soc*, 54(3):361–372, 1943.

[27] A. Gál and P. Gopalan. Lower bounds on streaming algorithms for approximating the length of the longest increasing subsequence. In *Proc. IEEE Symposium on Foundations of Computer Science*, 2007.

[28] S. Ganguly. Polynomial estimators for high frequency moments. *CoRR*, abs/1104.4552, 2011.

[29] S. Ganguly. A lower bound for estimating high moments of a data stream. *CoRR*, abs/1201.0253, 2012.

[30] A. Gronemeier. Asymptotically optimal lower bounds on the nih-multi-party information complexity of the and-function and disjointness. In *Symposium on Theoretical Aspects of Computer Science*, pages 505–516, 2009.

[31] S. Guha and Z. Huang. Revisiting the direct sum theorem and space lower bounds in random order streams. In *Proc. International Colloquium on Automata, Languages, and Programming*, 2009.

[32] N. J. A. Harvey, J. Nelson, and K. Onak. Sketching and streaming entropy via approximation theory. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 489–498, 2008.

[33] Z. Huang, K. Yi, and Q. Zhang. Randomized algorithms for tracking distributed count, frequencies, and ranks. *CoRR*, abs/1108.3413, 2011.

[34] P. Indyk and D. Woodruff. Optimal approximations of the frequency moments of data streams. In *Proc. ACM Symposium on Theory of Computing*, 2005.

[35] P. Indyk and D. P. Woodruff. Tight lower bounds for the distinct elements problem. In *FOCS*, pages 283–288, 2003.

[36] T. S. Jayram. Hellinger strikes back: A note on the multi-party information complexity of and. In *APPROX-RANDOM*, pages 562–573, 2009.

[37] B. Kalyanasundaram and G. Schintger. The probabilistic communication complexity of set intersection. *SIAM Journal on Discrete Mathematics*, 5:545–557, 1992.

[38] D. M. Kane, J. Nelson, E. Porat, and D. P. Woodruff. Fast moment estimation in data streams in optimal space. In *STOC*, pages 745–754, 2011.

[39] D. M. Kane, J. Nelson, and D. P. Woodruff. An optimal algorithm for the distinct elements problem. In *Proc. ACM Symposium on Principles of Database Systems*, pages 41–52, 2010.

[40] R. Keralapura, G. Cormode, and J. Ramamirtham. Communication-efficient distributed monitoring of thresholded counts. In *Proc. ACM SIGMOD International Conference on Management of Data*, 2006.

[41] E. Kushilevitz and N. Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[42] F. Magniez, C. Mathieu, and A. Nayak. Recognizing well-parenthesized expressions in the streaming model. In *Proc. ACM Symposium on Theory of Computing*, pages 261–270, 2010.

[43] A. Manjhi, V. Shkapenyuk, K. Dhamdhere, and C. Olston. Finding (recently) frequent items in distributed data streams. In *Proc. IEEE International Conference on Data Engineering*, 2005.

[44] J. Matousek and J. Vondrák. *The probabilistic method.* Lecture Notes, 2008.

[45] J. M. Phillips, E. Verbin, and Q. Zhang. Lower bounds for number-in-hand multiparty communication complexity, made easy. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 2012.

[46] A. A. Razborov. On the distributional complexity of disjointness. In *Proc. International Colloquium on Automata, Languages, and Programming*, 1990.

[47] A. A. Sherstov. The communication complexity of gap hamming distance. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:63, 2011.

[48] S. Tirthapura and D. P. Woodruff. Optimal random sampling from distributed streams revisited. In *The International Symposium on Distributed Computing*, pages 283–297, 2011.

[49] T. Vidick. A concentration inequality for the overlap of a vector on a large set, with application to the communication complexity of the gap-hamming-distance problem. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:51, 2011.

[50] D. Woodruff. Optimal space lower bounds for all frequency moments. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 2004.

[51] D. Woodruff and Q. Zhang. Tight Bounds for Distributed Functional Monitoring. In `http://arxiv.org/abs/1112.5153`, 2011.

[52] A. C.-C. Yao. Probabilistic computations: Toward a unified measure of complexity (extended abstract). In *FOCS*, 1977.

[53] K. Yi and Q. Zhang. Optimal tracking of distributed heavy hitters and quantiles. In *Proc. ACM Symposium on Principles of Database Systems*, 2009.