Robust Lower Bounds for Communication and Stream Computation

Amit Chakrabarti*
Dartmouth College
ac@cs.dartmouth.edu

Graham Cormode
AT&T Labs–Research
graham@research.att.com

Andrew McGregor
UC San Diego
andrewm@ucsd.edu

ABSTRACT

We study the communication complexity of evaluating functions when the input data is randomly allocated (according to some known distribution) amongst two or more players, possibly with information overlap. This naturally extends previously studied variable partition models such as the best-case and worst-case partition models [32, 29]. We aim to understand whether the hardness of a communication problem holds for almost every allocation of the input, as opposed to holding for perhaps just a few atypical partitions.

A key application is to the heavily studied data stream model. There is a strong connection between our communication lower bounds and lower bounds in the data stream model that are "robust" to the ordering of the data. That is, we prove lower bounds for when the order of the items in the stream is chosen not adversarially but rather uniformly (or near-uniformly) from the set of all permuations. This random-order data stream model has attracted recent interest, since lower bounds here give stronger evidence for the inherent hardness of streaming problems.

Our results include the first random-partition communication lower bounds for problems including multi-party set disjointness and gap-Hamming-distance. Both are tight. We also extend and improve previous results [19, 7] for a form of pointer jumping that is relevant to the problem of selection (in particular, median finding). Collectively, these results yield lower bounds for a variety of problems in the random-order data stream model, including estimating the number of distinct elements, approximating frequency moments, and quantile estimation.

Categories and Subject Descriptors:
F.2.2**[Theory of Computation]**: ANALYSIS OF ALGORITHMS AND PROBLEM COMPLEXITY

General Terms: Theory

Keywords: Communication Complexity, Lower Bounds, Data Streams

*Work supported by an NSF CAREER award and by Dartmouth College startup funds

STOC'08, May 17–20, 2008, Victoria, British Columbia, Canada.

1. INTRODUCTION

Since its introduction in 1979 by Yao, communication complexity [37, 28] has proven to be a powerful technique for proving lower bounds in a variety of settings, including the cell-probe and data stream models, circuit and decision tree complexity and VLSI design. The majority of results in this area involve a fixed-partition model of communication complexity, where the goal is for two or more players to evaluate a function of an input that has been partitioned between them in a particular way, e.g., computing $f(x, y)$ when one player holds x and the other has y. Many functions can be shown to require a large amount of communication to evaluate when the input is partitioned between the players in this manner. These can imply lower bounds for various models of computation, via arguments that such partitions necessarily arise in the course of the computation.

To a lesser extent, variable-partition models, such as best-case and worst-case partition, have also been studied: see, e.g., [2, 29, 32] and [28, Chap. 7] for a survey. For example, understanding the best-case partition complexity, where the data is partitioned in the most advantageous manner (subject to constraints such as each player receiving an equal amount of the input), is important for understanding various problems in VSLI design [2]. Another kind of worst-case partition arises when the corresponding bits of two equal-length input strings are written on opposite sides of opaque cards (the "two-sided card model" [11, 33]). However, a natural question that, to the best of our knowledge, has not been explored to date, is what happens when the input is partitioned amongst the players *at random*. In other words, does evaluating a given function require significant communication for only a few pathological partitions or does such a requirement apply to an overwhelming fraction of all partitions?

In this paper we initiate a study of communication complexity under random partitions of the input. In fact, we consider more general allocations of the input to the players, possibly allowing information overlap, where bits of data may be known to more than one player. A particularly interesting case is when each *token* of data is given to a player chosen uniformly at random; this provides a convenient way to count "bad" partitions. We consider a communication lower bound to be *robust* if it applies to all but a small fraction of possible partitions. One can think of our work as a form of average-case analysis. However, it is important to note that our work stands in contrast to the usual notion of distributional complexity: rather than considering a random input, we consider worst-case inputs allocated randomly amongst the players.

Data Stream Computation: A strong motivation for our study is the goal of proving robust lower bounds for problems in the data stream model. The data stream model has enjoyed significant attention in recent years owing to some influential work in the late

1990s [3, 22, 13]. Study of this model has thrived both because of the rich theoretical questions it raises and its applicability to numerous real world applications such as network monitoring and query planning in databases. Consequently, it is important to understand the complexity of problems not just in worst-case but also in "average-case" settings. To this end we prove lower bounds in the setting that the ordering of tokens in the data stream is chosen not adversarially but randomly, from the set of all permutations. Arguably, such a lower bound provides a stronger indication that a problem cannot be solved efficiently in the data stream model than a "fragile" lower bound that might depend on a clever adversarial ordering. (For further, more detailed, justification see the recent papers [17, 7]).

Random-order data streams were considered by Munro and Paterson [31] in one of the first studies of the data stream model. In recent years there has a been a resurgence of interest in this model for a variety of reasons [7, 10, 17, 21, 19, 20]. Uniform or near-uniform orderings can arise in a number of ways, such as when processing a stream of samples that are drawn independently from a non-time-varying distribution. For problems such as quantile estimation and finding frequent items it has been shown that there is a considerable difference between processing random-order stream and adversarial streams. In particular, streaming algorithms to find the median using polylog space require exponentially fewer passes if the stream is ordered randomly [17] and this is tight [19, 7].

In this paper, we use robust lower bounds on communication complexity in order to deduce robust data stream lower bounds. Once the communication bounds have been shown, the data stream bounds follow by simple reductions to appropriate instances of communication. Where such bounds were known before, our method yields much cleaner proofs and tighter bounds. It also yields a number of new bounds for random-order data streams.

Our Results and Overview: We begin in Section 2 with a formal definition of our model and introduce some techniques and terminology. We prove the following results:

- *Multi-Party Set Disjointess:* We consider the problem of t-party set disjointness where each entry of the relevant $t \times n$ matrix is given to one of p players chosen uniformly at random. If $p = \Omega(t^2)$ then we show that any randomized protocol requires $\Omega(n/t)$ communication. See Section 3.
- *Pointer Jumping and Selection:* We consider a natural variant of tree pointer jumping, called weight-based tree pointer jumping, that is related to the problem of selection. In this problem, instead of an explicit pointer at each node, we have a binary string at each node whose weight encodes the pointer. We consider t-ary trees of depth $p+1$ and show that if the bits of these strings are distributed uniformly between multiple players, we require about $\Omega(n^{(2+\varepsilon)^{-p}})$ bits of communication for a p-round protocol. See Section 4.
- *Hamming Distance and Index:* For $x, y \in \{0,1\}^n$, let $\Delta(x,y) := \{i \in [n] : x_i \neq y_i\}$ denote the Hamming distance between x and y. We show that, for some constant c, any one-way protocol that can distinguish between the cases $\Delta(x,y) \leq n/2 - c\sqrt{n}$ and $\Delta(x,y) \geq n/2 + c\sqrt{n}$ requires $\Omega(n)$ communication if the $2n$ input bits are split uniformly between two players. We also show that a one-way protocol for the index problem — $\text{INDEX}(x,j) := x_j$, with $x \in \{0,1\}^n$, $j \in [n]$ — requires $\Omega(n)$ communication if the $n+1$ tokens (j being a single token) are split uniformly between two players. See Section 5.

The above communication lower bounds lead to a wide variety of lower bounds for data stream problems in the random-order model. In Section 6, we deduce such bounds, many of which are tight, for approximating frequency moments, the number of distinct values, entropy, information divergences, selection, and graph connectivity. Two of these bounds deserve particular emphasis. For the kth frequency moment, we obtain a robust lower bound of $\Omega(n^{1-3/k})$, which comes close to the optimal $\Omega(n^{1-2/k})$ bound under adversarial ordering. For the problem of median finding, our framework greatly simplifies the proof of a recent $\Omega(\log\log n)$ lower bound [7] on the number of passes required to achieve polylogarithmic space. Further, our pass-space tradeoff for this problem greatly improves the results of [7]: for instance, with two passes, we obtain a space lower bound of $\Omega(n^{1/10})$ as opposed to their $\Omega(n^{3/80})$.

2. NOTATION AND PRELIMINARIES

We summarize some notation that we need repeatedly. Define the *weight* $|x|$ of a Boolean vector $x \in \{0,1\}^N$ to be $|\{i : x_i = 1\}|$. Let $\mathbf{e}_i$ denote the vector that is 1 at location i and 0 elsewhere. For random variables X and Y: $\mathrm{E}[X]$ denotes the expectation and $\mathrm{H}(X)$ the entropy of X, $\mathrm{H}(X \mid Y)$ the conditional entropy of X given Y and $\mathrm{I}(X:Y)$ the mutual information between X and Y. We write $X \sim \mu$ to indicate that X is drawn from the probability distribution μ, and $X \equiv Y$ to indicate that X and Y have the same distribution. We denote by $\mathrm{V}(\mu,\nu)$ the total variation distance between the distributions μ and ν, i.e., $\mathrm{V}(\mu,\nu) = \frac{1}{2}\|\mu - \nu\|_1$. We denote the product distribution of μ and ν by $\mu \otimes \nu$.

The Binomial distribution with parameters n (number of trials) and p (success probability) is denoted $\mathcal{B}(n,p)$. The notation $X \in_R S$ indicates that X is chosen uniformly at random from the set S. For an integer k, $\binom{S}{k}$ denotes the set of all k-subsets of S and 2^S denotes the power set of S. We say that Q' is an (ε,δ)-approximation for Q if $\Pr[|Q' - Q| > \varepsilon Q] \leq \delta$.

2.1 The Communication Model

Traditionally, a two-party communication problem (between Alice and Bob, say) is formalised as a function, or partial function, on a domain of the form $X \times Y$, where the finite set X (resp. Y) is the set of Alice's (resp. Bob's) possible inputs. For our purposes, it is helpful to think of the input domain represented differently. We shall think of an input as an m-tuple of *tokens*, where the tokens are given to the players according to a random *allocation* drawn from a known distribution. Thus, it will help to represent the input domain as $X_1 \times X_2 \times \cdots \times X_m$, where X_i is the set of possible values for the ith token. Typically, each X_i will be either the set $\{0,1\}$ or the set $[N] := \{1,2,\ldots,N\}$, for some positive integer N. An allocation amongst p players is then a function $\sigma : [m] \to 2^{[p]}$.

A natural and interesting special case of an allocation is a *split*, where each token is given to exactly one player selected at random from amongst all players. It will be convenient to think of splits as functions $\sigma : [m] \to [p]$. A further special case is that of a *uniform split*, where each token is equally likely to go to each of the players: we let $\mathcal{U}_p$ denote the probability distribution of a uniform split amongst p players.

DEFINITION 2.1. *A* random-partition communication problem *for p players consists of a function $f : X_1 \times \cdots \times X_m \to Z$ and a probability distribution ν on allocations $\sigma : [m] \to 2^{[p]}$. A traditional communication problem is a special case, where ν is supported on a single allocation (that often happens to be a split). For a random-partition protocol P, let $P(x,\sigma)$ denote the (possibly random)* transcript *of P, and* $\mathrm{out}(P,x,\sigma)$ *the output of P, on input x allocated according to σ. For a traditional protocol, where σ has only one possible value, we drop σ from these notations.*

DEFINITION 2.2 (ERROR, COST, COMPLEXITY). *Let P be a protocol for a random-partition communication problem (f, ν). We define the error*

$$\mathrm{err}(P, f, \nu) := \max_x \Pr[\mathrm{out}(P, x, S) \neq f(x)],$$

where $S \sim \nu$. If μ is a distribution on the inputs to f, we define the distributional error

$$\mathrm{err}_\mu(P, f, \nu) := \Pr[\mathrm{out}(P, X, S) \neq f(X)],$$

where $X \sim \mu$ and $S \sim \nu$. Let $\mathrm{cost}(P) := \max_{x,\sigma} |P(x, \sigma)|$ denote the communication cost of P.

Define the δ-error communication complexity of (f, ν) to be

$$\mathrm{R}_\delta(f, \nu) := \min\{\mathrm{cost}(P) : \mathrm{err}(P, f, \nu) \leq \delta\}$$

and the δ-error μ-distributional complexity to be

$$\mathrm{R}_{\mu,\delta}(f, \nu) := \min\{\mathrm{cost}(P) : \mathrm{err}_\mu(P, f, \nu) \leq \delta\}.$$

Let $\mathrm{R}^{\rightarrow}$ and R^k denote the restrictions of these notions to one-way and k-round protocols, respectively. For traditional communication problems, we drop ν from these notations.

Informally, a communication lower bound is *robust* if it applies to $\mathrm{R}_\delta(f, \nu)$ or $\mathrm{R}_{\mu,\delta}(f, \nu)$ for some high-entropy distribution ν, such as the aforementioned $\mathcal{U}_p$.

2.2 Technique Preliminaries

In this section we introduce some of the main techniques that we use to establish our results. These are all based on considering random input in addition to random splits.

The notion of information complexity has been used on many occasions in the study of communication protocols [9, 5, 8, 25]. Loosely speaking, information complexity is used to establish a direct sum result, which reduces the problem of lower bounding the complexity of a "compound" problem (here, disjointness) to that of lower bounding the complexity of a simpler "base" problem (here, the AND function). The direct sum result follows from a *simulation argument*, where we design a protocol for the base problem that randomly pads its input to generate an artificial input for the compound problem and then simulates a protocol for the compound problem. Here, for our robust lower bounds for set disjointness, we need to extend the methods of Bar-Yossef et al. [5] to handle public coin protocols. This is a subtle matter: we must condition on the public coin to have a meaningful notion of information complexity. At the same time, we must be careful about how the public coin is used in the simulation argument, ensuring that we do not introduce undesirable correlations in the random padding.

DEFINITION 2.3 (INFORMATION COST AND COMPLEXITY). *For a traditional private coin protocol P and a distribution μ on its inputs, we define $\mathrm{icost}_\mu(P) = \mathrm{I}(X : P(X))$, where $X \sim \mu$. If D is a random variable (possibly correlated with X), we define the D-conditional μ-information cost $\mathrm{icost}_\mu(P \mid D) = \mathrm{I}(X : P(X) \mid D)$. We extend these notions to public coin protocols thus: if P^R is a public coin protocol that uses a public random string R, we define*

$$\mathrm{icost}^{\mathrm{pub}}_\mu(P^R) = \mathrm{I}(X : P^R(X) \mid R)\,,$$

where $X \sim \mu$ and

$$\mathrm{icost}^{\mathrm{pub}}_\mu(P^R \mid D) = \mathrm{I}(X : P^R(X) \mid D, R)\,.$$

For each information cost measure above, we define a corresponding information complexity measure in the natural way, e.g., for a communication problem f,

$$\mathrm{IC}_{\mu,\delta}(f) = \inf\{\mathrm{icost}_\mu(P) : \mathrm{err}(P, f) \leq \delta\}\,.$$

We write $\mathrm{IC}^{\mathrm{pub}}$ and $\mathrm{IC}^{\mathrm{pub},\rightarrow}$ for the information complexity of public coin protocols, and public coin one-way protocols, respectively.

We also consider random inputs $X \sim \mu$ in another setting. Some of our lower bounds will use a reduction from a communication problem in the fixed-partition model to one where the partition $\sigma \sim \nu$. In these reductions, the players choose σ using public random bits, but then distributing the input tokens according to σ would seem to necessitate communicating a large fraction of the data and this would render the reduction useless. The solution is to use distributional lower bounds on fixed-partition problems. This suggests that the players may "guess" data that they do not know. Unfortunately, the issue that arises is that this guessing may be correlated to the distribution of σ. However, the following lemma connects us back to the "usual" situation, when inputs and allocations are independent of each other, provided this correlation is sufficiently weak.

LEMMA 2.4. *If a protocol P satisfies $\Pr_{(x,\sigma)\sim\xi}[\mathrm{out}(P, x, \sigma) \neq f(x)] \leq \delta$, for some joint distribution ξ, then $\mathrm{err}_\mu(P, f, \nu) \leq \delta + \mathrm{V}(\mu \otimes \nu, \xi)$.*

PROOF. Simply observe that

$$\begin{aligned}\mathrm{err}_\mu(P, f, \nu) &= \Pr_{x\sim\mu,\sigma\sim\nu}[\mathrm{out}(P, x, \sigma) \neq f(x)] \\ &\leq \Pr_{(x,\sigma)\sim\xi}[\mathrm{out}(P, x, \sigma) \neq f(x)] + \mathrm{V}(\mu \otimes \nu, \xi)\,.\end{aligned}$$

□

3. MULTI-PARTY SET DISJOINTNESS

Let $\mathrm{DISJ}_{n,t} : \{0,1\}^{nt} \to \{0,1\}$ denote the following problem. The input is an (nt)-tuple of bits denoted $\{x_{ij}\}_{i\in[t],j\in[n]}$, to be thought of as the entries of a $t \times n$ Boolean matrix. The input satisfies a *unique intersection promise*, namely, each column of the matrix has weight in $\{0, 1, t\}$ and at most one column has weight t. The desired output is $\bigvee_{j=1}^n \bigwedge_{i=1}^t x_{ij}$. Chakrabarti, Khot and Sun [8] show that $\mathrm{R}_\delta(\mathrm{DISJ}_{n,t}) = \Omega(n/(t \log t))$ and $\mathrm{R}^{\rightarrow}_\delta(\mathrm{DISJ}_{n,t}) = \Omega(n/t)$, under a t-player split where each player receives one row of the matrix.

Let $\mathrm{AND}_t : \{0,1\}^t \to \{0,1\}$ be shorthand for $\mathrm{DISJ}_{1,t}$. Let $D \in_R [t]$ and $X \in_R \{0, \mathbf{e}_D\}$. Denote the resulting joint distribution of (X, D) by λ and the marginal distribution of X by μ. The lower bound of [8] follows by carefully analysing $\mathrm{IC}_{\mu,\delta}(\mathrm{AND}_t \mid D)$ and using the direct sum techniques of Bar-Yossef et al. [5] to link this quantity with $\mathrm{IC}_{\mu^n,\delta}(\mathrm{DISJ}_{n,t} \mid D^n)$.

Here, we consider the random-partition communication problem $(\mathrm{DISJ}_{n,t}, \mathcal{U}_p)$ for some suitably large number, p, of players. We now prove a robust lower bound on its complexity by extending the earlier techniques. We start with the following well-known fact.

FACT 3.1 (BIRTHDAY PROBLEM). *For $t, p \in \mathbb{N}^+$, let $\alpha(t, p)$ denote the probability that t independent random variables, each drawn uniformly from $[p]$, do not take t distinct values. Then*

$$1 - e^{-t(t-1)/(2p)} \leq \alpha(t, p) \leq t(t-1)/(2p)\,.$$

LEMMA 3.2. *Let $\delta' = \delta + \alpha(t, p)$. Then*

$$\mathrm{R}_\delta(\mathrm{DISJ}_{n,t}, \mathcal{U}_p) \geq n \cdot \mathrm{IC}^{\mathrm{pub}}_{\mu,\delta'}(\mathrm{AND}_t \mid D).$$

PROOF. Let P be an optimal δ-error protocol for $(\text{DISJ}_{n,t}, \mathcal{U}_p)$, i.e., a protocol that achieves $\text{cost}(P) = \text{R}_\delta(\text{DISJ}_{n,t}, \mathcal{U}_p)$. Consider n independent pairs of random variables $(X_1, D_1), \ldots, (X_n, D_n)$, each drawn from λ. Then $X := X_1 X_2 \ldots X_n \sim \mu^n$ is a suitable random input for $\text{DISJ}_{n,t}$. Let $S \sim \mathcal{U}_p$ be a random split. Then, by standard information theoretic arguments, we have

$$\begin{aligned}
\text{cost}(P) &= \max_{x,\sigma} |P(x,\sigma)| \geq \text{H}(P(X,S)) \\
&\geq \text{I}(X : P(X,S) \mid D_1 D_2 \ldots D_n, S) \\
&\geq \sum_{j \in [n]} \text{I}(X_j : P(X,S) \mid D_1 D_2 \ldots D_n, S) \qquad (1) \\
&= \sum_{j \in [n]} \text{E}_d[\text{I}(X_j : P(X,S) \mid D_j, S, D_{-j} = d)],
\end{aligned}$$

where (1) holds because the X_js are independent even after conditioning on $D_1 D_2 \ldots D_n$ and S. Here, D_{-j} denotes the vector $(D_1, \ldots, D_{j-1}, D_{j+1}, \ldots, D_n)$ and the final expectation is over d drawn uniformly from $[t]^{[n]\setminus\{j\}}$. To finish the proof, it suffices to show that

$$c_{j,d} := \text{I}(X_j : P(X,S) \mid D_j, S, D_{-j} = d) \geq \text{IC}^{\text{pub}}_{\mu,\delta'}(\text{AND}_t \mid D),$$

for each $j \in [n]$ and each $d \in [t]^{[n]\setminus\{j\}}$. To this end, we shall design a certain δ'-error t-party traditional protocol $Q^S_{j,d}$ for AND_t, parametrised by j and d, that uses S as a public random string. Further, for each possible value σ of S, the transcript $Q^\sigma_{j,d}(X_j)$ is either constant or distributed identically to $(P(X,\sigma) \mid D_{-j} = d)$. Then, as required, we shall have

$$\begin{aligned}
\text{IC}^{\text{pub}}_{\mu,\delta'}(\text{AND}_t \mid D) &\leq \text{icost}^{\text{pub}}_\mu(Q^S_{j,d} \mid D_j) \\
&= \text{I}(X_j : Q^S_{j,d}(X_j) \mid D_j, S) \\
&\leq c_{j,d}.
\end{aligned}$$

The protocol $Q^\sigma_{j,d}$ works as follows. On input $x = (x_1, \ldots, x_t) \in \{0,1\}^t$, the players create a random virtual input $\{Z_{ik}\}_{i,k} \in \{0,1\}^{t \times n}$ for $\text{DISJ}_{n,t}$, pretend that this input has been split according to σ amongst p virtual players, and then, if possible, simulate the behaviour of these virtual players when they execute P on the virtual input. The virtual input is obtained by embedding x into the jth column of a random Boolean matrix drawn from $(\mu^n | D_{-j} = d)$. To wit:

$$Z_{ik} \in_R \begin{cases} \{x_i\}, & \text{if } k = j, \\ \{0\}, & \text{if } k \neq j \text{ and } d(k) \neq i, \\ \{0,1\}, & \text{if } k \neq j \text{ and } d(k) = i. \end{cases}$$

Therefore, the simulation is possible iff σ assigns each of the inputs $(Z_{1j}, \ldots, Z_{tj})$ to a distinct virtual player; we shall say that σ *ramifies* if this condition is met. If σ does not ramify, the players abort, leading to a constant empty transcript and an error probability of 1. If σ does ramify, then Player i plays the role of that virtual player who is assigned Z_{ij} by σ. The crucial observation that makes this role-playing possible is that all the *other* bits assigned to that virtual player are available to Player i, because they are either set to 0 or can be drawn uniformly at random from $\{0,1\}$ using Player i's private coin. All virtual players who are not assigned any of the inputs $\{Z_{ij}\}_{i \in [t]}$ are simulated by Player 1 (say). Thus, if σ ramifies, then $Q^\sigma_{j,d}(X_j) \equiv (P(X,\sigma) \mid D_{-j} = d)$. Finally, $Q^S_{j,d}$ is indeed a δ'-error protocol, because

$$\begin{aligned}
\text{err}(Q^S_{j,d}, \text{AND}_t) &\leq \Pr[\sigma \text{ does not ramify}] + \text{err}(P, \text{DISJ}_{n,t}, \mathcal{U}_p) \\
&= \alpha(t,p) + \delta = \delta'.
\end{aligned}$$

□

LEMMA 3.3. *If $\delta \leq 1/20$, then*

$$\text{IC}^{\text{pub}}_{\mu,\delta}(\text{AND}_t \mid D) = \Omega(1/(t \log t))$$
$$\textit{and} \quad \text{IC}^{\text{pub},\rightarrow}_{\mu,\delta}(\text{AND}_t \mid D) = \Omega(1/t).$$

PROOF. From the work of Chakrabarti, Khot and Sun [8] we can deduce that for a *private* coin traditional protocol P such that $\text{err}(P, \text{AND}_t) \leq 1/10$, we have $\text{icost}_\mu(P \mid D) = \Omega(1/(t \log t))$. Now, consider a public coin δ-error protocol Q^S for AND_t that uses a public random string S. For each possible value σ of S, define $c_\sigma := \text{icost}_\mu(Q^\sigma \mid D)$, so that $\text{E}_\sigma[c_\sigma] = \text{icost}^{\text{pub}}_\mu(Q^S \mid D)$ and $\text{E}_\sigma[\text{err}(Q^\sigma, \text{AND}_t)] \leq \delta$.

Suppose $\delta \leq 1/20$. Call a particular split σ "good" if

$$\text{err}(Q^\sigma, \text{AND}_t) \leq 2\delta \leq 1/10.$$

By Markov's inequality, $\Pr[\sigma \text{ is good}] \geq 1/2$. For each good σ, considering the private coin protocol Q^σ shows $c_\sigma = \Omega(1/(t \log t))$. Thus, $\text{E}_\sigma[c_\sigma] = \Omega(1/(t \log t))$. We conclude that

$$\text{IC}^{\text{pub}}_{\mu,\delta}(\text{AND}_t \mid D) = \Omega(1/(t \log t)).$$

The proof for one way protocols follows similarly. □

Putting together Fact 3.1, Lemma 3.2 and Lemma 3.3 yields the following theorem.

THEOREM 3.4. *For $\delta \leq 1/40$ and $p \geq 20t^2$, we have the robust lower bounds*

$$\text{R}_\delta(\text{DISJ}_{n,t}, \mathcal{U}_p) = \Omega(n/(t \log t))$$
$$\textit{and} \quad \text{R}^{\rightarrow}_\delta(\text{DISJ}_{n,t}, \mathcal{U}_p) = \Omega(n/t).$$

□

We note that in order to get this kind of robust lower bound for $\text{DISJ}_{n,t}$ under $\mathcal{U}_p$ that increases linearly with n, we *must* make p, the number of players, as large as $\Omega(t^2)$. This is because when an input x such that $\text{DISJ}_{n,t}(x) = 1$ is allocated to p players, with probability $\alpha(t,p)$ there exists a player that receives at least two tokens from the all-ones column. Therefore, a simple $O(p)$-communication protocol, where each player announces whether or not they have received two 1s from the same column, has error probability at most $1 - \alpha(t,p)$. By Fact 3.1, we now have $\text{R}_\delta(\text{DISJ}_{n,t}, \mathcal{U}_p) = O(p)$ for $p \leq t(t-1)/(2\ln(1/\delta)) = O(t^2)$.

4. POINTER JUMPING AND SELECTION

We now consider the *tree pointer jumping* problem $\text{TPJ}_{k,t}$, defined as follows. Consider a complete k-level t-ary tree, T, rooted at v_0. The input is a function $\phi : V(T) \to [t]$, with $\phi(v) \in \{0,1\}$ if v is a leaf of T. Define $g(v)$ to be the $\phi(v)$-th child of v if v is an internal node, and $\phi(v)$ if v is a leaf. The desired output is $\text{TPJ}_{k,t}(\phi) := g^{(k)}(v_0) = g(g(\cdots g(v_0) \cdots))$.

There are at least two natural ways to make a traditional communication problem out of $\text{TPJ}_{k,t}$, both of which are of interest to us. The first way is to have two players, Alice and Bob, with Alice (resp. Bob) receiving the values of $\phi(v)$ for odd-level (resp. even-level) vertices v; we use the convention that leaves are at level 1.

The second way is to have k players, with Player i receiving the values of $\phi(v)$ for vertices v on level i. When speaking of communication problems, we shall use $\text{TPJ}_{k,t}$ to denote the former, and $\text{M-TPJ}_{k,t}$ to denote the latter ("M" for "multi-player"). For $k = 2$ the two definitions coincide and we obtain the well-studied INDEX problem, for which strong one-way lower bounds are known [1], with numerous implications for stream computation. In particular, Guha and McGregor [19] use a reduction from INDEX to obtain a tight (up to logarithmic factors) space lower bound for estimating the median of a randomly ordered stream of numbers in one pass. This lower bound was recently extended to multiple passes by Chakrabarti, Jayram and Pătraşcu [7] via a rather different (and intricate) proof.

Here, we give a considerably simpler proof of a multi-pass lower bound for median finding,[1] and in fact improve upon previous bounds, by using a suitable reduction from $\text{TPJ}_{k,t}$. As an intermediate step, we consider a problem we call *weight-based tree pointer jumping*, or $\text{W-TPJ}_{k,n}$. This problem is closely related to $\text{TPJ}_{k,t}$ (with n determined by k and t) but the input is presented differently: instead of specifying $\phi(v)$ directly, the input includes a binary string $x_v \in \{0,1\}^{a_i}$ for each level-i node of T, where the weight $|x_v|$ determines $\phi(v)$. The lengths a_i are parameters that will be fixed later. The encoding works as follows. If v is a leaf ($i = 1$), then $x_v = \phi(v)$. Otherwise, x_v is any string with

$$|x_v| = \frac{1}{2}a_i + \left(\frac{1}{2}t - \phi(v) + \frac{1}{2}\right) b_{i-1},$$

where b_i is the total length of all strings associated with nodes in the subtree of a level-i node, i.e., $b_i = a_i + tb_{i-1}$ and $b_1 = 1$.[2]

Let $x \in \{0,1\}^n$ be the concatenation of all x_v. Define

$$\text{W-TPJ}_{k,n}(x) := \text{TPJ}_{k,t}(\phi).$$

The proof of the next theorem involves a reduction from W-TPJ to MEDIAN similar to that in [18].

THEOREM 4.1. *Let $\text{MEDIAN}_{m,N}$ denote the random-partition communication problem where the input consists of m tokens $(x_1, \ldots, x_m) \in [N]^m$ and the desired output is the median of this collection of tokens. For any $\delta > 0$, any allocation distribution ν, and any number $p \geq 1$ of rounds of communication, we have $\text{R}^p_\delta(\text{MEDIAN}_{n,\Theta(n)}, \nu) \geq \text{R}^p_\delta(\text{W-TPJ}_{k,n}, \nu)$.* □

PROOF. The proof follows from the following reduction from W-TPJ to MEDIAN. We start by defining some notation:

1. Let $v[i_1, \ldots, i_j]$ denote the i_jth child of $v[i_1, \ldots, i_{j-1}]$ where $v[]$ is the root of the tree.
2. Let the $(p+1)$ tuple $\langle h_p, \ldots, h_0 \rangle$ denote the base $(t+2)$ representation of $\sum_{i=0}^{p} h_i (t+2)^i$.

The reduction proceeds as follows:

1. For each internal node of level j, e.g., $v = v[i_p, \ldots, i_j]$, with associated binary string x_v, we generate a set of values $A(v)$ containing

$$|x_v| \text{ copies of } \langle i_p, \ldots, i_j, 0, 0, \ldots 0 \rangle$$
$$\text{and } a_i - |x_v| \text{ copies of } \langle i_p, \ldots, i_j, t+1, 0, \ldots 0 \rangle .$$

   This can be done by generating a copy of $\langle i_p, \ldots, i_j, 0, 0, \ldots 0 \rangle$ for each bit of x_v that is 1 and then generating a copy of $\langle i_p, \ldots, i_j, t+1, 0, \ldots 0 \rangle$ for each bit of x_v that is 0.
2. For leaf node, e.g., $v = v[i_p, \ldots, i_1]$, we generate a single element $(i_p, \ldots, i_1, f(v))$.

By construction, the least significant bit of median$(\cup_{v \in V(T)} A(v))$ equals $\text{W-TPJ}(f)$. □

4.1 A Robust Two-Player Lower Bound

Our starting point is a bounded-round lower bound for the traditional two-player communication problem $\text{TPJ}_{k,t}$ described above, where a "round" consists of one message from either Alice or Bob. This bound can be deduced from the work of Klauck et al. [27], who in fact studied the problem in the more general *quantum* communication setting. The underlying intuition is that of *round elimination* à la Miltersen et al. [30] and Sen [34].

THEOREM 4.2. *We have $\text{R}^p_{\mu,1/3}(\text{TPJ}_{p+1,t}) = \Omega(t/p^2)$, where μ is the uniform distribution over inputs.* □

To obtain the desired robust lower bound for W-TPJ, we use a reduction from TPJ that introduces a slight correlation between input and split, and then appeal to Lemma 2.4 to correct for this.

THEOREM 4.3. *We have*

$$\text{R}^p_{1/24}(\text{W-TPJ}_{p+1,n}, \mathcal{U}_2) = \Omega\big(n^{\frac{1}{((p-1)2^{p+1}+2)}} \cdot (\log n)^{\frac{-1}{(2(p-1))}} \cdot p^{-2}\big).$$

Thus, for any constant $\varepsilon > 0$, for n and p large enough with $p = O(\log \log n)$, we have

$$\text{R}^p_{1/24}(\text{W-TPJ}_{p+1,n}, \mathcal{U}_2) = \Omega\big(n^{(2+\varepsilon)^{-p}}\big) .$$

PROOF. Let P be a protocol for $(\text{W-TPJ}, \mathcal{U}_2)$, between players Carol and Dave, such that $\text{err}(P, \text{W-TPJ}, \mathcal{U}_2) \leq \frac{1}{24}$. Consider a uniform random instance ϕ of TPJ that Alice and Bob must solve. We construct a protocol Q for this. In Q, Alice and Bob use public randomness to construct a random input for W-TPJ together with a random split of its tokens between Carol and Dave. They then proceed to simulate P on this instance, with Alice and Bob simulating Carol and Dave, respectively. Define

$$a_i = (ct^{2(p+2)} \log n)^{2^{i-1}-1} t^{-2(3 \cdot 2^{i-1} - i - 2)}$$

for some large constant c to be determined. For each internal node v in level i, using public randomness:

- Alice and Bob pick $d_{1v} \sim \mathcal{B}\left(\frac{1}{2}a_i, \frac{1}{2}\right)$, $d_{2v} \sim \mathcal{B}\left(\frac{1}{2}a_i, \frac{1}{2}\right)$ and $S_v \in_R \binom{[a_i]}{d_{1v} + d_{2v}}$.
- Assume level(v) is even. Alice determines $x_{v,j}$ for $j \in S_v$ and, uniformly at random, sets d_{1v} of these tokens to 1 and the remaining d_{2v} tokens to 0. Bob determines $x_{v,j}$ for $j \notin S_v$ and, uniformly at random, sets $|x_v| - d_{1v}$ of these tokens to 1 and the remaining $a_i - |x_v| - d_{2v}$ tokens to 0. If level(v) is odd then Alice and Bob's roles are reversed.

For each leaf node v, using public randomness:

- With probability $\frac{1}{2}$, Alice determines $x_{v,1} = \phi(v)$. Otherwise Bob determines $x_{v,1} \in_R \{0,1\}$.

The resulting instance of W-TPJ consists of the random input x so generated together with the random split σ where Carol receives all the tokens determined by Alice, and Dave receives all those determined by Bob. This completes the description of Q. Note that, with probability $3/4$, $\text{W-TPJ}(x) = \text{TPJ}(\phi)$.

[1] Our results, like the earlier ones [19, 7], apply to the more general problem of selection.

[2] Note that for $|x_v|$ to be a positive integer this implies that $a_i/b_{i-1} \in \{t-1, t+1, t+3, \ldots\}$.

It remains to show that x and σ are sufficiently close to being independent. Note that the marginals are correct: we do have $\sigma \sim \mathcal{U}_2$ and the values of $x_{v,j}$ are indeed chosen according to a uniform setting of $\phi(v)$. The issue is that the joint distribution is not a product distribution. However, note that had $d_{1,v}$ and $d_{2,v}$ been chosen according to $\mathcal{B}(|x_v|, \frac{1}{2})$ and $\mathcal{B}(a_i - |x_v|, \frac{1}{2})$, respectively, then σ and x would have been independent. For each internal node v at level i, let

$$\tilde{A}_v := \mathcal{B}\left(\frac{1}{2}a_i, \frac{1}{2}\right), \quad \tilde{B}_v := \mathcal{B}\left(\frac{1}{2}a_i, \frac{1}{2}\right),$$
$$A_v := \mathcal{B}\left(|x_v|, \frac{1}{2}\right), \quad B_v := \mathcal{B}\left(a_i - |x_v|, \frac{1}{2}\right).$$

Hence, we need to show that the product distribution of all $\tilde{A}_v$ and $\tilde{B}_v$ is sufficiently close to that of all A_v and B_v. Using Lemma A.1, we can bound the total variation distance in terms of a_i and b_i as follows,

$$\begin{aligned}\mathrm{V}\Big(\bigotimes_v(\tilde{A}_v \otimes \tilde{B}_v), \bigotimes_v(A_v \otimes B_v)\Big) &\le \sum_v \mathrm{V}\big(\tilde{A}_v, A_v\big) + \sum_v \mathrm{V}\big(\tilde{B}_v, B_v\big) \\ &\le O(\sqrt{\log n}) \sum_{i=2}^{p+1} \frac{t^{p+2-i} b_{i-1}}{\sqrt{a_i}}\end{aligned}$$

where the first inequality follows from the triangle inequality. Noting that $b_{i-1} \le 2a_{i-1}$ and substituting in the value for a_i, the distance can be made less than $\frac{1}{24}$ for sufficiently large constant c. By Lemma 2.4,

$$\mathrm{err}_\mu(Q, \text{TPJ}_{p+1,t}) \le \frac{1}{4} + \frac{1}{24} + \mathrm{err}(P, \text{W-TPJ}_{p+1,n}, \mathcal{U}_2) \le \frac{1}{3}.$$

Therefore, by Theorem 4.2,

$$\mathrm{R}^p_{1/24}(\text{W-TPJ}_{p+1,n}, \mathcal{U}_2) = \Omega(t/p^2).$$

Note that

$$n = b_{p+1} = O((ct^{2(p+2)} \log n)^{2^p - 1} t^{-2(3\cdot 2^p - p - 3)})$$

and hence

$$\begin{aligned} t &= \Omega\Big(n^{\frac{1}{(p-1)2^{p+1}+2}} \big/ (c \log n)^{\frac{2^p-1}{(p-1)2^{p+1}+2}}\Big) \\ &\ge \Omega\Big(n^{\frac{1}{(p-1)2^{p+1}+2}} \big/ (c \log n)^{\frac{1}{2(p-1)}}\Big). \end{aligned}$$

□

4.2 A Robust Multi-Player Lower Bound

The two-player lower bound above is already sufficient to improve upon previous data stream lower bounds for selection in randomly ordered streams. We now prove a multi-player variant of Theorem 4.2 that gives even tighter data stream lower bounds.

THEOREM 4.4. *Let $\mathcal{V}_p$ be the (non-uniform) split distribution that gives each token to Player 1 with probability $\frac{1}{2}$ and to Player i with probability $\gamma := 1/(2p)$ for each $i \in \{2 \ldots, p+1\}$. Then, we have*

$$\mathrm{R}^p_{1/10}(\text{W-TPJ}_{p+1,n}, \mathcal{V}_p) = \Omega\big(n^{\frac{1}{((p-1)2^{p+1}+2)}} \cdot (\log n)^{\frac{-1}{(2(p-1))}} \cdot p^{-2}\big).$$

As before, for any constant $\varepsilon > 0$, for n and p large enough with $p = O(\log\log n)$, we have

$$\mathrm{R}^p_{1/10}(\text{W-TPJ}_{p+1,n}, \mathcal{V}_p) = \Omega\big(n^{(2+\varepsilon)^{-p}}\big).$$

The proof is similar to that of Theorem 4.3 so we outline the main differences. The starting point is the traditional $(p+1)$-player problem $\text{M-TPJ}_{p+1,t}$, for which we can prove the following lower bound, analogous to Theorem 4.2. The proof uses Sen's version of the round elimination lemma [34]; the details are in the full version of the paper.

THEOREM 4.5. *We have $\mathrm{R}^p_{\mu,1/3}(\text{M-TPJ}_{p+1,t}) = \Omega(t/p^2)$ where μ is the uniform distribution over inputs. Here, a "round" consists of one message from each player, in the order Player 1, …, Player $(p+1)$.* □

The construction of the reduction differs as follows. For a node v with $i = \text{level}(v)$, using public randomness, the players all pick $d_{1v} \sim \mathcal{B}\left(\frac{a_i}{2}, 1-\gamma\right)$, $d_{2v} \sim \mathcal{B}\left(\frac{a_i}{2}, 1-\gamma\right)$, and $S_v \in_R \binom{[a_i]}{d_{1v}+d_{2v}}$. The players other than the ith player set d^j_{1v} values of $\{x_{v,k} : k \in S_v\}$ to 1 and the rest to 0.

If σ and the data were independent they should be distributed as $\mathcal{B}(|x_v|, 1-\gamma)$ and $\mathcal{B}(a_i - |x_v|, 1-\gamma)$ respectively. However, if a_i is chosen as

$$a_i = (cp^2 t^{2(p+2)} \log n)^{2^{i-1}-1} t^{-2(3\cdot 2^{i-1} - i - 2)}$$

for some sufficiently large constant, then by appealing to Lemma A.1, the total variation distance can be made arbitrarily small. The extra p^{2^i-2} term in a_i has only a constant factor effect on the bound as

$$n = b_{p+1} = O((cp^2 t^{2(p+2)} \log n)^{2^p - 1} t^{-2(3\cdot 2^p - p - 3)})$$

and, because $p^{\frac{1}{2(p-1)}} = O(1)$ for $p \ge 2$,

$$\begin{aligned} t &= \Omega\Big(n^{\frac{1}{(p-1)2^{p+1}+2}} \big/ \left(c \log n p^2\right)^{\frac{2^p-1}{(p-1)2^{p+1}+2}}\Big) \\ &\ge \Omega\Big(n^{\frac{1}{(p-1)2^{p+1}+2}} \big/ (c \log n)^{\frac{1}{2(p-1)}}\Big). \end{aligned}$$

5. HAMMING DISTANCE AND INDEX

In this section, we prove robust lower bounds for INDEX and HAM-DIST_G, in the one-way communication model. For our purposes, we define the INDEX problem over inputs $x \in [n] \times \{0,1\}^n$ as follows: $\text{INDEX}(x) := x_j$ where $j := x_0$. Traditionally, one considers the worst-case partition where Alice (the player who speaks) holds $x_1 \ldots x_n$ and Bob holds j. Strong randomized lower bounds are known in this setting [1]. HAM-DIST_G [26, 35, 23] is defined based on the function $\Delta(x) = |\{i : x_{2i} \ne x_{2i-1}\}|$ over inputs $x \in \{0,1\}^{2n}$. With the promise that $\Delta(x)$ does not fall between $n/2 - G$ and $n/2 + G$,

$$\text{HAM-DIST}_G(x) := \begin{cases} 0, & \text{if } \Delta(x) \ge n/2 + G, \\ 1, & \text{if } \Delta(x) \le n/2 - G. \end{cases}$$

5.1 Hamming Distance

The main idea is to create an instance of HAM-DIST in the fixed partition model, and then pad this with carefully chosen random bits so that the resulting split appears almost uniform.

THEOREM 5.1. *There exists a constant c_3 such that*

$$\mathrm{R}^{\rightarrow}_{1/4}(\text{HAM-DIST}_{c_3\sqrt{n}}, \mathcal{U}_2) = \Omega(n).$$

PROOF. We reduce the traditional one-way INDEX problem to our HAM-DIST problem. Suppose Alice holds a string $x \in \{0,1\}^{n'}$

with $n' = c_2 n$ and Bob holds $j \in [n']$, where $c_2 < 1$ will be a constant to be fixed later. We know that $R^{\rightarrow,\text{pub}}_{0.49}(\text{INDEX}) = \Omega(n)$.

Suppose there exists a one-way protocol P such that

$$\text{err}_\mu(P, \text{HAM-DIST}_{c_3\sqrt{n}}, \mathcal{U}_2) \leq 1/4\,,$$

where μ is the uniform distribution over inputs. Let $r \in_R \{-1, 1\}^{n'}$ be determined by public random bits. Define the indicator random variables T_{i1} and T_{i2} for the events "$\sum_{i=1}^{n'} r_i x_i > 0$" and "$r_j > 0$," respectively. It can be shown (see [26] for a proof) that, for some constant $c_1 > 0$,

$$\Pr\left[T_{i1} = T_{i2}\right] = \begin{cases} 1/2 - c_1/\sqrt{n'}, & \text{if INDEX}(x, j) = 0\,, \\ 1/2 + c_1/\sqrt{n'}, & \text{if INDEX}(x, j) = 1\,. \end{cases}$$

The players now generate an instance y of HAM-DIST using shared randomness. They first pick a split $\sigma \sim \mathcal{U}_2$. For each i such that $\sigma(2i) \neq \sigma(2i-1)$, with probability $p = c_2^{1/4}$, the players set (y_{2i}, y_{2i-1}) based on T_{i1} and T_{i2}: since T_{i1} is known to Alice, she sets whichever input bit was allocated to her as T_{i1}, and Bob similarly uses T_{i2}. Otherwise, set $(y_{2i}, y_{2i-1}) \in_R \{0, 1\}^2$. Define $\Delta = \Delta(y) = |\{i : y_{2i} \neq y_{2i-1}\}|$.

CLAIM 5.2. *For sufficiently small c_2,*

$$(x_j = 0) \Rightarrow \Pr\left[\frac{\Delta}{n} > \frac{1}{2} + \frac{c_1}{5\sqrt{n'}}\right] \geq 0.99,$$
$$\textit{and } (x_j = 1) \Rightarrow \Pr\left[\frac{\Delta}{n} < \frac{1}{2} - \frac{c_1}{5\sqrt{n'}}\right] \geq 0.99.$$

PROOF OF CLAIM. Let t be the number of times Alice and Bob insert bits from T into their constructed strings. Note that $\text{E}[t] = pn$ and, by an application of the Chernoff bound, for sufficiently large n, we have $\Pr\left[t \leq np/2\right] \leq 1/1000$.

$$(x_j = 1) \Rightarrow \Pr\left[\tfrac{\Delta}{n} \leq \tfrac{1}{2} - \tfrac{c_1}{5\sqrt{n'}}\right] = \Pr\left[\tfrac{\Delta}{n} - \text{E}\left[\tfrac{\Delta}{n}\right] \leq -\tfrac{c_1}{5\sqrt{n'}}\right] \leq \exp\left(\frac{-c_1^2}{25c_2^{1/4}}\right)$$
$$(x_j = 0) \Rightarrow \Pr\left[\tfrac{\Delta}{n} \geq \tfrac{1}{2} - \tfrac{2c_1}{5\sqrt{n'}}\right] = \Pr\left[\tfrac{\Delta}{n} - \text{E}\left[\tfrac{\Delta}{n}\right] \geq +\tfrac{c_1}{5\sqrt{n'}}\right] \leq \exp\left(\frac{-c_1^2}{25c_2^{1/4}}\right)$$

Hence the claim holds true for sufficiently small c_2. □

While σ is not fully independent of y, it has sufficient independence, as shown by the following claim:

CLAIM 5.3. *For sufficiently small c_2, with probability at least 5/8, P answers $\text{HAM-DIST}_{c_3\sqrt{n}}$ correctly on y.*

PROOF OF CLAIM. Let μ_p be the distribution over $y \in \{0, 1\}^{2n}$. For $p = 0$ both y and the partition, are uniformly and independently chosen. We argue that $|\mu_p - \mu_0| \leq 1/8$ for sufficiently small c_2 and so by Lemma 2.4, P would answer $\text{HAM-DIST}_{c_3\sqrt{n}}$ with probability at least $3/4 - 1/8 = 5/8$ as required.

Define $I = \{i : \sigma(2i) \neq \sigma(2i-1)\}$. For $i \notin I$, $(y_{2i-1}, y_{2i}) \in_R \{0, 1\}^2$ under both μ_0 and μ_p. For $i \in I$, define the probability that a pair of bits differ as

$$q = \Pr_{\mu_p}[y_{2i} \neq y_{2i-1} | i \in I] = 1/2 - pc_1/\sqrt{n'}.$$

Therefore

$$\begin{aligned} |\mu_p - \mu_0| &= \sum_y |\Pr_{\mu_p}[y] - \Pr_{\mu_0}[y]| \\ &\leq \sum_y |2^{-n} q^\Delta (1-q)^{n-\Delta} - 2^{-2n}| \\ &= \sum_{d \in [n]} \binom{n}{d} |q^d (1-q)^{n-d} - 2^{-n}| \end{aligned}$$

By appealing to Lemma A.1 we can make this smaller than 1/8 by choosing c_2 sufficiently small. □

Hence if $c_3\sqrt{n} \leq n\frac{c_1}{5\sqrt{n'}}$, i.e., $c_3 \geq \frac{c_1}{5\sqrt{c_2}}$, the linear lower bound holds for $\text{HAM-DIST}_{c_3\sqrt{n}}$: otherwise, by Claim 5.2, $\text{HAM-DIST}_{c_3\sqrt{n}}$ on y reveals $\text{INDEX}(x, j)$ with probability at least $5/8 - 1/100 > 51/100$.

5.2 Index

In the usual fixed-partition model, INDEX can be thought of as a special case of $\text{DISJ}_{n,2}$, where one string is of the form $\mathbf{e}_i$. This is no longer the case under uniform splits, since the zeros in $\mathbf{e}_i$ get spread between the players, and leak information about which indices are not of interest. For INDEX, we prove a bound for a more general distribution ν that allocates multiple copies of input items amongst the players. This generalization is needed for proving subsequent data stream bounds.

THEOREM 5.4. *For $a, b = O(1)$, $R_{1/2^{a+b+2}}(\text{INDEX}, \nu) = \Omega(n)$ where ν is the distribution that distributions a copies of each x_i ($i \in [n]$) and b copies of x_0 uniformly between two players.*

PROOF. The proof is by reduction from INDEX when player 1 holds $y = y_1 \ldots y_n \in \{0, 1\}^n$ and player 2 holds index $x_0 = j$. Let μ be the uniform distribution over all possible inputs. Even when the players share public random bits, any one-way protocol succeeding with probability $1/2 + 1/2^{a+b+2}$ (for a, b constant) for instances of INDEX drawn from μ requires $\Omega(n)$ bits to be communicated [28].

Suppose there exists a one-way protocol P with the property that $\text{err}(P, \text{INDEX}, \mu, \nu) \leq 1/2^{a+b+2}$. The players agree on a partition $\sigma \sim \nu$ using their public random bits. Let B be the event that $\{1\} \subseteq \sigma(0)$ or that $\{2\} \subseteq \sigma(x_0)$, and note that $\Pr[B] = 1 - 1/2^{a+b}$. If B occurs then player 2 outputs 0 with probability 1/2 and 1 otherwise. Otherwise, using public random bits, the players choose a string r where $r_i \in_R \{0, 1\}$. They construct the string y' where $y'_i = r_i$ for $i \geq 1, \{2\} \subseteq \sigma(i)\}$ and $y'_i = y_i$ otherwise. They run protocol P for σ and y'.

The new protocol is correct with probability

$$\Pr[B]/2 + \Pr[\neg B \text{ and } P \text{ is correct}] \geq 1/2 + 1/2^{a+b+2}$$

and therefore the protocol must communicate $\Omega(n)$ bits. □

6. ROBUST LOWER BOUNDS FOR DATA STREAM COMPUTATION

Finally, we use our results on communication complexity to derive robust lower bounds for problems in the data stream model.

Frequency Moments: These are some of the most well-studied problems in the data stream model [3]. The stream comprises a sequence of m values $a_j \in [n]$. Define $f_i = |\{j : a_j = i\}|$. The kth frequency moment is

$$F_k := \sum_{i \in [n]} f_i^k\,.$$

We consider constant $k \geq 3$. It is known that any $O(1)$-pass algorithm that returns an (ε, δ)-approximation of F_k requires $\tilde{\Omega}(n^{1-2/k})$ space and that this is tight under worst-case orderings [24, 8]. However, it was observed that for random orderings and $m = \tilde{\Omega}_\varepsilon(an)$ there exists a single pass $\tilde{O}((n/a)^{1-2/k})$-space algorithm that (ε, δ)-approximates F_k [20]. The following theorem is a consequence of Theorem 3.4 by setting $t = n^{1/k}$. Since the result in Theorem 3.4 bounds the total amount of communication, the per-message bound implied scales with the reciprocal of the number of players (here, $\Omega(t^2)$).

THEOREM 6.1. *Any one-pass $(1/10, 1/10)$-approximation for F_k of a randomly ordered stream requires $\Omega(n^{1-3/k})$ space. If we assume that $m = \Omega(an)$ then $\Omega(n^{1-3/k}/a^3)$ space is required. For $O(1)$-pass algorithms, we have the corresponding lower bounds of $\Omega(n^{1-3/k}/\log n)$ and $\Omega(n^{1-3/k}/(a^3 \log n))$, respectively.*

Distinct Elements and Entropy: The number of distinct elements in a stream is $F_0 := |\{i : f_i \neq 0\}|$, and the empirical entropy is $H := \sum_{i \in [n]} (f_i/m) \log(m/f_i)$. One-pass, $\tilde{O}(\varepsilon^{-2})$-space, (ε, δ)-approximation algorithms are known for both problems [6, 14, 4]. We prove that the known algorithms are essentially tight even under random order. These results follow from Theorem 5.1 and the reductions in [6, Theorem 2] and [35, Section 3.2].

THEOREM 6.2. *For constant $k \neq 1$, a one-pass (ε, δ)-approximation for F_k of a randomly ordered stream requires $\Omega(\varepsilon^{-2})$ space. A one-pass (ε, δ)-approximation for H of a randomly ordered stream requires $\Omega(\varepsilon^{-2}/\log^2 \varepsilon^{-1})$ space.*

Selection: Selection is one of the most well studied problems in the data stream model [31, 15]. The following result improves upon the previous best single and multi-pass lower bounds [18, 7]. As an example, our theorem implies a $\tilde{\Omega}(n^{1/10})$ space lower bound for 3-pass algorithms whereas the best previous result was $\tilde{\Omega}(n^{3/80})$ [7]. The following theorem is immediate from Theorem 4.4.

THEOREM 6.3. *Any p-pass algorithm to return the median of a length-m randomly ordered stream which succeeds with probability at least $3/4$ requires $\Omega(m^{1/((p-1)2^{p+1}+2)}/(p2^p))$ space.*

Graph Streaming: We now consider bounds on estimating graph problems given a stream of edges in arbitrary order. Using Theorem 5.4 and reductions from [12, 22] it is possible to show:

THEOREM 6.4. *Given a stream of edges in random order, $\Omega(n)$ space is required to determine if the resulting graph is connected. Furthermore, any t-approximation of the distance between two nodes requires $O(n^{1+1/t})$ space.*

Information Divergences: The next theorem extends a result by Guha et al. [16] on the approximation of information divergences. The results follows from Theorem 5.4 using a variant of the reduction from [16].

THEOREM 6.5. *Given a randomly ordered stream defining two empirical distributions p and q on $[n]$, $\Omega(n)$ space is required to find an $\sqrt{1/2 + a/2}$ multiplicative approximation to*

$$Hellinger(p, q) = \sum (\sqrt{p_i} - \sqrt{q_i})^2$$

with probability at least $1 - 2^{-a-3}$ (for some even $a \in \mathbb{N}^+$.)

7. REFERENCES

[1] F. Ablayev. Lower bounds for one-way probabilistic communication complexity and their application to space complexity. *Theoretical Computer Science*, 175(2):139–159, 1996.

[2] A. V. Aho, J. D. Ullman, and M. Yannakakis. On notions of information transfer in VLSI circuits. In *STOC*, pages 133–139, 1983.

[3] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.

[4] Z. Bar-Yossef, T. Jayram, R. Kumar, D. Sivakumar, and L. Trevisan. Counting distinct elements in a data stream. In *Proc. 6th International Workshop on Randomization and Approximation Techniques in Computer Science*, pages 1–10, 2002.

[5] 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(4):702–732, 2004.

[6] A. Chakrabarti, G. Cormode, and A. McGregor. A near-optimal algorithm for computing the entropy of a stream. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 328–335, 2007.

[7] A. Chakrabarti, T. Jayram, and M. Pătraşcu. Tight lower bounds for selection in randomly ordered streams. In *ACM-SIAM Symposium on Discrete Algorithms*, 2008.

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

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

[10] E. D. Demaine, A. López-Ortiz, and J. I. Munro. Frequency estimation of internet packet streams with limited space. In *European Symposium on Algorithms*, pages 348–360, 2002.

[11] J. Edmonds and R. Impagliazzo. Manuscript. Unpublished, 1994.

[12] J. Feigenbaum, S. Kannan, A. McGregor, S. Suri, and J. Zhang. Graph distances in the data-stream model. In *ACM-SIAM Symposium on Discrete Algorithms*, pages pp. 745-754, 2007.

[13] J. Feigenbaum, S. Kannan, M. Strauss, and M. Viswanathan. An approximate L^1 difference algorithm for massive data streams. *SIAM Journal on Computing*, 32(1):131–151, 2002.

[14] P. Flajolet and G. N. Martin. Probabilistic counting algorithms for data base applications. *J. Comput. Syst. Sci.*, 31(2):182–209, 1985.

[15] M. Greenwald and S. Khanna. Efficient online computation of quantile summaries. In *ACM International Conference on Management of Data*, pages 58–66, 2001.

[16] S. Guha, P. Indyk, and A. McGregor. Sketching information divergences. In *Conference on Learning Theory*, pages 424–438, 2007.

[17] S. Guha and A. McGregor. Approximate quantiles and the order of the stream. In *ACM Symposium on Principles of Database Systems*, pages 273–279, 2006.

[18] S. Guha and A. McGregor. A general approach to multi-pass stream lower-bounds. *Manuscript*, 2007.

[19] S. Guha and A. McGregor. Lower bounds for quantile estimation in random-order and multi-pass streaming. In *International Colloquium on Automata, Languages and Programming*, pages 704–715, 2007.

[20] S. Guha and A. McGregor. Space-efficient sampling. In *AISTATS*, pages 169–176, 2007.

[21] S. Guha, A. McGregor, and S. Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 733–742, 2006.

[22] M. R. Henzinger, P. Raghavan, and S. Rajagopalan. Computing on data streams. *External memory algorithms*, pages 107–118, 1999.

[23] P. Indyk and D. P. Woodruff. Tight lower bounds for the distinct elements problem. *IEEE Symposium on Foundations of Computer Science*, pages 283–288, 2003.

[24] P. Indyk and D. P. Woodruff. Optimal approximations of the frequency moments of data streams. In *ACM Symposium on Theory of Computing*, pages 202–208, 2005.

[25] T. S. Jayram, R. Kumar, and D. Sivakumar. Two applications of information complexity. In *ACM Symposium on Theory of Computing*, pages 673–682, 2003.

[26] T. S. Jayram, R. Kumar, and D. Sivakumar. The one-way communication complexity of gap hamming distance. In *Manuscript*, 2007.

[27] H. Klauck, A. Nayak, A. Ta-Shma, and D. Zuckerman. Interaction in quantum communication and the complexity of set disjointness. In *ACM Symposium on Theory of Computing*, pages 124–133, 2001.

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

[29] T. W. Lam and W. L. Ruzzo. Results on communication complexity classes. *J. Comput. Syst. Sci.*, 44(2):324–342, 1992.

[30] P. B. Miltersen, N. Nisan, S. Safra, and A. Wigderson. On data structures and asymmetric communication complexity. *J. Comput. Syst. Sci.*, 57(1):37–49, 1998.

[31] J. I. Munro and M. Paterson. Selection and sorting with limited storage. *Theor. Comput. Sci.*, 12:315–323, 1980.

[32] C. H. Papadimitriou and M. Sipser. Communication complexity. *J. Comput. Syst. Sci.*, 28(2):260–269, 1984.

[33] P. Pudlák and J. Sgall. An upper bound for a communication game related to time-space tradeoffs. *Electronic Colloquium on Computational Complexity (ECCC)*, 2(10), 1995.

[34] P. Sen. Lower bounds for predecessor searching in the cell probe model. In *IEEE Conference on Computational Complexity*, pages 73–83, 2003.

[35] D. P. Woodruff. Optimal space lower bounds for all frequency moments. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 167–175, 2004.

[36] D. P. Woodruff. Distinct elements is hard even for random streams. *Manuscript*, 2008.

[37] A. C. Yao. Some complexity questions related to distributive computing (preliminary report). *ACM Symposium on Theory of Computing*, pages 209–213, 1979.

APPENDIX

A. VARIATIONAL DISTANCE BETWEEN BINOMIAL DISTRIBUTIONS

LEMMA A.1. *There exist constants $c_1, c_2 > 0$ such that, for $a \in \mathbb{N}$ sufficiently large, $q = 1-\gamma \in [1/2, 1-o(a)]$, and $w \in [a]$,*

$$\mathrm{V}\left(\mathcal{B}(a,q), \mathcal{B}(a-w,q)\right) \;\le\; c_1 w\sqrt{\ln(a)/(\gamma\, a)}\,,$$

and for $q = 1/2+\delta$,

$$\mathrm{V}\left(\mathcal{B}(a,1/2), \mathcal{B}(a,q)\right) \;\le\; c\delta^2 a\,.$$

PROOF. We may assume that $w = O(\sqrt{a\gamma\,\ln(a/w)})$ because otherwise the bound is trivial. Then by an application of the Chernoff bounds, there exists a constant c_1' such that

$$\max\Big(\Pr\Big[|\mathcal{B}(a,q)-aq| \ge c_1'\sqrt{a\gamma\,\ln\tfrac{a}{w}}\Big],\; \Pr\big[|\mathcal{B}(a-w,q)-aq| \ge c_1'\sqrt{a\gamma\,\ln(a/w)}\big]\Big) \;\le\; \tfrac{w}{\sqrt{a}}\,.$$

Let $t = c_1'\sqrt{a\gamma\,\ln(a/w)}$. Then,

$$\begin{aligned}
&\mathrm{V}(\mathcal{B}(a,q), \mathcal{B}(a-w,q))\\
&\le \tfrac{2w}{\sqrt{a}} + \textstyle\sum_{r=aq-t}^{aq+t}\left|\binom{a}{r}q^r(1-q)^{a-r} - \binom{a-w}{r}q^r(1-q)^{a-w-r}\right|\\
&\le \tfrac{2w}{\sqrt{a}} + \max_{r\in aq\pm t}\left|\tfrac{a!(a-w-r)!}{(a-r)!(a-w)!}(1-q)^w - 1\right|\\
&= \tfrac{2w}{\sqrt{a}} + \max_{r\in aq\pm t}\left|\tfrac{a(a-1)\ldots(a-w+1)}{(a-r)(a-r-1)\ldots(a-w-r+1)}(1-q)^w - 1\right|\\
&\le \tfrac{2w}{\sqrt{a}} + \max\left\{\left|\left(\tfrac{a}{a-aq+t}\gamma\right)^w - 1\right|, \left|\left(\tfrac{a-w+1}{a-aq-t-w+1}\gamma\right)^w - 1\right|\right\}\\
&= \tfrac{2w}{\sqrt{a}} + \max\left\{\left|\left(1-\tfrac{t}{\gamma\,a+t}\right)^w - 1\right|, \left|\left(1+\tfrac{qw-q+t}{\gamma\,a-t-w+1}\right)^w - 1\right|\right\}\\
&\le 2\tfrac{c_1''w}{\sqrt{a}} + \max\left\{\tfrac{tw}{\gamma\,a+t}, \exp\left(\tfrac{qw^2-qw+tw}{\gamma\,a-t-w+1}\right) - 1\right\}\\
&= O(1)\cdot w\sqrt{\ln(a)/(\gamma\,a)}\,.
\end{aligned}$$

For the second part of lemma, we proceed in a similar fashion. By Chernoff bounds, there exists a constant c_2' such that

$$\max\Big(\Pr\Big[|\mathcal{B}(a,1/2)-a/2| \ge c_2'\sqrt{a\ln\tfrac{1}{\delta a}}\Big],\; \Pr\Big[|\mathcal{B}(a,q')-a/2| \ge c_2'\sqrt{a\ln\tfrac{1}{\delta^2 a}}\Big]\Big) \;\le\; \delta^2 a\,,$$

where we have assumed that $\delta a = O(\sqrt{a\ln\frac{1}{\delta^2 a}})$, since otherwise the bound is trivial. Let $s = c_2'\sqrt{a\ln\frac{1}{\delta^2 a}}$. Then,

$$\begin{aligned}
\mathrm{V}(\mathcal{B}(a,1/2), \mathcal{B}(a,q)) &\le \sum_{r\in[a]}\binom{a}{r}|1/2^a - q^r(1-q)^{a-r}|\\
&= 2\delta^2 a + \max_{r\in a/2\pm s}\left|(1+2\delta)^r(1-2\delta)^{a-r} - 1\right|\\
&\le 2\delta^2 a + \max_{u\in\pm s}\left|(1+2\delta)^{1/2+u/a}(1-2\delta)^{1/2-u/a} - 1\right|^a\\
&\le 2\delta^2 a + \max_{u\in\pm s}\left|(1-4\delta^2)^{1/2}\left(1+\frac{4\delta}{1-2\delta}\right)^{u/a} - 1\right|^a\\
&= O(1)\cdot\left(\delta^2 a + \delta\sqrt{a\ln(\delta a)^{-1}}\right)
\end{aligned}$$

□