2016 IEEE 57th Annual Symposium on Foundations of Computer Science

Strong Fooling Sets for Multi-Player Communication with Applications to Deterministic Estimation of Stream Statistics

Amit Chakrabarti and Sagar Kale
Department of Computer Science, Dartmouth College
Hanover, NH, USA
Email: {ac,sag} (at) cs.dartmouth.edu

***Abstract*— We develop a paradigm for studying multi-player deterministic communication, based on a novel combinatorial concept that we call a *strong fooling set*. Our paradigm leads to optimal lower bounds on the per-player communication required for solving multi-player EQUALITY problems in a private-message setting. This in turn gives a very strong—$O(1)$ versus $\Omega(n)$—separation between private-message and one-way blackboard communication complexities.**

Applying our communication complexity results, we show that for deterministic data streaming algorithms, even loose estimations of some basic statistics of an input stream require large amounts of space. For instance, approximating the frequency moment F_k within a factor α requires $\Omega(n/\alpha^{1/(1-k)})$ space for $k<1$ and roughly $\Omega(n/\alpha^{k/(k-1)})$ space for $k>1$. In particular, approximation within any *constant* factor α, however large, requires *linear* space, with the trivial exception of $k=1$. This is in sharp contrast to the situation for randomized streaming algorithms, which can approximate F_k to within $(1\pm\varepsilon)$ factors using $\widetilde{O}(1)$ space for $k\leqslant 2$ and $o(n)$ space for all finite k and all constant $\varepsilon>0$. Previous linear-space lower bounds for deterministic estimation were limited to small factors α, such as $\alpha<2$ for approximating F_0 or F_2.

We also provide certain space/approximation tradeoffs in a deterministic setting for the problems of estimating the empirical entropy of a stream as well as the size of the maximum matching and the edge connectivity of a streamed graph.

I. Introduction

This paper introduces a new paradigm for studying multi-player number-in-hand deterministic communication complexity, uses the paradigm to obtain some new communication lower bounds centered around the EQUALITY problem, and applies these results to derive a number of lower bounds in the data streaming model. These latter results address a very basic topic in the theory of data stream algorithms.

Data stream algorithms offer a convincing demonstration of the power of randomization. Many problems that call for summarization of "big data" admit remarkably efficient *sublinear*-space streaming algorithms, provided such algorithms are allowed randomness. Restricted to determinism, such sublinear-space solutions usually do not exist. This dichotomy between determinism and randomization is among the first lessons one learns in the study of streaming algorithms. Indeed, as shown in the pioneering work of Alon, Matias, and Szegedy [1], one encounters it in the most basic problem of estimating the number of distinct elements in a stream, as well as the more general problem of estimating the stream's frequency moments.

Applying our nuanced understanding of the multi-player EQUALITY problem, we establish very sharp contrasts between the deterministic and randomized space complexities of several important problems for data streams. The problems we study are already known to admit very efficient and accurate randomized estimations. For most of them, past work shows that *accurate* deterministic estimations require linear space. An important takeaway from our present work is that even *loose* deterministic estimations require linear space: randomization is even more crucial to these problems than was previously realized.

Returning to the technical level of communication complexity, our lower-bounding paradigm, which drives these results, should be of independent interest. It sharply separates the "blackboard" and "private-message" communication models, an issue previously highlighted by Gál and Gopalan [2].

Contributions to Data Streaming: Take the case of estimating F_0, the number of distinct elements, in a stream of elements from the universe $[n]:=\{1,\ldots,n\}$. The randomized algorithm of Kane et al. [3] makes one pass over the stream, using $O(\varepsilon^{-2}+\log n)$ bits of space, and computes an estimate that lies in $[(1-\varepsilon)F_0,(1+\varepsilon)F_0]$ with probability $\geqslant\frac{2}{3}$. In contrast, Alon et al. [1] show that, for $\alpha=1.1$, a deterministic estimator that returns a value in $[\alpha^{-1}F_0,\alpha F_0]$ using $O(1)$ passes must use $\Omega(n)$ bits of space. One consequence of our results is that this $\Omega(n)$ space bound holds for *every* constant α, no matter how large.

We go on to show that for every frequency moment F_k, apart from $k=1$, deterministic constant-pass constant-factor estimation requires $\Omega(n)$ space. Stated in full (Section IV), our results give detailed tradeoffs between the space usage, the number of passes, and the quality of estimation for computing the frequency moments and the empirical entropy of a stream, and for estimating graph parameters (Section V) such as the size of a maximum matching and the edge connectivity, when the input is a stream of undirected edges.

This work was supported in part by the National Science Foundation (NSF), under Award CCF-1217375.

0272-5428/16 $31.00 © 2016 IEEE
DOI 10.1109/FOCS.2016.14

Contributions to Communication Complexity: Our results ultimately derive from the phenomenon that, in a communication complexity setting, determining equality between strings is very hard deterministically but easy with randomization. However, this alone does not lead to the strong streaming lower bounds we are claiming: the standard two-player EQUALITY (EQ) problem is insufficient. Instead, we study multi-player EQ problems with certain strong promises (Section III) that greatly separate YES-instances from NO-instances. In one version, each of t players is given an n-bit string: these strings are either all equal or all distinct. Another variant gives them fixed-sized sets, either all equal or with large "spread," i.e., large union size.

To effectively lower-bound the complexities of these promise versions of EQ, we need to consider *discreet protocols*, wherein players can only communicate via private messages. This is in contrast to the better-studied *blackboard* protocols, which allow players to write their messages on a public blackboard. The latter model makes our promise-EQ problems too easy. Gál and Gopalan [2] had previously shown a separation between the discreet and one-way blackboard models, but for an artificial[1] function and with a bespoke proof for the discreet lower bound. Here, we separate these models using what may be *the* natural separating problem (q.v. the end of Section III). More importantly, we introduce an abstract paradigm for studying discreet protocols, based on a novel combinatorial concept that we call a *strong fooling set* (Definition II.7) and a key technical result that we call the "strong fooling set bound" (Lemma II.8).

Other Contributions: In the process of designing reductions from our promise-EQUALITY problems to obtain our data streaming lower bounds, we establish a sharp concentration result for power sums of a collection of independent binomial random variables. This result (Lemma IV.4), though basic-looking and proved via elementary means, appears to be novel. Indeed, other seemingly basic questions about moments of the binomial distribution seem to have been addressed only recently [4].

Other Related Work: The general topic of the communication complexity of EQUALITY problems, in their many variants, has occupied researchers for decades, beginning with Yao's seminal work [5]. Some key results for 2-player EQ address its amortized complexity [6], simultaneous-message complexity [7], [8], [9], direct sum properties [10], [11], round complexity [12], and information complexity [13]. Multi-player versions of EQ, which are interesting only in a number-in-hand setting, have been receiving attention recently, as has the discreet model of communication. Liang and Vaidya [14] consider the basic version of distinguishing all-equal inputs from not-all-equal ones. This is easily shown to require $nt/2$ bits of deterministic communication in total; they show a nontrivial upper bound of cnt for a specific constant $c < 1$. A combinatorial construction of Alon, Moitra, and Sudakov [15] improves this to $c = 1/2+\varepsilon$. Chattopadhyay, Radhakrishnan, and Rudra [16] consider another (also non-promise) variant: ELEMENT-DISTINCTNESS, where the players must distinguish all-distinct inputs from not-all-distinct ones. They study the effect of the communication topology—who may send messages to whom—on the complexities of this, and other, problems. Our lower bounds in this work hold regardless of topology.

In the topology-restricted setting, the *coordinator* model has attracted a lot of study [17], [18], [19]. Building on techniques developed for this model, Woodruff and Zhang [20], [21] gave near-optimal lower bounds for estimating stream statistics and graph parameters in a randomized distributed setting. Separately, they gave strong lower bounds for randomized *exact* computation of such statistics under general topology [22]. This complements our streaming results, which concern *deterministic* weakly-approximate computation.

Our results on estimation of graph parameters contribute to a fast-developing body of literature on streaming and sketching algorithms for graphs. We discuss the relevant background in Section V.

II. Definitions and Preliminaries: Communication Complexity

All our logarithms are to the base 2. The notation $f\colon A \rightsquigarrow B$ denotes a partial function f with domain A and codomain B; formally this is a function $f\colon A \to B \cup \{\star\}$, where $\star \notin B$ is a special do-not-care value. We say that f *is constant* on $A' \subseteq A$ if $|f(A') \setminus \{\star\}| \leqslant 1$.

Let $\mathcal{X}_1, \ldots, \mathcal{X}_t$ be finite sets. Put $\boldsymbol{\mathcal{X}} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_t$. Consider a communication game between players $\text{PLR}_1, \ldots, \text{PLR}_t$ given by a partial function $f\colon \boldsymbol{\mathcal{X}} \rightsquigarrow \mathcal{Z}$. An input for this game is a tuple $\mathbf{x} = (x_1, \ldots, x_t)$. At the start of the game, PLR_i receives the *input fragment* $x_i \in \mathcal{X}_i$, for each i; the players then communicate according to a deterministic protocol Π that ends with an output $\Pi^o(\mathbf{x}) \in \mathcal{Z}$. We say that Π computes f if

$$\forall \mathbf{x} \in \boldsymbol{\mathcal{X}}\colon \ f(\mathbf{x}) \neq \star \implies \Pi^o(\mathbf{x}) = f(\mathbf{x}).$$

In the above context, a *discreet protocol* is a sequence $(s_1, r_1, M_1), \ldots, (s_L, r_L, M_L)$, where the jth tuple describes the action during the jth *step* of the protocol: the sender, PLR_{s_j}, sends a *private* message to the receiver, PLR_{r_j}, using the message function $M_j\colon \{0,1\}^* \times \mathcal{X}_{s_j} \to \{0,1\}^*$. Specifically, if h_j is the concatenation of the messages received by PLR_{s_j} during the first $j-1$ steps, then $M_j(h_j, x_{s_j})$ is the message she sends in the jth step. We require that the range of each M_j be a prefix code; this makes such concatenations self-punctuating. Notice that our discreet protocols are "oblivious" in the sense that the communication pattern (including

[1] The Gál–Gopalan function is arguably artificial as a *communication* problem. It was used in aid of space lower bounds for the (natural) data-streaming problem of estimating the length of the longest increasing subsequence of an input stream.

L, the number of steps) is input-independent. The final (Lth) message is defined to be the output of the protocol, using some canonical map from $\{0,1\}^*$ to $\mathcal{Z}$; the "receiver" r_L is a dummy value.

Let Π be a discreet protocol formalized as above. For an input $\mathbf{x}$ and a player PLR_i, the *local transcript* $\Pi_i(\mathbf{x})$ is defined to be concatenation (in order of occurrence) of the messages received and sent by PLR_i when Π is run on $\mathbf{x}$. The protocol Π is *B-bounded* if, for all i and $\mathbf{x}$, $|\Pi_i(\mathbf{x})| \leqslant B$, i.e., each player sends and receives a total of at most B bits. We define

$$\text{cost}(\Pi) := \max_{i,\mathbf{x}} |\Pi_i(\mathbf{x})| = \min\{B : \Pi \text{ is } B\text{-bounded}\};$$
$$\text{DD}(f) := \min\{\text{cost}(\Pi) : \Pi \text{ computes } f\}.$$

Another, better-studied, kind of protocol for such multi-player communication games is a *blackboard protocol*. Here, players communicate by writing messages on a blackboard visible to all players. The communication history determines which player will write the next bit on the blackboard, and the bit written is a function of this history and the writer's input fragment. Let $\Pi_i^{\text{w}}(\mathbf{x})$ denote the concatenation of all messages written by PLR_i when a blackboard protocol Π is run on an input $\mathbf{x}$. We define $\text{cost}(\Pi) := \max_{i,\mathbf{x}} |\Pi_i^{\text{w}}(\mathbf{x})|$, $\text{cost}^{\text{tot}}(\Pi) := \max_{\mathbf{x}} \left(|\Pi_1^{\text{w}}(\mathbf{x})| + \cdots + |\Pi_t^{\text{w}}(\mathbf{x})|\right)$, and

$$\text{BB}(f) := \min\{\text{cost}(\Pi) : \text{blackboard protocol } \Pi \text{ computes } f\},$$
$$\text{BB}^{\text{tot}}(f) := \min\{\text{cost}^{\text{tot}}(\Pi) : \text{blackboard protocol } \Pi \text{ computes } f\}.$$

Notice that discreet protocols can be thought of as special cases of blackboard protocols: ones that are oblivious and wherein each message is ignored by all players except its receiver. This translates a B-bounded t-player discreet protocol Π into a blackboard protocol Π' with $\text{cost}(\Pi') \leqslant B$ and $\text{cost}^{\text{tot}}(\Pi') \leqslant tB/2$.

We recall the well-known concepts of *rectangles* and *fooling sets* used in the analysis of deterministic two-player protocols [23] and, by an easy extension [2], [24], to deterministic t-player blackboard protocols. A (combinatorial) rectangle in $\boldsymbol{\mathcal{X}}$ is a set $\mathcal{Y}_1 \times \cdots \times \mathcal{Y}_t$, where each $\mathcal{Y}_i \subseteq \mathcal{X}_i$. The *span* of a set $\boldsymbol{\mathcal{F}} \subseteq \boldsymbol{\mathcal{X}}$, denoted $\text{span}(\boldsymbol{\mathcal{F}})$, is the minimal rectangle that includes $\boldsymbol{\mathcal{F}}$. Let Π be some t-player blackboard protocol on input space $\boldsymbol{\mathcal{X}}$. The *transcript* of Π on input $\mathbf{x} \in \boldsymbol{\mathcal{X}}$ is defined to be $(\Pi_1^{\text{w}}(\mathbf{x}), \ldots, \Pi_t^{\text{w}}(\mathbf{x}))$. Two inputs that generate the same transcript are *equivalent*: this relation partitions $\boldsymbol{\mathcal{X}}$ into equivalence classes.

Lemma II.1 (Rectangle Property (folklore))**.** *Each equivalence class of Π is a rectangle in $\boldsymbol{\mathcal{X}}$. In particular, if $\boldsymbol{\mathcal{F}} \subseteq \boldsymbol{\mathcal{X}}$ lies within an equivalence class, then so does $\text{span}(\boldsymbol{\mathcal{F}})$. Consequently, if Π computes a partial function f, then f is constant on $\text{span}(\boldsymbol{\mathcal{F}})$.* ∎

Let $f : \boldsymbol{\mathcal{X}} \rightsquigarrow \mathcal{Z}$ specify a communication game and let $\boldsymbol{\mathcal{F}} \subseteq \boldsymbol{\mathcal{X}}$. We say that $\boldsymbol{\mathcal{F}}$ is a *K-weak-fooling set* for f if, for all $\boldsymbol{\mathcal{F}}' \subseteq \boldsymbol{\mathcal{F}}$ with $|\boldsymbol{\mathcal{F}}'| > K$, f is nonconstant on $\text{span}(\boldsymbol{\mathcal{F}}')$. The standard notion of "fooling set" used in a number of two-player communication complexity lower bounds would be a 1-weak-fooling set under this terminology.

It follows, using Lemma II.1, that if f and $\boldsymbol{\mathcal{F}}$ are as above, and Π is a blackboard protocol that computes f, then no subset of $\boldsymbol{\mathcal{F}}$ of size $> K$ can lie within an equivalence class of Π. So Π must have at least $|\boldsymbol{\mathcal{F}}|/K$ equivalence classes. This gives us the following basic lower bounds, which (we emphasize) we are stating for contrast with what is to follow.

Lemma II.2 (Weak fooling set bound)**.** *Suppose that $f : \boldsymbol{\mathcal{X}} \rightsquigarrow \mathcal{Z}$ specifies a t-player communication game and that f has a K-weak-fooling set $\boldsymbol{\mathcal{F}}$. Then $\text{BB}^{tot}(f) \geqslant \log(|\boldsymbol{\mathcal{F}}|/K)$, and so, $\text{DD}(f) \geqslant (2/t)\log(|\boldsymbol{\mathcal{F}}|/K)$.* ∎

A. The Strong Fooling Set Bound for Discreet Protocols

Having set the stage, we now focus on discreet protocols. We shall analyze such protocols by using more nuanced notions of equivalence of inputs, and then strengthening the above notion of weak fooling sets.

Let Π be a t-player discreet protocol on input space $\boldsymbol{\mathcal{X}}$. Inputs $\mathbf{x}$ and $\mathbf{y}$ are *i-equivalent* if $\Pi_i(\mathbf{x}) = \Pi_i(\mathbf{y})$; they are *equivalent* if they are i-equivalent for all $i \in [t]$. These relations partition the input space $\boldsymbol{\mathcal{X}}$ into the *i-equivalence classes* and the *equivalence classes* of Π, respectively. Clearly, $\Pi^o(\mathbf{x}) = \Pi^o(\mathbf{y})$ whenever $\mathbf{x}$ and $\mathbf{y}$ are equivalent.

Let $\boldsymbol{\mathcal{G}} \subseteq \boldsymbol{\mathcal{X}}$ be nonempty. A *neighborhood within $\boldsymbol{\mathcal{G}}$* is a t-tuple $\mathcal{N} = (\boldsymbol{\mathcal{H}}_1, \ldots, \boldsymbol{\mathcal{H}}_t)$ where each $\boldsymbol{\mathcal{H}}_i \subseteq \boldsymbol{\mathcal{G}}$ and its *core*, defined by $\text{core}(\mathcal{N}) := \boldsymbol{\mathcal{H}}_1 \cap \cdots \cap \boldsymbol{\mathcal{H}}_t$, is nonempty. For an input $\mathbf{x} = (x_1, \ldots, x_t)$, we put $\text{proj}_i \mathbf{x} := x_i$. We extend this notation to sets, defining $\text{proj}_i \boldsymbol{\mathcal{G}} := \{\text{proj}_i \mathbf{x} : \mathbf{x} \in \boldsymbol{\mathcal{G}}\}$. We define the width and the span of the neighborhood $\mathcal{N}$ as follows:

$$\text{wid}(\mathcal{N}) := \min\{|\boldsymbol{\mathcal{H}}_1|, \ldots, |\boldsymbol{\mathcal{H}}_t|\},$$
$$\text{span}(\mathcal{N}) := \text{proj}_1 \boldsymbol{\mathcal{H}}_1 \times \cdots \times \text{proj}_t \boldsymbol{\mathcal{H}}_t.$$

The definitions immediately imply that $\text{core}(\mathcal{N}) \subseteq \text{span}(\mathcal{N}) \subseteq \boldsymbol{\mathcal{X}}$ and that $\text{core}(\mathcal{N}) \subseteq \boldsymbol{\mathcal{G}}$.

We say that a protocol Π is *smooth* on $\mathcal{N}$ if, for each $i \in [t]$, $\boldsymbol{\mathcal{H}}_i$ lies within an i-equivalence class of Π. This notion allows us to generalize the rectangle property from Lemma II.1.

Lemma II.3 (Generalized Rectangle Property)**.** *Let Π be a discreet protocol on input space $\boldsymbol{\mathcal{X}}$. Let $\mathcal{N}$ be a neighborhood within $\boldsymbol{\mathcal{X}}$ such that Π is smooth on $\mathcal{N}$. Then $\text{span}(\mathcal{N})$ lies within an equivalence class of Π. Consequently, if Π computes a partial function f, then f is constant on $\text{span}(\mathcal{N})$.*

To see that the above is a generalization, consider the neighborhood $(\boldsymbol{\mathcal{G}}, \ldots, \boldsymbol{\mathcal{G}})$, where $\boldsymbol{\mathcal{G}}$ lies within an equivalence class of some blackboard protocol.

The following helper lemma will help us prove the generalized rectangle property.

Lemma II.4. *With respect to a protocol Π, suppose that inputs $\mathbf{x} := (x_1, \ldots, x_t)$ and $\mathbf{y} := (y_1, \ldots, y_t)$ are i-equivalent, for some $i \in [t]$. Then $\mathbf{x}$ and $\mathbf{x}' := (x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_t)$ are equivalent.*

Proof: Think of Π as a "virtual" two-player protocol between PLR_i and the rest of the players, combined into a single entity. The local transcript $\Pi_i(\mathbf{x})$ is the transcript of this virtual protocol on input $\mathbf{x}$. Since $\Pi_i(\mathbf{x}) = \Pi_i(\mathbf{y})$, Lemma II.1 (the usual rectangle property) applied to this virtual protocol tells us that $\Pi_i(\mathbf{x}) = \Pi_i(\mathbf{x}')$.

Consider a switch of input from $\mathbf{x}$ to $\mathbf{x}'$, and consider an arbitrary $j \in [t]$ with $j \neq i$. Since $\Pi_i(\mathbf{x}) = \Pi_i(\mathbf{x}')$, the switch affects neither the input fragment nor any messages received by PLR_j. Therefore $\Pi_j(\mathbf{x}) = \Pi_j(\mathbf{x}')$. We conclude that $\mathbf{x}$ and $\mathbf{x}'$ are equivalent. ∎

Proof of Lemma II.3: Fix an input $\mathbf{x} := (x_1, \ldots, x_t) \in \text{core}(\mathcal{N})$. We shall prove that every input in $\text{span}(\mathcal{N})$ is equivalent to $\mathbf{x}$. Let $\mathbf{y} = (y_1, \ldots, y_t) \in \text{span}(\mathcal{N})$. Put

$$\mathbf{x}^i := (y_1, \ldots, y_{i-1}, x_i, \ldots, x_t), \quad \text{for } i \in [t+1].$$

Since Π is smooth on $\mathcal{N}$, for $i \in [t]$, the set $\{\Pi_i(\mathbf{x}') : \mathbf{x}' \in \boldsymbol{\mathcal{H}}_i\}$ is a singleton; let π_i be its lone element.

We shall prove by induction on i that $\mathbf{x}$ and $\mathbf{x}^i$ are equivalent. This will imply that $\mathbf{x}^1 = \mathbf{x}$ and $\mathbf{x}^{t+1} = \mathbf{y}$ are equivalent, as required. The base case, $i = 1$, is trivial. Since $\mathbf{y} \in \text{span}(\mathcal{N})$, we have $y_i \in \text{proj}_i \boldsymbol{\mathcal{H}}_i$ and there is some $\mathbf{z} := (z_1, \ldots, z_{i-1}, y_i, z_{i+1}, \ldots, z_t) \in \boldsymbol{\mathcal{H}}_i$. So $\Pi_i(\mathbf{z}) = \pi_i$. Also, $\mathbf{x} \in \text{core}(\mathcal{N}) \subseteq \boldsymbol{\mathcal{H}}_i$, so $\Pi_i(\mathbf{x}) = \pi_i$. Thus, $\mathbf{z}$ and $\mathbf{x}$ are i-equivalent. Using Lemma II.4, we get that $\mathbf{x}$ and $(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_t) =: \mathbf{x}'$ are equivalent (paste the ith coordinate of $\mathbf{z}$, which is y_i, into that of $\mathbf{x}$). By the inductive hypothesis, $\mathbf{x}$ and $\mathbf{x}^i$ are equivalent. So $\mathbf{x}'$ and $\mathbf{x}^i$ are equivalent, and hence, i-equivalent. Using Lemma II.4 again, we get that $\mathbf{x}^i$ and $(y_1, \ldots, y_i, x_{i+1}, \ldots, x_t) = \mathbf{x}^{i+1}$ are equivalent. This completes the inductive step. ∎

By a pigeonhole argument, a low-cost blackboard protocol entails a large rectangle lying within an equivalence class (a.k.a. "monochromatic"). We prove the following stronger result for discreet protocols.

Lemma II.5. *Let Π be a B-bounded t-player discreet protocol on input space $\boldsymbol{\mathcal{X}}$. Let $\boldsymbol{\mathcal{G}} \subseteq \boldsymbol{\mathcal{X}}$. Then there exists a neighborhood $\mathcal{N}$ within $\boldsymbol{\mathcal{G}}$ such that Π is smooth on $\mathcal{N}$ and $\text{wid}(\mathcal{N}) \geqslant |\boldsymbol{\mathcal{G}}|/(t2^B)$.*

Proof: We use the probabilistic method. We begin with an observation, readily proved by counting.

Observation II.6. *Suppose the finite set S is partitioned into L blocks and $s \in_R S$ is picked uniformly at random. For every real $A > 0$, $\Pr[s \text{ lies in a block of size} < |S|/(AL)] < 1/A$.*

For each $i \in [t]$, the i-equivalence classes of Π partition $\boldsymbol{\mathcal{X}}$, and hence $\boldsymbol{\mathcal{G}}$, into at most 2^B blocks. Pick $\mathbf{x} \in_R \boldsymbol{\mathcal{G}}$ uniformly at random. Put $[\![\mathbf{x}]\!]_i := \{\mathbf{y} \in \boldsymbol{\mathcal{G}} : \mathbf{y} \text{ is } i\text{-equivalent to } \mathbf{x}\}$. Then $\mathcal{N}_\mathbf{x} := ([\![\mathbf{x}]\!]_1, \ldots, [\![\mathbf{x}]\!]_t)$ is a neighborhood within $\boldsymbol{\mathcal{G}}$ (its core is nonempty because it contains $\mathbf{x}$) on which Π is smooth. By the above observation and a union bound, we have

$$\begin{aligned}\Pr\left[\text{wid}(\mathcal{N}_\mathbf{x}) \geqslant \frac{|\boldsymbol{\mathcal{G}}|}{t2^B}\right] &= 1 - \Pr\left[\exists i : \ \left|[\![\mathbf{x}]\!]_i\right| < \frac{|\boldsymbol{\mathcal{G}}|}{t2^B}\right] \\ &\geqslant 1 - \sum_{i=1}^{t} \Pr\left[\left|[\![\mathbf{x}]\!]_i\right| < \frac{|\boldsymbol{\mathcal{G}}|}{t2^B}\right] \\ &> 1 - \sum_{i=1}^{t} \frac{1}{t} = 0.\end{aligned}$$ ∎

We now strengthen our earlier notion of weak fooling sets and prove our main technical lemma, a stronger communication lower bound in terms of these "strong" fooling sets.

Definition II.7. Let $f : \boldsymbol{\mathcal{X}} \rightsquigarrow \mathcal{Z}$ specify a communication game and let $\boldsymbol{\mathcal{F}} \subseteq \boldsymbol{\mathcal{X}}$. We say that $\boldsymbol{\mathcal{F}}$ is a *K-fooling set* for f if, for every neighborhood $\mathcal{N}$ within $\boldsymbol{\mathcal{F}}$, $\text{wid}(\mathcal{N}) > K \implies f$ is nonconstant on $\text{span}(\mathcal{N})$.

Lemma II.8 (Strong fooling set bound)**.** *Suppose that $f : \boldsymbol{\mathcal{X}} \rightsquigarrow \mathcal{Z}$ specifies a t-player communication game and that f has a K-fooling set $\boldsymbol{\mathcal{F}}$. Then*

$$\text{DD}(f) \geqslant \log \frac{|\boldsymbol{\mathcal{F}}|}{tK}.$$

Proof: Let Π be a B-bounded discreet protocol for f. By Lemma II.5, there exists a neighborhood $\mathcal{N}$ within $\boldsymbol{\mathcal{F}}$ with $\text{wid}(\mathcal{N}) \geqslant |\boldsymbol{\mathcal{F}}|/(t2^B)$ such that Π is smooth on $\mathcal{N}$. By Lemma II.3, f is constant on $\text{span}(\mathcal{N})$. In view of Definition II.7, we must have $\text{wid}(\mathcal{N}) \leqslant K$, which implies $|\boldsymbol{\mathcal{F}}|/(t2^B) \leqslant K$.

The lemma follows by rearranging the latter inequality. ∎

III. Three Lower Bounds for Multi-Player Equality

We shall now use our strong fooling set bound to analyze two promise versions of the EQUALITY problem, as alluded to in Section I. Each is given by a partial function of the form $f : \mathcal{X}^t \rightsquigarrow \{0,1\}$, where $\mathcal{X} = \{0,1\}^n$. In the Equal-vs-Distinct problem, the goal is to distinguish the case when all players hold the same n-bit string from the case when they hold distinct strings. This is formalized by the following partial function:

$$\text{EQ-DIST}_{n,t}(x_1, \ldots, x_t) = \begin{cases} 1, & \text{if } x_1 = \cdots = x_t, \\ 0, & \text{if } x_i \neq x_j \\ & \quad \text{whenever } 1 \leqslant i < j \leqslant t. \end{cases}$$

In the Equal-vs-Spread problem, each player receives a $\lceil \beta n \rceil$-subset of $[n]$ and they must distinguish the case when all of these subsets are equal from the case when these subsets are sufficiently spread out. Formally, we interpret $\mathcal{X}$

as the power set $2^{[n]}$ and use the following partial function:

$$\text{EQ-SPRD}^{\beta,\gamma}_{n,t}(x_1,\ldots,x_t) = \begin{cases} 1, & \text{if for } i\in[t], |x_i| = \lceil \beta n\rceil \\ & \text{and } x_1 = \cdots = x_t\,, \\ 0, & \text{if for } i\in[t], |x_i| = \lceil \beta n\rceil \\ & \text{and } |x_1\cup\cdots\cup x_t| \geqslant \gamma n. \end{cases}$$

This problem is nontrivial when the parameters satisfy $0 < \beta < \gamma \leqslant 1$ and $\gamma n > \lceil \beta n\rceil$.

We now give strong, essentially optimal, communication lower bounds for discreet deterministic protocols that solve these problems. In each case, when the problem's input space is $\mathcal{X}^t$ and $x\in\mathcal{X}$ is an input fragment, we denote the input $(x,x,\ldots,x)\in\mathcal{X}^t$ by $x^{\otimes t}$.

The following observation will aid some of our estimations.

Observation III.1. *For all integral values $0\leqslant k\leqslant \ell\leqslant n$,* $\binom{n}{k}\Big/\binom{\ell}{k} = \frac{n}{\ell}\cdot\frac{n-1}{\ell-1}\cdot\cdots\cdot\frac{n-k+1}{\ell-k+1} \geqslant \left(\frac{n}{\ell}\right)^k$.

Theorem III.2 (Lower bound for Equal-vs-Distinct)**.** $\mathrm{DD}(\text{EQ-DIST}_{n,t}) \geqslant n - 2\log t$.

Proof: Put $f := \text{EQ-DIST}_{n,t}$. We claim that the set $\mathcal{F} := \{x^{\otimes t} : x\in\{0,1\}^n\}$ is a $(t-1)$-fooling set for f. Indeed, let $\mathcal{N} = (\mathcal{H}_1,\ldots,\mathcal{H}_t)$ be a neighborhood within $\mathcal{F}$ such that $\mathrm{wid}(\mathcal{N}) > t-1$. Since $f(\mathbf{x}) = 1$ for all $\mathbf{x}\in\mathcal{F}$, our earlier observations that $\varnothing \neq \mathrm{core}(\mathcal{N}) \subseteq \mathcal{F}$ and $\mathrm{core}(\mathcal{N}) \subseteq \mathrm{span}(\mathcal{N})$ imply that f takes the value 1 at some point in $\mathrm{span}(\mathcal{N})$. On the other hand, consider the point $\mathbf{y} = (y_1,\ldots,y_t)$ constructed by the following procedure:

- Having chosen $y_1,\ldots,y_{i-1}$, where $1\leqslant i\leqslant t$, choose an arbitrary fragment y_i such that $y_i^{\otimes t}\in\mathcal{H}_i\setminus\{y_1^{\otimes t},\ldots,y_{i-1}^{\otimes t}\}$. This choice is possible because $|\mathcal{H}_i| \geqslant \mathrm{wid}(\mathcal{N}) > t-1 \geqslant i-1$.

This ensures that $\mathbf{y}\in\mathrm{span}(\mathcal{N})$ and $f(\mathbf{y}) = 0$. Therefore f is nonconstant on $\mathrm{span}(\mathcal{N})$, proving the claim.

Applying Lemma II.8, we get $\mathrm{DD}(f) \geqslant \log\frac{|\mathcal{F}|}{t(t-1)} \geqslant \log\frac{2^n}{t^2} = n - 2\log t$. ∎

Theorem III.3 (Lower bound for Equal-vs-Spread)**.** *For all values $0<\beta<\gamma\leqslant 1$ and sufficiently large n, if $t\geqslant\gamma n$, then* $\mathrm{DD}(\text{EQ-SPRD}^{\beta,\gamma}_{n,t}) \geqslant (\beta\log(1/\gamma))n - \log t$.

Proof: Put $f := \text{EQ-SPRD}^{\beta,\gamma}_{n,t}$ and $w = \lceil\beta n\rceil$. We claim that the set $\mathcal{F} := \{x^{\otimes t} : x\in\{0,1\}^n, |x| = w\}$ is a $\binom{\lfloor\gamma n\rfloor}{w}$-fooling set for f. Indeed, let $\mathcal{N} = (\mathcal{H}_1,\ldots,\mathcal{H}_t)$ be a neighborhood within $\mathcal{F}$ such that $\mathrm{wid}(\mathcal{N}) > \binom{\lfloor\gamma n\rfloor}{w}$. Since $f(\mathbf{x}) = 1$ for all $\mathbf{x}\in\mathcal{F}$, as in the proof of Theorem III.2, f takes the value 1 at some point in $\mathrm{span}(\mathcal{N})$. On the other hand, consider the point $\mathbf{y} = (y_1,\ldots,y_t)$ constructed by the following procedure:

- Having chosen $y_1,\ldots,y_{i-1}$, where $1\leqslant i\leqslant t$, let $U_{i-1} := y_1\cup\cdots\cup y_{i-1}$. If $|U_{i-1}|\geqslant\gamma n$, then choose an arbitrary fragment y_i such that $y_i^{\otimes t}\in\mathcal{H}_i$.
- Otherwise, let $B_i := \{x\in\{0,1\}^n : x\subseteq U_{i-1}, |x| = w\}$. Choose an arbitrary fragment $y_i\notin B_i$ such that $y_i^{\otimes t}\in\mathcal{H}_i$. This choice is possible because

$$|B_i| = \binom{|U_{i-1}|}{w} \leqslant \binom{\lfloor\gamma n\rfloor}{w} < \mathrm{wid}(\mathcal{N}) \leqslant |\mathcal{H}_i|\,.$$

Then $\mathbf{y}\in\mathrm{span}(\mathcal{N})$. Notice that $|U_i| > |U_{i-1}|$ whenever the second case occurs while choosing y_i. We make $t\geqslant\gamma n$ choices in total, which ensures that $|U_t|\geqslant\gamma n$, implying that $f(\mathbf{y}) = 0$. Therefore f is nonconstant on $\mathrm{span}(\mathcal{N})$, proving the claim.

Applying Lemma II.8 and Observation III.1,

$$\begin{aligned} \mathrm{DD}(f) &\geqslant \log\frac{|\mathcal{F}|}{t\binom{\lfloor\gamma n\rfloor}{w}} \\ &= \log\frac{\binom{n}{w}}{t\binom{\lfloor\gamma n\rfloor}{w}} \\ &\geqslant \log\frac{\left(\frac{n}{\lfloor\gamma n\rfloor}\right)^w}{t} \\ &\geqslant w\log\frac{1}{\gamma} - \log t\,. \end{aligned}$$ ∎

Theorem III.3 gives a lower bound for "large" t. In particular, if the spread threshold γn is to be $\Omega(n)$, then we have to take $t = \Omega(n)$ in order to apply the theorem. We now give an alternate lower bound for EQ-SPRD that holds in a different parameter regime, where t could be "small."

Theorem III.4. *For all values $t\geqslant 2$, $\beta > 0$, $\gamma = \beta t(1 - e\beta t) > \beta$, and sufficiently large integral n, we have* $\mathrm{DD}(\text{EQ-SPRD}^{\beta,\gamma}_{n,t}) \geqslant 2e\beta^2 n - 2\log t - \Theta(1)$.

We prove this by reducing from EQ-DIST, using a coding-style argument. Define an (r,s,n)-packing to be set system $\mathcal{C}\subseteq 2^{[n]}$ such that (i) for all $A\in\mathcal{C}$, $|A| = s$, and (ii) for all $A,B\in\mathcal{C}$ with $A\neq B$, $|A\cap B|\leqslant r$. We shall need the following bound, which can be inferred from Proposition 2.1 in Erdös, Frankl, and Füredi [25].

Lemma III.5. *For all values $0\leqslant r\leqslant s\leqslant n$, there exists an (r,s,n)-packing $\mathcal{C}$ with $|\mathcal{C}| \geqslant \binom{n}{r}\Big/\binom{s}{r}^2$.* ∎

Proof of Theorem III.4: Let $\mathcal{C}$ be a maximum-sized (r,s,n)-packing, with $s = \lceil\beta n\rceil$ and $r = 2\lceil e\beta s\rceil$. By Lemma III.5, Observation III.1, and the estimation $\binom{s}{r}\leqslant (es/r)^r$, we have $|\mathcal{C}| \geqslant \binom{n}{r}\Big/\binom{s}{r}^2 \geqslant \binom{n}{s}^r\left(\frac{r}{es}\right)^r = \left(\frac{nr}{es^2}\right)^r$.

By our choice of parameters,

$$\begin{aligned} r\log\frac{nr}{es^2} &\geqslant r\log\frac{2e\beta sn}{es^2} \\ &= r\log\frac{2\beta n}{\lceil\beta n\rceil} \\ &= r\left(1 - \Theta\left(\frac{1}{n}\right)\right) \\ &= 2e\beta^2 n - \Theta(1)\,. \end{aligned}$$

Therefore, we can find an injection from $\{0,1\}^N$ to $\mathcal{C}$ provided $N \leqslant 2e\beta^2 n - \Theta(1)$. We choose the largest possible N satisfying this bound and fix such an injection.

To solve EQ-DIST$_{N,t}$, the players encode their respective input fragments using this injection and then solve EQ-SPRD$^{\beta,\gamma}_{n,t}$ on the encoded input. Recall that $\gamma = \beta t(1 - e\beta t)$. We now argue that this reduction is correct. A 1-input for EQ-DIST$_{N,t}$ is, rather obviously, mapped to a 1-input for EQ-SPRD$^{\beta,\gamma}_{n,t}$. Suppose a 0-input for EQ-DIST$_{N,t}$ maps to $(x_1,\ldots,x_t)$. Then $x_1,\ldots,x_t$ are distinct sets in $\mathcal{C}$. By the packing property,

$$\begin{aligned}|x_1 \cup \cdots \cup x_t| &\geqslant ts - \tbinom{t}{2} r \\ &= ts - t(t-1)\lceil e\beta s\rceil \\ &\geqslant ts - t^2 e\beta s \\ &= \lceil \beta n\rceil t(1 - e\beta t) \\ &\geqslant \gamma n,\end{aligned}$$

where the second inequality holds once n (and hence, s) is sufficiently large. So EQ-SPRD$^{\beta,\gamma}_{n,t}(x_1,\ldots,x_t) = 0$.

Theorem III.2 gives DD(EQ-SPRD$^{\beta,\gamma}_{n,t}$) $\geqslant$ DD(EQ-DIST$_{N,t}$) $\geqslant N - 2\log t \geqslant 2e\beta^2 n - 2\log t - \Theta(1)$. ∎

We conclude this section with some commentary on the lower bounds that we have just shown.

Optimality: Suppose that $t = \text{poly}(n)$. The trivial protocol, where PLR$_1$ sends his input to PLR$_2$, shows that DD(EQ-DIST$_{n,t}$) $\leqslant n+1$. This shows that Theorem III.2 is tight up to lower order terms. Another trivial protocol, where players efficiently encode $\lceil \beta n\rceil$-subsets of $[n]$ and each sends his input to the "next" player, shows that DD(EQ-SPRD$^{\beta,\gamma}_{n,t}$) $\leqslant 2\log\binom{n}{\lceil \beta n\rceil} \leqslant 2H(\beta)n$, for $\beta < 1/2$. Thus, Theorem III.3 and Theorem III.4 are both asymptotically tight in their dependence on n. Moreover, when $\gamma/\beta = O(1)$, Theorem III.3 is also tight in its dependence on β.

Separation Between Models: Our lower bounds also demonstrate a separation between one-way blackboard protocols and discreet protocols, an issue highlighted by Gál and Gopalan [2]. Consider the following blackboard protocol for EQ-DIST$_{n,t}$. Partition $[n]$ into $t-1$ blocks, each of size at most $\lceil n/(t-1)\rceil$. For each $j \in [t-1]$, PLR$_j$ announces his input fragment restricted to the jth block, provided all blocks from 1 to $j-1$ of his input fragment agree with previously announced blocks (if not, PLR$_j$ ends the protocol with output 0). If this protocol reaches PLR$_t$, he knows the entirety of PLR$_{t-1}$'s input fragment. He outputs 1 if his own fragment agrees with this, and outputs 0 otherwise. Based on the promise, this is a correct protocol for EQ-DIST$_{n,t}$. This protocol has max-cost $O(n/t)$, which is $O(1)$ when $t = \Theta(n)$. Yet, DD(EQ-DIST$_{n,\Theta(n)}$) $= n - \Theta(\log n)$.

IV. Stream Statistics

In the data stream model, an input consists of m elements of $[n]$ arriving in the form of a stream σ that may be read in one or more passes by a *streaming* algorithm. Formally, σ is a sequence $(a_1, a_2, \ldots, a_m)$, where each $a_j \in [n]$. The stream σ defines a frequency vector $\mathbf{f} = \mathbf{f}(\sigma) = (f_1,\ldots,f_n)$, where $f_i = |\{j \in [m] : a_j = i\}|$ for each $i \in [n]$. Stream statistics problems involve computing some function of $\mathbf{f}$, e.g., frequency moments and empirical entropy, which we consider in this section. The kth frequency moment and the empirical entropy are defined, respectively, as $F_k(\mathbf{f}) := \sum_{i\in[n]} f_i^k$ and $\text{ENT}(\mathbf{f}) = m^{-1}\sum_{i\in[n]} f_i \log(m/f_i)$. Note that $F_1(\mathbf{f}) = m$ and $F_0(\mathbf{f})$ is the number of distinct elements in σ.

Our focus is on *deterministic* streaming algorithms. An s-space p-pass streaming algorithm is one that uses $s = s(m,n)$ bits of space to process its input, which it reads in $p = p(m,n)$ passes. Consider such an algorithm $\mathcal{A}$. We denote its output, on input σ, by $\mathcal{A}(\sigma)$. By splitting σ into $t = t(m,n)$ sub-streams, we obtain a communication problem for which $\mathcal{A}$ naturally gives rise to a $2ps$-bounded discreet protocol, for every t. For a real quantity $\alpha = \alpha(m,n) \geqslant 1$, we say that $\mathcal{A}$ is an α-estimator for a quantity $Q(\sigma)$ if

$$\exists\, \kappa, \lambda \geqslant 1 \ (\kappa\lambda \leqslant \alpha \text{ and } \forall\, \sigma \ (\kappa^{-1} Q(\sigma) \leqslant \mathcal{A}(\sigma) \leqslant \lambda Q(\sigma)) \ ).$$

The main results in this section are lower bounds that trade off α against the product ps, for algorithms estimating frequency moments and empirical entropy. For F_k ($k \neq 1$), when $\alpha = O(1)$, we obtain the strongest possible bound: $ps = \Omega(n)$. We can also give simple estimators (upper bounds) for F_k and ENT (Section IV-C); for moments of "lower" order ($0 \leqslant k < 1$) these simple estimators show that our lower bounds are tight even in their dependence on α.

Throughout this section, asymptotic expressions may hide constants depending on k. For readability, we ignore floors and ceilings. This does not affect our (asymptotic) bounds.

A. Warm-Up: Distinct Elements and Basic Lower Bounds for Other Moments

We shall first obtain lower bounds for estimating F_k ($k \neq 1$) by reduction from EQ-SPRD$^{\beta,\gamma}_{n,t}$, for certain values of t, β, γ and invoking Theorems III.3 and III.4 to lower-bound DD(EQ-SPRD$^{\beta,\gamma}_{n,t}$). The bound we obtain is tight for F_0 (the distinct elements problem). Our bounds here are also tight for all k when $\alpha = O(1)$. In the next section, we use a more complicated analysis to obtain a tighter tradeoff between ps and α.

Theorem IV.1. *For each $k \in [0,1)$, every deterministic s-space p-pass α-estimator for F_k satisfies $ps = \Omega(\max\{n^{1-k}/\alpha, n/\alpha^{2/(1-k)}\})$. In particular, at $k = 0$ we have $ps = \Omega(n/\alpha)$.*

Proof: Let $\mathbf{x} = (x_1,\ldots,x_t)$ be an input for EQ-SPRD$^{\beta,\gamma}_{n,t}$. For each $j \in [t]$, PLR$_j$ turns his input fragment $x_j \in \{0,1\}^n$ into the stream of indices j where $x_j = 1$. The concatenations of the t such streams has frequency vector $\mathbf{f} = x_1 + \cdots + x_t$.

Note that

$$\text{EQ-SPRD}^{\beta,\gamma}_{n,t}(\mathbf{x}) = 1 \implies F_k(\mathbf{f}) = \beta t^k n; \quad (1)$$

$$\text{EQ-SPRD}^{\beta,\gamma}_{n,t}(\mathbf{x}) = 0 \implies F_0(\mathbf{f}) \geqslant \gamma n. \quad (2)$$

Also, $F_k(\mathbf{f}) \geqslant F_0(\mathbf{f})$. Thus, an α-estimator for F_k can separate these two cases provided $\gamma/(\beta t^k) > \alpha$. If such an estimator uses s bits of space and p passes then, as argued at the start of Section IV, we have $2ps \geqslant \text{DD}(\text{EQ-SPRD}^{\beta,\gamma}_{n,t})$. It remains to invoke a suitable communication lower bound.

Set $\gamma = 1/e$, $t = \gamma n$, and $\beta = \gamma/(\alpha t^k) - 1/n$. This ensures that $\gamma/(\beta t^k) > \alpha$ and optimizes the lower bound from Theorem III.3, giving $ps \geqslant \frac{1}{2}((n/e)^{1-k}(\log e)/\alpha - \log n) = \Omega(n^{1-k}/\alpha)$.

We could instead apply Theorem III.4 to estimate $\text{DD}(\text{EQ-SPRD}^{\beta,\gamma}_{n,t})$. We set $t = (2\alpha)^{1/(1-k)}$ and $\beta < 1/(2et)$; the theorem then requires $\gamma = \beta t(1 - e\beta t)$. Note that $\gamma/(\beta t^k) > \alpha$, as required. Applying the theorem gives $ps \geqslant e\beta^2 n - \log t - \Theta(1) = \Omega(n/t^2) = \Omega(n/\alpha^{2/(1-k)})$. ∎

For frequency moments F_k of "higher" order ($k > 1$), we can follow a similar proof template, but it takes more work to analyze the effect of the "spread" case in the Equal-vs-Spread problem. In particular, we use the following technical lemma whose proof (given in the full version of this paper) is based on convexity and Karamata's inequality.

Lemma IV.2. *Let $g\colon \mathbb{R} \to \mathbb{R}$ be a nondecreasing convex function with $g(0) = 0$ and let $\mathbf{f} \in \{0,1,\ldots,t\}^n$ where $t \geqslant 2$. Suppose that $F_1(\mathbf{f}) = m$, $F_0(\mathbf{f}) \geqslant r$, and $rt \geqslant m$. Then*

$$\sum_{i=1}^{n} g(f_i) \leqslant \ell g(t) + (r - \ell) g(1),$$

where $\ell = \lceil (m-r)/(t-1) \rceil$. ∎

Theorem IV.3. *For each $k > 1$, every deterministic s-space p-pass α-estimator for F_k has $ps = \Omega(n/\alpha^{2k/(k-1)})$.*

Proof sketch: We use Theorem III.4, setting $t = (2\alpha)^{1/(k-1)}$ and $\beta < 1/(3t^k)$. When $\text{EQ-SPRD}^{\beta,\gamma}_{n,t}(\mathbf{x}) = 1$, by eq. (1), $F_k(\mathbf{f}) = \beta t^k n$. When $\text{EQ-SPRD}^{\beta,\gamma}_{n,t}(\mathbf{x}) = 0$, eq. (2) tells us that $F_0(\mathbf{f})$ is "large," at least γn. Meanwhile, $F_1(\mathbf{f}) = \beta t n$. By Lemma IV.2, these facts imply that $F_k(\mathbf{f}) < 2\beta t n$, which shows a gap of $(\beta t^k n)/(2\beta t n) = \alpha$. ∎

B. Stronger Lower Bounds for Frequency Moments and Empirical Entropy

We shall now improve the lower bounds in Theorems IV.1 and IV.3, obtaining a tighter dependence on α. From a data-streaming perspective, the two new lower bounds for F_k estimation given in this section—one for $k > 1$ and one for $k < 1$—are the main theorems of this paper.

The improvements ultimately stem from sufficiently sharp concentration bounds for power sums of binomial random variables. Let $Y_1, \ldots, Y_n$ be independent random variables, each with binomial distribution $\mathcal{B}(t,q)$. Let $Z = Z(n,t,q,k) = Y_1^k + \cdots + Y_n^k$ be the kth power sum of this collection. In the full version of this paper [26], we prove the following upper tail bound for Z.

Lemma IV.4. *For each $k > 1$, there exist $b, c > 0$ such that the following holds. For each $q \in (0, 1/(2e^2))$, there exist integers n_0 and t_0 such that, for all $n \geqslant n_0$ and $t \geqslant t_0$,*

$$\Pr[Z > bq^k t^k n] \leqslant \exp\left(-\frac{cq^k tn}{(\log\log(1/q))^2}\right).$$

This tail bound does not follow from Chernoff-Hoeffding and Azuma-Hoeffding inequalities [27]; those give a much weaker upper bound of the form $\exp(-\Theta(n))$. Indeed, even a bound of $\exp(-\Theta(tn))$ would not be strong enough for our purposes. We need to understand how the coefficient in front of tn depends on q, and this seems to require a delicate partitioning of the large deviation event.

Our improved lower bounds for F_k estimation are obtained by reducing from $\text{EQ-DIST}_{N,t}$, using what we shall call an *F_k-separating mapping*, defined as follows. A function $R\colon \{0,1\}^N \to \{0,1\}^n$ is said to be F_k-separating with parameters (t,α) if

$$\frac{\min\{F_k(t \cdot R(x)) : x \in \{0,1\}^N\}}{\max\left\{F_k\left(\sum_{i\in[t]} R(x_i)\right) : \text{EQ-DIST}_{N,t}(x_1,\ldots,x_t) = 0\right\}} > \alpha, \quad \text{when } k > 1, \quad (3)$$

and

$$\frac{\min\left\{F_k\left(\sum_{i\in[t]} R(x_i)\right) : \text{EQ-DIST}_{N,t}(x_1,\ldots,x_t) = 0\right\}}{\max\{F_k(t \cdot R(x)) : x \in \{0,1\}^N\}} > \alpha, \quad \text{when } k < 1. \quad (4)$$

Suppose that such a mapping exists and that we have an s-space p-pass α-estimator for F_k over the universe $[n]$. Then, following the template from Section IV-A, a team of t players can solve $\text{EQ-DIST}_{N,t}$ by mapping their inputs to $\{0,1\}^n$ via R and converting the mapped inputs to streams over $[n]$. Applying Theorem III.2, we obtain $2ps \geqslant \text{DD}(\text{EQ-DIST}_{N,t}) = N - 2\log t$.

Theorem IV.5. *For each $k > 1$ and $\alpha \geqslant 1$, every deterministic s-space p-pass α-estimator for F_k satisfies $ps = \Omega(n/(\alpha^{k/(k-1)}(\log\log\alpha)^2))$.*

Proof: We follow the outline above. It remains to prove the existence of an F_k-separating mapping for a large enough $N = N(n,t,\alpha)$ and a not-too-large t.

We construct the mapping R at random, as follows. Generate a random $2^N \times n$ matrix whose entries are independent Bernoulli random variables, each equal to 1 with probability q. We shall fix q later. Then, for each $x \in \{0,1\}^N$, define $R(x)$ to be the xth row of this matrix. We claim that with positive probability both of the following events occur:

$$\mathcal{E}_1 := \left\{\min\{F_k(t \cdot R(x)) : x \in \{0,1\}^N\} \geqslant qt^k n/2\right\},$$

$$\mathcal{E}_2 := \left\{\max\{F_k(R(x_1) + \cdots + R(x_t)) : \text{EQ-DIST}_{N,t}(x_1,\ldots,x_t) = 0\} \leqslant bq^k t^k n\right\}.$$

Noting that $\mathbb{E}|R(x)| = qn$ for each $x \in \{0,1\}^N$, a standard Chernoff bound followed by a union bound gives $\Pr[\neg\mathcal{E}_1] = \Pr\left[\exists x \in \{0,1\}^N : |R(x)| < qn/2\right] \leqslant 2^N \exp(-qn/8)$. On the other hand, for each choice of distinct $x_1, \ldots, x_t \in \{0,1\}^N$, the quantity $F_k(R(x_1) + \cdots + R(x_t))$ is the kth power sum of n independent binomial random variables. By Lemma IV.4 and a union bound,

$$\begin{aligned}\Pr[\neg\mathcal{E}_2] &\leqslant \binom{2^N}{t} \exp\left(-\frac{cq^k tn}{(\log\log(1/q))^2}\right) \\ &\leqslant 2^{Nt} \exp\left(-\frac{cq^k tn}{(\log\log(1/q))^2}\right),\end{aligned}$$

for all large enough n and t. Therefore, setting $N = c'q^k n/(\log\log(1/q))^2$ for an appropriate constant c' ensures $\Pr[\neg\mathcal{E}_1 \vee \neg\mathcal{E}_2] < 1$.

Thus, there exists a specific R at which both $\mathcal{E}_1$ and $\mathcal{E}_2$ occur. For this R, the left-hand side of eq. (3) is at least $(qt^k n/2)/(bq^k t^k n) = 1/(2bq^{k-1})$. We set $q = O(1/\alpha^{1/(k-1)})$ so that this ratio exceeds α. Then eq. (3) is satisfied and R is F_k-separating.

Taking $t = n$ (say) gives us the bound $ps = \Omega(N - \log t) = \Omega(n/(\alpha^{k/(k-1)}(\log\log\alpha)^2))$. ∎

Analogous proofs, given in the full version of this paper [26] handle the estimation of frequency moments of "lower" order and empirical entropy.

Theorem IV.6. *For each $k \in [0,1)$ and $\alpha \geqslant 1$, α-estimating F_k requires $ps = \Omega(n/\alpha^{1/(1-k)})$.* ∎

Theorem IV.7. *For each $\varepsilon > 0$ and $\alpha \in [1, o(\log n)]$, α-estimating $\textsc{ent}(\mathbf{f})$ requires $ps = \Omega(n^{1/((1+\varepsilon)\alpha)})$.* ∎

C. Some Simple Upper Bounds

On the algorithmic side, we can *achieve* a tradeoff between space and approximation for F_k, in one deterministic pass. The algorithms are straightforward: the main idea is to coarsen the universe $[n]$ by bucketing (details appear in the full version of this paper [26]). Additionally, we have the following theorem giving an upper bound for estimating $\textsc{ent}(\mathbf{f})$.

Theorem IV.8. *There is a deterministic $O(\log m)$-space 2-pass $(1+\log n)$-estimator for the empirical entropy of a stream.* ∎

Theorem IV.9. *For integers $p \geqslant 1$, and reals $k \geqslant 0$ and $\alpha \geqslant 1$, there is a family of deterministic p-pass α-estimators for F_k, with the following guarantees on their space usage, s.*

- *When $k = 0$, we have $ps = \lceil n/\alpha \rceil + O(\log n)$.*
- *When $0 < k < 1$, we have $ps = O(n \log m/\alpha^{1/(1-k)})$.*
- *When $k = 1$, at $p = 1$ we have $s \leqslant \lceil \log m \rceil$, trivially.*
- *When $k > 1$, we have $ps = O(n \log m/\alpha^{1/(k-1)})$.* ∎

V. Graph Streams

A number of important data stream problems are graph-theoretic. The input graph $G_n = (V,E)$, where $|V| = n$, is described as a stream of edges (the edge-arrival model, our default). It is usually interesting, and nontrivial, to achieve space $\widetilde{O}(n)$ for most standard graph computations [28]. We focus on two particular graph problems: maximum matching size estimation (MMSE), described next, and a variant of edge connectivity, described in Section V-A.

The MMSE problem asks for an estimate of the number of edges in a maximum cardinality matching (MCM). For this problem, it is also natural to consider the vertex-arrival model, where the input is a bipartite graph $G_n = (V_1, V_2, E)$, with $|V_2| = n$ and $|V_1| = O(n)$, and the stream lists each vertex $u \in V_1$ with all its neighbors in V_2. This potentially makes an algorithm's task easier, so lower bounds proven in this model are stronger. We prove lower bounds in the vertex-arrival model for s-space p-pass α-estimators for MMSE; our bounds trade off α against the product ps.

Previous Work: The MCM problem asks to output (the edges of) a large matching. Outputting a *maximal* matching gives a 2-approximation for MCM in $\widetilde{O}(n)$ space; nothing better using $o(n^2)$ space is known. Kapralov [29] showed that a one-pass randomized MCM algorithm achieving approximation better than $e/(e-1)$ must use $n^{1+\Omega(1/\log\log n)}$ space, even in the vertex-arrival model. On the other hand, a $(1+\varepsilon)$-approximation is possible using $O(n\,\mathrm{polylog}\,n)$ space and $O_\varepsilon(1)$ passes, for every constant ε [28], [30]. Turning to MMSE, Kapralov et al. [31] gave a $O(\mathrm{polylog}\,n)$-space $O(\mathrm{polylog}\,n)$-estimator for randomly-ordered edge streams. Esfandiari et al. [32] gave $o(n)$-space estimators for graphs with bounded arboricity. They also give $\Omega(n)$ space lower bound for deterministic one-pass $(3/2-\varepsilon)$-estimators for MMSE; this bound should be compared with that in Theorem V.2. For constant α, no $o(n)$ space α-estimator is known.

Dobzinski, Nisan, and Oren [33] consider the bipartite MCM problem in the simultaneous message (SM) model; each player gets the neighbor set for a vertex in V_1 and they send a message of size at most ℓ to the coordinator who has to output a large matching. They show that for deterministic α-approximation protocols, $\ell = \Omega(n/\alpha)$. Our communication lower-bound techniques for discreet protocols also extend to SM protocols with essentially no change. So the lower bound obtained in Theorem V.2 below also applies to SM protocols and discreet protocols for MMSE defined appropriately as a communication problem. Thus it generalizes the deterministic lower bound by Dobzinski et al. from a star communication topology (for SM protocols) to arbitrary topology. This, in particular, yields data streaming lower bounds.

Our Results: First, we define a variant of the Equal-vs-Spread problem that we call Equal-vs-Distinct-Representatives. There are $t = \lfloor \gamma n \rfloor$ players. Each player

receives a $\lceil \beta n \rceil$-subset of $[n]$ (where $\beta < \gamma$) and they must distinguish the case when all of these subsets are equal from the case when each player can pick a representative element from her subset so that these representatives are distinct. Formally, we use the following partial function:

$$\text{EQ-DR}^{\beta}_{n,t}(x_1,\ldots,x_t) = \begin{cases} 1, & \text{if for } i \in [t], |x_i| = \lceil \beta n \rceil \\ & \quad \text{and } x_1 = \cdots = x_t, \\ 0, & \text{if for } i \in [t], |x_i| = \lceil \beta n \rceil \\ & \quad \text{and } \exists g \colon [t] \to \bigcup_{i \in [t]} x_i \\ & \quad \text{such that } g \text{ is injective} \\ & \quad \text{and } g(i) \in x_i \text{ for } i \in [t]. \end{cases}$$

It is not hard to see that the proof of Theorem III.3 applies to an analysis of this problem as well, after the substitution $\gamma = t/n$, due to the way in which $\mathbf{y} \in \text{span}(\mathcal{N})$ is constructed in that proof. This leads to the following lower bound.

Theorem V.1. *For all values $0 < \beta < 1$, $\varepsilon > 0$, and sufficiently large n, if $(\beta + \varepsilon)n \leqslant t < n$, then we have* $\text{DD}(\text{EQ-DR}^{\beta}_{n,t}) \geqslant (\beta \log(n/t))n - \log t$. ∎

Though the following theorem is stated for streaming algorithms for MMSE, it is a special case of a more general communication result as noted earlier.

Theorem V.2. *For a deterministic s-space p-pass α-estimator for* MMSE, *$ps \geqslant ((n/e\alpha)(\log e) - \log n)/2$.*

Proof: We reduce from $\text{EQ-DR}^{\beta}_{n,t}$ setting t and β later. The theorem can be proved in the vertex arrival model with $V_1 = \{u_1,\ldots,u_t\}$ and $V_2 = [n]$. Note that the lower bounds for vertex-arrival model also apply for the edge-arrival model. For $i \in [t]$, PLR_i adds edges $\{\{u_i, j\} : j \in x_i\}$. Call the resulting graph G_n. When $\text{EQ-DR}^{\beta}_{n,t}(x_1,\ldots,x_t) = 1$, an MCM in G_n has size βn. When $\text{EQ-DR}^{\beta}_{n,t}(x_1,\ldots,x_t) = 0$, an MCM in G_n has size t, because the definition of $\text{EQ-DR}^{\beta}_{n,t}$ guarantees existence of an injective mapping g from $[t]$ to $[n]$. So, if β and t are such that $t/\beta n > \alpha$, then a deterministic s-space p-pass α-estimator for MMSE can be used to give a $2ps$-bounded discreet protocol $\text{EQ-DR}^{\beta}_{n,t}$. To get the desired bound, we set $t = n/e$ and $\beta = 1/(\alpha e) - 1/n$, and use Theorem V.1. ∎

A. Edge Connectivity

We divert from multi party to two party communication complexity in this section. The dynamic graph connectivity problem XCONN is as follows. There are two players, Alice and Bob, who get inputs E_A and E_B which are sets of edges on the vertex set $[n]$. For two sets S and T, denote by $S \oplus T$ the set $(S \cup T) \setminus (S \cap T)$. Alice and Bob communicate to determine whether the graph $E_A \oplus E_B$ is connected.

We reduce $\text{EQ}_{n^2/4}$ to XCONN, where EQ is the well-known two-party equality problem. Alice adds a complete graph on $[n/2]$, Bob adds a complete graph on $[n] \setminus [n/2]$, and they encode the inputs for $\text{EQ}_{n^2/4}$ within the edges in $[n/2] \times ([n] \setminus [n/2])$. In case of equality, XCONN will evaluate to false, otherwise XCONN will evaluate to true; hence, communication complexity of XCONN is at least $n^2/4$.

There is a randomized protocol for XCONN. Alice can send Bob the sketch for connectivity given by Ahn, Guha, and McGregor [34] of size $O(n \log^3 n)$. Bob can solve XCONN using this sketch. This separates the randomized and deterministic communication complexity of XCONN.

By using error correcting codes (ECC), we can show that even the following version of XCONN with a strong promise is hard. Alice and Bob get inputs E_A and E_B with the promise that the graph $(E_A \cup E_B) \setminus (E_A \cap E_B)$ is disconnected or $(n/2-1)$-connected, i.e., at least $n/2-1$ edges need to be removed to disconnect it. We reduce from EQ_{N^2} where $N = \Omega(n)$. We use a binary ECC of size 2^{N^2}, block length $n^2/4$, and distance $n/2-1$. By Shannon's construction, we can construct such an ECC with $N = \Omega(n)$. Then we use the same construction as in the reduction from $\text{EQ}_{n^2/4}$ to XCONN to get E_A and E_B. In case of equality, $E_A \oplus E_B$ will be disconnected (no edge from $[n/2]$ to $[n] \setminus [n/2]$). In case of inequality, there will be at least $n/2-1$ edges from $[n/2]$ to $[n] \setminus [n/2]$. Since $E_A \oplus E_B$ has a complete graph within $[n/2]$ and within $[n] \setminus [n/2]$, it is at least $(n/2-1)$-connected. Hence, the communication complexity of strong-promise version of XCONN is at least $N^2 = \Omega(n^2)$.

REFERENCES

[1] N. Alon, Y. Matias, and M. Szegedy, "The space complexity of approximating the frequency moments," *J. Comput. Syst. Sci.*, vol. 58, no. 1, pp. 137–147, 1999, preliminary version in *Proc. 28th Annual ACM Symposium on the Theory of Computing*, pages 20–29, 1996.

[2] A. Gál and P. Gopalan, "Lower bounds on streaming algorithms for approximating the length of the longest increasing subsequence," in *Proc. 48th Annual IEEE Symposium on Foundations of Computer Science*, 2007, pp. 294–304.

[3] D. M. Kane, J. Nelson, and D. P. Woodruff, "On the exact space complexity of sketching and streaming small norms," in *Proc. 21st Annual ACM-SIAM Symposium on Discrete Algorithms*, 2010, pp. 1161–1178.

[4] Árpád Bényi and S. M. Manago, "A recursive formula for moments of a binomial distribution," *The College Mathematics Journal*, vol. 36, no. 1, pp. 68–72, 2005.

[5] A. C. Yao, "Some complexity questions related to distributive computing," in *Proc. 11th Annual ACM Symposium on the Theory of Computing*, 1979, pp. 209–213.

[6] T. Feder, E. Kushilevitz, M. Naor, and N. Nisan, "Amortized communication complexity," *SIAM J. Comput.*, vol. 24, no. 4, pp. 736–750, 1995, preliminary version in *Proc. 32nd Annual IEEE Symposium on Foundations of Computer Science*, pages 239–248, 1991.

[7] A. Ambainis, "Communication complexity in a 3-computer model," *Algorithmica*, vol. 16, no. 3, pp. 298–301, 1996.

[8] L. Babai and P. G. Kimmel, "Randomized simultaneous messages: Solution of a problem of Yao in communication complexity," in *Proc. 12th Annual IEEE Conference on Computational Complexity*, 1997, pp. 239–246.

[9] R. Bottesch, D. Gavinsky, and H. Klauck, "Equality, revisited," in *Proc. 40th International Symposium on Mathematical Foundations of Computer Science*, 2015, pp. 127–138.

[10] A. Chakrabarti, Y. Shi, A. Wirth, and A. C. Yao, "Informational complexity and the direct sum problem for simultaneous message complexity," in *Proc. 42nd Annual IEEE Symposium on Foundations of Computer Science*, 2001, pp. 270–278.

[11] M. Molinaro, D. Woodruff, and G. Yaroslavtsev, "Beating the direct sum theorem in communication complexity with implications for sketching," in *Proc. 24th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2013, p. to appear.

[12] J. Brody, A. Chakrabarti, R. Kondapally, D. P. Woodruff, and G. Yaroslavtsev, "Certifying equality with limited interaction," in *Proc. 18th International Workshop on Randomization and Approximation Techniques in Computer Science*, 2014, pp. 545–581.

[13] M. Braverman, "Interactive information complexity," in *Proc. 44th Annual ACM Symposium on the Theory of Computing*, 2012, pp. 505–524.

[14] G. Liang and N. H. Vaidya, "Multiparty equality function computation in networks with point-to-point links," in *Proc. 18th International Colloquium on Structural Information and Communication Complexity*, 2011, pp. 258–269.

[15] N. Alon, A. Moitra, and B. Sudakov, "Nearly complete graphs decomposable into large induced matchings and their applications," in *Proc. 44th Annual ACM Symposium on the Theory of Computing*, 2012, pp. 1079–1090.

[16] A. Chattopadhyay, J. Radhakrishnan, and A. Rudra, "Topology matters in communication," in *Proc. 55th Annual IEEE Symposium on Foundations of Computer Science*, 2014, pp. 631–640.

[17] G. Cormode, S. Muthukrishnan, and K. Yi, "Algorithms for distributed functional monitoring," in *Proc. 19th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2008, pp. 1076–1085.

[18] C. J. Arackaparambil, J. Brody, and A. Chakrabarti, "Functional monitoring without monotonicity," in *Proc. 36th International Colloquium on Automata, Languages and Programming*, 2009, pp. 95–106.

[19] J. M. Phillips, E. Verbin, and Q. Zhang, "Lower bounds for number-in-hand multiparty communication complexity, made easy," in *Proc. 23rd Annual ACM-SIAM Symposium on Discrete Algorithms*, 2012, pp. 486–501.

[20] D. P. Woodruff and Q. Zhang, "Tight bounds for distributed functional monitoring," in *Proc. 43rd Annual ACM Symposium on the Theory of Computing*, 2012, pp. 941–960.

[21] ——, "An optimal lower bound for distinct elements in the message passing model," in *Proc. 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2014, pp. 718–733.

[22] ——, "When distributed computation is communication expensive," in *Proc. 27th International Symposium on Distributed Computing*, 2013, pp. 16–30.

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

[24] F. Ergün and H. Jowhari, "On distance to monotonicity and longest increasing subsequence of a data stream," in *Proc. 19th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2008, pp. 730–736.

[25] P. Erdős, P. Frankl, and Z. Füredi, "Families of finite sets in which no set is covered by the union of r others," *Israel J. Math.*, vol. 51, pp. 79–89, 1985.

[26] A. Chakrabarti and S. Kale, "Strong fooling sets for multi-player communication with applications to deterministic estimation of stream statistics," ECCC, Tech. Rep. TR16-111, 2016.

[27] N. Alon and J. H. Spencer, *The Probabilistic Method*. New York, NY: Wiley-Interscience, 2000.

[28] J. Feigenbaum, S. Kannan, A. McGregor, S. Suri, and J. Zhang, "On graph problems in a semi-streaming model," *Theor. Comput. Sci.*, vol. 348, no. 2–3, pp. 207–216, 2005, preliminary version in *Proc. 31st International Colloquium on Automata, Languages and Programming*, pages 531–543, 2004.

[29] M. Kapralov, "Better bounds for matchings in the streaming model," in *Proc. 24th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2013, pp. 1679–1697.

[30] K. J. Ahn and S. Guha, "Linear programming in the semi-streaming model with application to the maximum matching problem," *Inf. Comput.*, vol. 222, pp. 59–79, 2013.

[31] M. Kapralov, S. Khanna, and M. Sudan, "Approximating matching size from random streams," in *Proc. 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2014, pp. 734–751.

[32] H. Esfandiari, M. T. Hajiaghayi, V. Liaghat, M. Monemizadeh, and K. Onak, "Streaming algorithms for estimating the matching size in planar graphs and beyond," in *Proc. 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2015, pp. 1217–1233.

[33] S. Dobzinski, N. Nisan, and S. Oren, "Economic efficiency requires interaction," in *Proc. 46th Annual ACM Symposium on the Theory of Computing*, 2014, pp. 233–242.

[34] K. J. Ahn, S. Guha, and A. McGregor, "Analyzing graph structure via linear measurements," in *Proc. 23rd Annual ACM-SIAM Symposium on Discrete Algorithms*, 2012, pp. 459–467.