Lower Bounds for Gap-Hamming-Distance and Consequences for Data Stream Algorithms

Joshua Brody and Amit Chakrabarti

Dartmouth College

24th CCC, 2009, Paris
### Counting Distinct Elements in a Data Stream

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>14</th>
<th>1</th>
<th>3</th>
<th>9</th>
<th>9</th>
<th>4</th>
<th>2</th>
<th>1</th>
<th>5</th>
<th>2</th>
<th>3</th>
<th>6</th>
</tr>
</thead>
</table>

Input: Stream of integers $\sigma = < a_1, \ldots, a_m >$

Output: $F_0 :=$ number of distinct elements in $\sigma$
### Counting Distinct Elements in a Data Stream

| 3 | 14 | 1 | 3 | 9 | 9 | 4 | 2 | 1 | 5 | 2 | 3 | 6 |

Input: Stream of integers $\sigma = \langle a_1, \ldots, a_m \rangle$

Output: $F_0 :=$ number of distinct elements in $\sigma$

Goal: Minimize space used to compute $F_0$
Previous Streaming Results

Frequency Moments: \( F_k = \sum_{i=1}^{n} \text{freq}(i)^k \)  
[Alon-Matias-Szegedy '96]
Frequency Moments: $F_k = \sum_{i=1}^{n} \text{freq}(i)^k$ [Alon-Matias-Szegedy '96]

- $\Omega(n)$ space unless randomization and approximation used
- Upper, lower bounds for randomized algorithms that approximate $F_k$
- Spawned lots of research, won 2005 Gödel Prize
Previous Streaming Results

Frequency Moments: $F_k = \sum_{i=1}^{n} \text{freq}(i)^k$ [Alon-Matias-Szegedy '96]

- $\Omega(n)$ space unless randomization and approximation used
- Upper, lower bounds for randomized algorithms that approximate $F_k$
- Spawned lots of research, won 2005 Gödel Prize

One-pass, randomized, $\varepsilon$-approximate: $\left| \frac{\text{output}}{\text{answer}} - 1 \right| \leq \varepsilon$
Previous Streaming Results

Frequency Moments: $F_k = \sum_{i=1}^{n} \text{freq}(i)^k$  

- $\Omega(n)$ space unless randomization and approximation used
- Upper, lower bounds for randomized algorithms that approximate $F_k$
- Spawned lots of research, won 2005 Gödel Prize

One-pass, randomized, $\varepsilon$-approximate: $\left| \frac{\text{output}}{\text{answer}} - 1 \right| \leq \varepsilon$

Status as of Jan 2009:

- Space upper bound: $\tilde{O}(\varepsilon^{-2})$
- Space lower bound: $\tilde{\Omega}(\varepsilon^{-2})$
- Also hold for other problems, e.g. empirical entropy

Do multiple passes help?
**Previous Streaming Results**

**Frequency Moments:** \( F_k = \sum_{i=1}^{n} \text{freq}(i)^k \)  

- \( \Omega(n) \) space unless randomization and approximation used
- Upper, lower bounds for randomized algorithms that approximate \( F_k \)
- Spawned lots of research, won 2005 Gödel Prize

One-pass, randomized, \( \varepsilon \)-approximate: 
\[ \left| \frac{\text{output}}{\text{answer}} - 1 \right| \leq \varepsilon \]

**Status as of Jan 2009:**
- Space upper bound: \( \tilde{O}(\varepsilon^{-2}) \)
- Space lower bound: \( \tilde{\Omega}(\varepsilon^{-2}) \)
- Also hold for other problems, e.g. empirical entropy

Do multiple passes help? If not, why not?

Joshua Brody
The Gap-Hamming-Distance Problem

Input: Alice gets $x \in \{0, 1\}^n$, Bob gets $y \in \{0, 1\}^n$.

Output:

- $\text{GHD}(x, y) = 1$ if $\Delta(x, y) > \frac{n}{2} + \sqrt{n}$
- $\text{GHD}(x, y) = 0$ if $\Delta(x, y) < \frac{n}{2} - \sqrt{n}$

Problem: Design randomized, constant error protocol to solve this

Cost: Worst case number of bits communicated

\[
\begin{align*}
x &= 0 \ 1 \ 0 \ 0 \ 1 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 1 \\
y &= 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 1 \ 1 \ 0 \ 0 \ 1 \\
n &= 12; \quad \Delta(x, y) = 3 \in [6 - \sqrt{12}, 6 + \sqrt{12}] \end{align*}
\]
The Reductions

E.g., Distinct Elements (Other problems: similar)

Alice:  $x \mapsto \sigma = \langle (1, x_1), (2, x_2), \ldots, (n, x_n) \rangle$
Bob:  $y \mapsto \tau = \langle (1, y_1), (2, y_2), \ldots, (n, y_n) \rangle$

Notice:  $F_0(\sigma \circ \tau) = n + \Delta(x, y) = \begin{cases} < \frac{3n}{2} - \sqrt{n}, & \text{or} \\ > \frac{3n}{2} + \sqrt{n}. \end{cases}$

Set $\varepsilon = \frac{1}{\sqrt{n}}$. 

Joshua Brody
Communication to Streaming

$p$-pass streaming algorithm $\implies (2p - 1)$-round communication protocol

messages = memory contents of streaming algorithm

And Thus

Previous results [Indyk-Woodruff’03], [Woodruff’04], [C.-Cormode-McGregor’07]:

- For one-round protocols, $R \to (GHD) = \Omega(n)$
- Implies the $\tilde{\Omega}(\varepsilon^{-2})$ streaming lower bounds
Communication to Streaming

$p$-pass streaming algorithm $\implies (2p - 1)$-round communication protocol

messages = memory contents of streaming algorithm

And Thus

Previous results [Indyk-Woodruff’03], [Woodruff’04], [C.-Cormode-McGregor’07]:

- For one-round protocols, $R \rightarrow (\text{GHD}) = \Omega(n)$
- Implies the $\widetilde{\Omega}(\varepsilon^{-2})$ streaming lower bounds

Key open questions:

- What is the unrestricted randomized complexity $R(\text{GHD})$?
- Better algorithm for Distinct Elements (or $F_k$, or $H$) using two passes?
Previous Results (Communication):

- One-round (one-way) lower bound: \( R^{-}(\text{GHD}) = \Omega(n) \) [Woodruff’04]
- Simplification, clever reduction from \textsc{Index} [Jayram-Kumar-Sivakumar]
- Multi-round case: \( R(\text{GHD}) = \Omega(\sqrt{n}) \) [Folklore]
Our Results

Previous Results (Communication):

• One-round (one-way) lower bound: $R^→(\text{GHD}) = \Omega(n)$ [Woodruff’04]

• Simplification, clever reduction from INDEX [Jayram-Kumar-Sivakumar]

  Hard distribution “contrived,” non-uniform

• Multi-round case: $R(\text{GHD}) = \Omega(\sqrt{n})$ [Folklore]
Our Results

Previous Results (Communication):

- One-round (one-way) lower bound: $R^{-}(\text{GHD}) = \Omega(n)$ [Woodruff’04]
- Simplification, clever reduction from INDEX [Jayram-Kumar-Sivakumar]
  Hard distribution “contrived,” non-uniform
- Multi-round case: $R(\text{GHD}) = \Omega(\sqrt{n})$ [Folklore]
  Reduction from DISJOINTNESS using “repetition code”
  Hard distribution again far from uniform
Our Results

Previous Results (Communication):

- One-round (one-way) lower bound: $R(\text{GHD}) = \Omega(n)$ [Woodruff'04]
- Simplification, clever reduction from INDEX [Jayram-Kumar-Sivakumar]
  Hard distribution “contrived,” non-uniform
- Multi-round case: $R(\text{GHD}) = \Omega(\sqrt{n})$ [Folklore]
  Reduction from DISJOINTNESS using “repetition code”
  Hard distribution again far from uniform

What we show:

- Theorem 1: $\Omega(n)$ lower bound for any $O(1)$-round protocol
  Holds under uniform distribution
Previous Results (Communication):

- One-round (one-way) lower bound: $R \leftarrow (\text{GHD}) = \Omega(n)$ \cite{Woodruff'04}
- Simplification, clever reduction from INDEX \cite{Jayram-Kumar-Sivakumar}
  Hard distribution “contrived,” non-uniform
- Multi-round case: $R(\text{GHD}) = \Omega(\sqrt{n})$ \cite{Folklore}
  Reduction from DISJOINTNESS using “repetition code”
  Hard distribution again far from uniform

What we show:

- Theorem 1: $\Omega(n)$ lower bound for any $O(1)$-round protocol
  Holds under uniform distribution
- Theorem 2: one-round, deterministic: $D \leftarrow (\text{GHD}) = n - \Theta(\sqrt{n \log n})$
- Theorem 3: $R \leftarrow (\text{GHD}) = \Omega(n)$ (simpler proof, uniform distrib)
  (independently proved by \cite{Woodruff'09})
Technique: Round Elimination

**Base Case Lemma:** There is no “nice” 0-round GHD protocol.

**Round Elimination Lemma:** If there is a “nice” \(k\)-round GHD protocol, then there is a “nice” \(((k - 1))\)-round GHD protocol.
**Technique: Round Elimination**

**Base Case Lemma:** There is no 0-round GHD protocol with error $\varepsilon < \frac{1}{2}$.

**Round Elimination Lemma:** If there is a “nice” $k$-round GHD protocol, then there is a “nice” $(k - 1)$-round GHD’ protocol.
Technique: Round Elimination

**Base Case Lemma:** There is no 0-round GHD protocol with error $\varepsilon < \frac{1}{2}$.

**Round Elimination Lemma:** If there is a “nice” $k$-round GHD protocol, then there is a “nice” $(k - 1)$-round GHD’ protocol.

- The $(k - 1)$-round protocol will be solving a “simpler” problem
- Parameters degrade with each round elimination step
The problem:

\[
GHD_{c,n}(x, y) = \begin{cases} 
1, & \text{if } \Delta(x, y) \geq n/2 + c\sqrt{n}, \\
0, & \text{if } \Delta(x, y) \leq n/2 - c\sqrt{n}, \\
\star, & \text{otherwise.}
\end{cases}
\]
The problem:

\[
GHD_{c,n}(x, y) = \begin{cases} 
1, & \text{if } \Delta(x, y) \geq n/2 + c\sqrt{n}, \\
0, & \text{if } \Delta(x, y) \leq n/2 - c\sqrt{n}, \\
\ast, & \text{otherwise.}
\end{cases}
\]

Hard input distribution:

\[\mu_{c,n} : \text{uniform over } (x, y) \text{ such that } |\Delta(x, y) - n/2| \geq c\sqrt{n}\]
The problem:

\[ G_{c,n}(x, y) = \begin{cases} 
1, & \text{if } \Delta(x, y) \geq n/2 + c\sqrt{n}, \\
0, & \text{if } \Delta(x, y) \leq n/2 - c\sqrt{n}, \\
\ast, & \text{otherwise.}
\end{cases} \]

Hard input distribution:

\[ \mu_{c,n} : \text{uniform over } (x, y) \text{ such that } |\Delta(x, y) - n/2| \geq c\sqrt{n} \]

Protocol assumptions (eventually, will lead to contradiction):

- Deterministic \( k \)-round protocol for \( G_{c,n} \)
- Each message is \( s \ll n \) bits
- Error probability \( \leq \varepsilon \), under distribution \( \mu_{c,n} \)
Round Elimination

**Main Construction:** Given $k$-round protocol $\mathcal{P}$ for $GHD_{c,n}$, construct $(k - 1)$-round protocol $\mathcal{Q}$ for $GHD_{c',n'}$
**Main Construction:** Given $k$-round protocol $\mathcal{P}$ for $\text{GHD}_{c,n}$, construct $(k-1)$-round protocol $\mathcal{Q}$ for $\text{GHD}_{c',n'}$

First Attempt:

- Fix Alice’s first message $m$ in $\mathcal{P}$, suitably
Round Elimination

Main Construction: Given $k$-round protocol $\mathcal{P}$ for $\text{GHD}_{c,n}$, construct $(k-1)$-round protocol $\mathcal{Q}$ for $\text{GHD}_{c',n'}$

First Attempt:

- Fix Alice’s first message $m$ in $\mathcal{P}$, suitably
- Protocol $\mathcal{Q}_1$:
  - Input: $x', y' \in \{0, 1\}^A$ where $A \subseteq [n]$, $|A| = n'$
  - Extend $x' \rightarrow x$ s.t. Alice sends $m$ on input $x$
  - Extend $y' \rightarrow y$ uniformly at random
  - Output $\mathcal{P}(x, y)$; Note: first message unnecessary
Main Construction: Given $k$-round protocol $\mathcal{P}$ for $\text{GHD}_{c,n}$, construct $(k - 1)$-round protocol $\mathcal{Q}$ for $\text{GHD}_{c',n'}$

First Attempt:

• Fix Alice’s first message $m$ in $\mathcal{P}$, suitably

• Protocol $\mathcal{Q}_1$:
  - Input: $x', y' \in \{0, 1\}^A$ where $A \subseteq [n]$, $|A| = n'$
  - Extend $x' \rightarrow x$ s.t. Alice sends $m$ on input $x$
  - Extend $y' \rightarrow y$ uniformly at random
  - Output $\mathcal{P}(x, y)$; Note: first message unnecessary

• Errors: $\mathcal{Q}_1$ correct, unless
  - $BAD_1$: $\text{GHD}_{c',n'}(x', y') \neq \text{GHD}_{c,n}(x, y)$.
  - $BAD_2$: $\text{GHD}_{c,n}(x, y) \neq \mathcal{P}(x, y)$. 

Joshua Brody
Main Construction: Given $k$-round protocol $\mathcal{P}$ for $GHD_{c,n}$, construct $(k - 1)$-round protocol $\mathcal{Q}$ for $GHD_{c',n'}$

First Attempt:

- Fix Alice’s first message $m$ in $\mathcal{P}$, suitably
- Protocol $\mathcal{Q}_1$:
  - Input: $x', y' \in \{0, 1\}^A$ where $A \subseteq [n]$, $|A| = n'$
  - Extend $x' \rightarrow x$ s.t. Alice sends $m$ on input $x$ (why possible?)
  - Extend $y' \rightarrow y$ uniformly at random
  - Output $\mathcal{P}(x, y)$; Note: first message unnecessary
- Errors: $\mathcal{Q}_1$ correct, unless
  - $BAD_1$: $GHD_{c',n'}(x', y') \neq GHD_{c,n}(x, y)$.
  - $BAD_2$: $GHD_{c,n}(x, y) \neq \mathcal{P}(x, y)$.
Fixing Alice’s first message:

- Call $x$ good if $\Pr_y[\mathcal{P}(x, y) \neq \text{GHD}_{c,n}(x, y)] \leq 2\varepsilon$
  
  Then $\#\{\text{good } x\} \geq 2^{n-1}$ (Markov)

- Let $M = M_m = \{\text{good } x : \text{Alice sends } m \text{ on input } x\}$.

- Fix $m$ to maximize $|M|$; then $|M| \geq 2^{n-1-s}$. 
Fixing Alice’s first message:

- Call \( x \) good if \( \Pr_y[\mathcal{P}(x, y) \neq \text{GHD}_{c,n}(x, y)] \leq 2\varepsilon \)
  
  Then \( \#\{\text{good } x\} \geq 2^{n-1} \) (Markov)

- Let \( M = M_m = \{\text{good } x : \text{Alice sends } m \text{ on input } x\} \).

- Fix \( m \) to maximize \( |M| \); then \( |M| \geq 2^{n-1-s} \).

Shattering:

- Say \( S \subseteq \{0, 1\}^n \) shatters \( A \subseteq [n] \) if \( \#\{x|_A : x \in S\} = 2^{|A|} \)

- \( \text{VCD}(S) := \text{size of largest } A \text{ shattered by } S \)
VC-Dimension

Fixing Alice’s first message:

• Call $x$ good if $\Pr_y[\mathcal{P}(x, y) \neq \text{GHD}_{c,n}(x, y)] \leq 2\varepsilon$
  Then $\#\{\text{good } x\} \geq 2^{n-1}$ (Markov)
• Let $M = M_m = \{\text{good } x : \text{Alice sends } m \text{ on input } x\}$.
• Fix $m$ to maximize $|M|$; then $|M| \geq 2^{n-1-s}$.

Shattering:

• Say $S \subseteq \{0, 1\}^n$ shatters $A \subseteq [n]$ if $\#\{x|_A : x \in S\} = 2^{|A|}$
• $\text{VCD}(S) := \text{size of largest } A \text{ shattered by } S$

Sauer’s Lemma: If $\text{VCD}(S) < \alpha n$ then $|S| < 2^{nH(\alpha)}$. 
Fixing Alice’s first message:

- Call $x$ good if $\Pr_y[\mathcal{P}(x,y) \neq \text{GHD}_{c,n}(x,y)] \leq 2\varepsilon$
  
  Then $\#\{\text{good } x\} \geq 2^{n-1}$ (Markov)

- Let $M = M_m = \{\text{good } x : \text{Alice sends } m \text{ on input } x\}$.

- Fix $m$ to maximize $|M|$; then $|M| \geq 2^{n-1-s}$.

Shattering:

- Say $S \subseteq \{0,1\}^n$ shatters $A \subseteq [n]$ if $\#\{x|_A : x \in S\} = 2^{|A|}$

- $\text{VCD}(S) := \text{size of largest } A \text{ shattered by } S$

**Sauer’s Lemma**: If $\text{VCD}(S) < \alpha n$ then $|S| < 2^{nH(\alpha)}$.

**Corollary**: $\text{VCD}(M) \geq n' := n/3$ (Because $s \ll n$)
Fixing Alice’s first message:

- Call $x$ good if $\Pr_y[P(x, y) \neq \text{GHD}_{c,n}(x, y)] \leq 2\varepsilon$

Then $\#\{\text{good } x\} \geq 2^{n-1}$ (Markov)

- Let $M = M_m = \{\text{good } x : \text{Alice sends } m \text{ on input } x\}$

- Fix $m$ to maximize $|M|$; then $|M| \geq 2^{n-1-s}$.

Shattering:

- Say $S \subseteq \{0, 1\}^n$ shatters $A \subseteq [n]$ if $\#\{x | A : x \in S\} = 2^{|A|}$

- $\text{VCD}(S) := \text{size of largest } A \text{ shattered by } S$

**Sauer’s Lemma:** If $\text{VCD}(S) < \alpha n$ then $|S| < 2^{nH(\alpha)}$.

**Corollary:** $\text{VCD}(M) \geq n' := n/3$ (Because $s \ll n$)

Extend $x' \rightarrow x$: pick $x \in M$ such that $x' = x |_A$
The First Bad Event

Recall $BAD_1$: $\text{GHD}_{c',n'}(x', y') \neq \text{GHD}_{c,n}(x, y)$.

Notation: $x = x' \circ \bar{x}$, $y = y' \circ \bar{y}$, $n = n' + \bar{n}$. 
The First Bad Event

Recall $BAD_1$: $\text{GHD}_{c',n'}(x', y') \neq \text{GHD}_{c,n}(x, y)$.

Notation: $x = x' \circ \bar{x}$, $y = y' \circ \bar{y}$, $n = n' + \bar{n}$.

Definition: $\bar{x}, \bar{y}$ nearly orthogonal if $|\Delta(\bar{x}, \bar{y}) - \bar{n}/2| < 2\sqrt{\bar{n}}$. 
The First Bad Event

Recall $BAD_1$: $\text{GHD}_{c',n'}(x', y') \neq \text{GHD}_{c,n}(x, y)$.

Notation: $x = x' \circ \bar{x}$, $y = y' \circ \bar{y}$, $n = n' + \bar{n}$.

Definition: $\bar{x}, \bar{y}$ nearly orthogonal if $|\Delta(\bar{x}, \bar{y}) - \bar{n}/2| < 2\sqrt{n}$.

Lemma: $\Pr[\bar{x}, \bar{y} \text{ nearly orthogonal}] > 7/8$. (Binom distrib tail)
Recall $BAD_1$: $\text{GHD}_{c',n'}(x', y') \neq \text{GHD}_{c,n}(x, y)$.

Notation: $x = x' \circ \bar{x}$, $y = y' \circ \bar{y}$, $n = n' + \bar{n}$.

Definition: $\bar{x}, \bar{y}$ nearly orthogonal if $|\Delta(\bar{x}, \bar{y}) - \bar{n}/2| < 2\sqrt{\bar{n}}$.

**Lemma:** $\Pr[\bar{y}[\bar{x}, \bar{y} \text{ nearly orthogonal}]] > 7/8$. (Binom distrib tail)

**Lemma:** If $\bar{x}, \bar{y}$ nearly orthogonal and $c' \geq 2c$, then

- $\text{GHD}_{c',n'}(x', y') = 1 \implies \text{GHD}_{c,n}(x, y) = 1$
- $\text{GHD}_{c',n'}(x', y') = 0 \implies \text{GHD}_{c,n}(x, y) = 0$
The First Bad Event

Recall $BAD_1$: $\text{GHD}_{c',n'}(x', y') \neq \text{GHD}_{c,n}(x, y)$.

Notation: $x = x' \circ \bar{x}$, $y = y' \circ \bar{y}$, $n = n' + \bar{n}$.

Definition: $\bar{x}, \bar{y}$ nearly orthogonal if $|\Delta(\bar{x}, \bar{y}) - \bar{n}/2| < 2\sqrt{\bar{n}}$.

Lemma: $\Pr[\bar{x}, \bar{y} \text{ nearly orthogonal}] > 7/8$. (Binom distrib tail)

Lemma: If $\bar{x}, \bar{y}$ nearly orthogonal and $c' \geq 2c$, then

- $\text{GHD}_{c',n'}(x', y') = 1 \implies \text{GHD}_{c,n}(x, y) = 1$
- $\text{GHD}_{c',n'}(x', y') = 0 \implies \text{GHD}_{c,n}(x, y) = 0$

Corollary: $\Pr[BAD_1] < 1/8$. 
Recall $BAD_2$: $GHD_{c,n}(x, y) \neq P(x, y)$.

Bounding $\Pr[BAD_2]$ is subtle:

- $x$ is good, so $\Pr[P \text{ errs} \mid x] \leq 2\varepsilon$
  - But this requires $(x, y) \sim \mu_{c,n}$
- Random extension $(x', y') \rightarrow (x, y)$ is not $\sim \mu_{c,n}$.
Recall \( \text{BAD}_2: \ GHD_{c,n}(x, y) \neq \mathcal{P}(x, y) \).

Bounding \( \Pr[\text{BAD}_2] \) is subtle:

- \( x \) is good, so \( \Pr[\mathcal{P} \text{ errs} \mid x] \leq 2\varepsilon \)
  - But this requires \( (x, y) \sim \mu_{c,n} \)
- Random extension \( (x', y') \rightarrow (x, y) \) is not \( \sim \mu_{c,n} \).

Lemma:
\[
\Pr[\text{BAD}_2] = O(\varepsilon).
\]
Recall $BAD_2$: $\text{GHD}_{c,n}(x, y) \neq \mathcal{P}(x, y)$.

Bounding $\Pr[BAD_2]$ is subtle:

- $x$ is good, so $\Pr[\mathcal{P} \text{ errs } | x] \leq 2\varepsilon$
  - But this requires $(x, y) \sim \mu_{c,n}$
- Random extension $(x', y') \rightarrow (x, y)$ is not $\sim \mu_{c,n}$.
The Second Bad Event

Recall $BAD_2$: $\text{GHD}_{c,n}(x, y) \neq \mathcal{P}(x, y)$.

Bounding $\Pr[BAD_2]$ is subtle:

- $x$ is good, so $\Pr[\mathcal{P} \text{ errs } | x] \leq 2\varepsilon$
  - But this requires $(x, y) \sim \mu_{c,n}$
- Random extension $(x', y') \rightarrow (x, y)$ is not $\sim \mu_{c,n}$.
- Actual distrib (fixed $x$, random $y$):
  - $(x, y) \sim (\mu_{c',n'} | x) \otimes \text{Unif}_{\overline{n}}$
  - $y$ uniform over a subset of $\{0,1\}^n$, just like in $\mu_{c,n}$
The Second Bad Event

Recall $BAD_2$: $\text{GHD}_{c,n}(x, y) \neq \mathcal{P}(x, y)$.

Bounding $\text{Pr}[BAD_2]$ is subtle:

- $x$ is good, so $\text{Pr}[\mathcal{P} \text{ errs } | x] \leq 2\varepsilon$
  - But this requires $(x, y) \sim \mu_{c,n}$

- Random extension $(x', y') \rightarrow (x, y)$ is not $\sim \mu_{c,n}$.

- Actual distrib (fixed $x$, random $y$):
  - $(x, y) \sim (\mu_{c',n'} | x) \otimes \text{Unif}_n$
  - $y$ uniform over a subset of $\{0, 1\}^n$, just like in $\mu_{c,n}$

\[
\begin{array}{cccccccccccccccc}
& y_1 & y_2 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & y_{2^n} \\
x & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet
\end{array}
\]
The Second Bad Event

Recall $BAD_2$: $\text{GHD}_{c,n}(x, y) \neq \mathcal{P}(x, y)$.

Bounding $\Pr[BAD_2]$ is subtle:

- $x$ is good, so $\Pr[\mathcal{P} \text{ errs} \mid x] \leq 2\varepsilon$
  - But this requires $(x, y) \sim \mu_{c,n}$

- Random extension $(x', y') \rightarrow (x, y)$ is not $\sim \mu_{c,n}$.

- Actual distrib (fixed $x$, random $y$):
  - $(x, y) \sim (\mu_{c',n'} \mid x) \otimes \text{Unif}_{\bar{n}}$
  - $y$ uniform over a subset of $\{0,1\}^n$, just like in $\mu_{c,n}$
The Second Bad Event

Recall $BAD_2$: $\text{GHD}_{c,n}(x, y) \neq P(x, y)$.

Bounding $\Pr[BAD_2]$ is subtle:

- $x$ is good, so $\Pr[P \text{ errs} \mid x] \leq 2\varepsilon$
  - But this requires $(x, y) \sim \mu_{c,n}$
- Random extension $(x', y') \rightarrow (x, y)$ is not $\sim \mu_{c,n}$.
- Actual distrib (fixed $x$, random $y$):
  - $(x, y) \sim (\mu_{c',n'} \mid x) \otimes \text{Unif}_n$
  - $y$ uniform over a subset of $\{0, 1\}^n$, just like in $\mu_{c,n}$

Lemma: $\Pr[BAD_2] = O(\varepsilon)$. 
Round Elimination, First Attempt (Recap)

Putting it together:

- $\mathcal{P}$ is $k$-round $\varepsilon$-error protocol for $\text{GHD}_{c,n}$
- $Q_1$ is $(k - 1)$-round $\varepsilon'$-error protocol for $\text{GHD}_{c',n'}$ with
  - $c' = 2c$, $n' = n/3$
  - $\varepsilon' = 1/8 + O(\varepsilon)$
Round Elimination, First Attempt (Recap)

Putting it together:

- \( P \) is \( k \)-round \( \varepsilon \)-error protocol for \( \text{GHD}_{c,n} \)
- \( Q_1 \) is \((k - 1)\)-round \( \varepsilon' \)-error protocol for \( \text{GHD}_{c',n'} \) with
  - \( c' = 2c \), \( n' = n/3 \)
  - \( \varepsilon' \leq 1/8 + 16\varepsilon \) ← Can’t repeat this argument!
Putting it together:

- \( \mathcal{P} \) is \( k \)-round \( \varepsilon \)-error protocol for \( \text{GHD}_{c,n} \)
- \( \mathcal{Q}_1 \) is \((k - 1)\)-round \( \varepsilon' \)-error protocol for \( \text{GHD}_{c',n'} \) with
  - \( c' = 2c, n' = n/3 \)
  - \( \varepsilon' \leq 1/8 + 16\varepsilon \) \( \leftarrow \) Can’t repeat this argument!

Second attempt: protocol \( \mathcal{Q} \):

- Repeat \( \mathcal{Q}_1 \) \( 2^{O(k)} \) times in parallel, take majority
- Blows up communication by \( 2^{O(k)} \)
- Error analysis even more subtle: not just a Chernoff bound
Round Elimination, Second Attempt

Putting it together:

- $P$ is $k$-round $\varepsilon$-error protocol for $\text{GHD}_{c,n}$
- $Q_1$ is $(k - 1)$-round $\varepsilon'$-error protocol for $\text{GHD}_{c',n'}$ with
  - $c' = 2c$, $n' = n/3$
  - $\varepsilon' \leq 1/8 + 16\varepsilon$ ← Can’t repeat this argument!

Second attempt: protocol $Q$:

- Repeat $Q_1$ $2^{O(k)}$ times in parallel, take majority
- Blows up communication by $2^{O(k)}$
- Error analysis even more subtle: not just a Chernoff bound

**Lemma:** $\Pr[Q \text{ errs}] = O(\varepsilon)$. 
Eventual Round Elimination Lemma

**Lemma**: If there is a $k$-round, $\varepsilon$-error protocol for $\text{GHD}_{c,n}$ in which each player sends $s \ll n$ bits, then there is a $(k-1)$-round, $O(\varepsilon)$-error protocol for $\text{GHD}_{2c,n/3}$ in which each player sends $2^{O(k)}s$ bits.

Recall Base Case Lemma: There is no zero-round protocol with error $< 1/2$. 

Eventual Round Elimination Lemma

**Lemma:** If there is a $k$-round, $\varepsilon$-error protocol for $\text{GHD}_{c,n}$ in which each player sends $s \ll n$ bits, then there is a $(k - 1)$-round, $O(\varepsilon)$-error protocol for $\text{GHD}_{2c,n/3}$ in which each player sends $2^{O(k)}s$ bits.

Recall Base Case Lemma: There is no zero-round protocol with error $< 1/2$.

Consequence: Main Theorem

**Theorem:** There is no $o(n)$-bit, $\frac{1}{3}$-error, $O(1)$-round randomized protocol for $\text{GHD}_{c,n}$. In other words, $R^{O(1)}(\text{GHD}) = \Omega(n)$. 
**Eventual Round Elimination Lemma**

**Lemma:** If there is a $k$-round, $\varepsilon$-error protocol for $\text{GHD}_{c,n}$ in which each player sends $s \ll n$ bits, then there is a $(k - 1)$-round, $O(\varepsilon)$-error protocol for $\text{GHD}_{2c,n/3}$ in which each player sends $2^{O(k)}s$ bits.

Recall Base Case Lemma: There is no zero-round protocol with error $< 1/2$.

**Consequence: Main Theorem**

**Theorem:** There is no $o(n)$-bit, $\frac{1}{3}$-error, $O(1)$-round randomized protocol for $\text{GHD}_{c,n}$. In other words, $R^{O(1)}(\text{GHD}) = \Omega(n)$.

More Specific: $R^{k}(\text{GHD}) = n/2^{O(k^2)}$. 
Multi-pass lower bounds for Distinct Elements and $F_k$ has been an important open question since at least 2003. Why did it remain open for so long?
Why Did This Take So Long?

Multi-pass lower bounds for Distinct Elements and $F_k$ has been an important open question since at least 2003. Why did it remain open for so long?

Underlying communication problem thorny!
Why Did This Take So Long?

Multi-pass lower bounds for Distinct Elements and $F_k$ has been an important open question since at least 2003. Why did it remain open for so long?

Underlying communication problem thorny! Resists the “usual” attacks:

- Rectangle-based methods (discrepancy/corruption)

- Approximate polynomial degree

- Pattern matrix, Factorization norms [Sherstov’08], [Linial-Shraibman’07]

- Information complexity [C.-Shi-Wirth-Yao’01], [BarYossef-J.-K.-S.’02]
Why Did This Take So Long?

Multi-pass lower bounds for Distinct Elements and $F_k$ has been an important open question since at least 2003. Why did it remain open for so long?

Underlying communication problem thorny! Resists the “usual” attacks:

- Rectangle-based methods (discrepancy/corruption)
  
  Matrix has large near-monochromatic rectangles

- Approximate polynomial degree

- Pattern matrix, Factorization norms [Sherstov’08], [Linial-Shraibman’07]

- Information complexity [C.-Shi-Wirth-Yao’01], [BarYossef-J.-K.-S.’02]
Why Did This Take So Long?

Multi-pass lower bounds for Distinct Elements and $F_k$ has been an important open question since at least 2003. Why did it remain open for so long?

Underlying communication problem thorny! Resists the “usual” attacks:

- Rectangle-based methods (discrepancy/corruption)
  
  Matrix has large near-monochromatic rectangles

- Approximate polynomial degree
  
  Underlying predicate has approx degree $\tilde{O}(\sqrt{n})$

- Pattern matrix, Factorization norms [Sherstov’08], [Linial-Shraibman’07]

- Information complexity [C.-Shi-Wirth-Yao’01], [BarYossef-J.-K.-S.’02]
Why Did This Take So Long?

Multi-pass lower bounds for Distinct Elements and $F_k$ has been an important open question since at least 2003. Why did it remain open for so long?

Underlying communication problem thorny! Resists the “usual” attacks:

- Rectangle-based methods (discrepancy/corruption)
  
  Matrix has large near-monochromatic rectangles

- Approximate polynomial degree
  
  Underlying predicate has approx degree $\tilde{O}(\sqrt{n})$

- Pattern matrix, Factorization norms [Sherstov’08], [Linial-Shraibman’07]
  
  Quantum communication upper bound $O(\sqrt{n} \log n)$

- Information complexity [C.-Shi-Wirth-Yao’01], [BarYossef-J.-K.-S.’02]
Why Did This Take So Long?

Multi-pass lower bounds for Distinct Elements and $F_k$ has been an important open question since at least 2003. Why did it remain open for so long?

Underlying communication problem thorny! Resists the “usual” attacks:

- Rectangle-based methods (discrepancy/corruption)
  Matrix has large near-monochromatic rectangles

- Approximate polynomial degree
  Underlying predicate has approx degree $\tilde{O}(\sqrt{n})$

- Pattern matrix, Factorization norms [Sherstov’08], [Linial-Shraibman’07]
  Quantum communication upper bound $O(\sqrt{n} \log n)$

- Information complexity [C.-Shi-Wirth-Yao’01], [BarYossef-J.-K.-S.’02]
  Hmm! Can’t see a concrete obstacle
Why Did This Take So Long?

Multi-pass lower bounds for Distinct Elements and $F_k$ has been an important open question since at least 2003. Why did it remain open for so long?

Underlying communication problem thorny! Resists the “usual” attacks:

- Rectangle-based methods (discrepancy/corruption)
  Matrix has large near-monochromatic rectangles

- Approximate polynomial degree
  Underlying predicate has approx degree $\tilde{O}(\sqrt{n})$

- Pattern matrix, Factorization norms [Sherstov’08], [Linial-Shraibman’07]
  Quantum communication upper bound $O(\sqrt{n \log n})$

- Information complexity [C.-Shi-Wirth-Yao’01], [BarYossef-J.-K.-S.’02]
  Hmm! Can’t see a concrete obstacle
  We’re biased (Amit helped invent it, so it’s his pet technique)
Open Problems

1. The key problem here: Settle $R(GHD)$.


3. This should help with other streaming problems, e.g., longest increasing subsequence.

Contact jbrody@cs.dartmouth.edu