Tight Bounds for Distributed Functional Monitoring

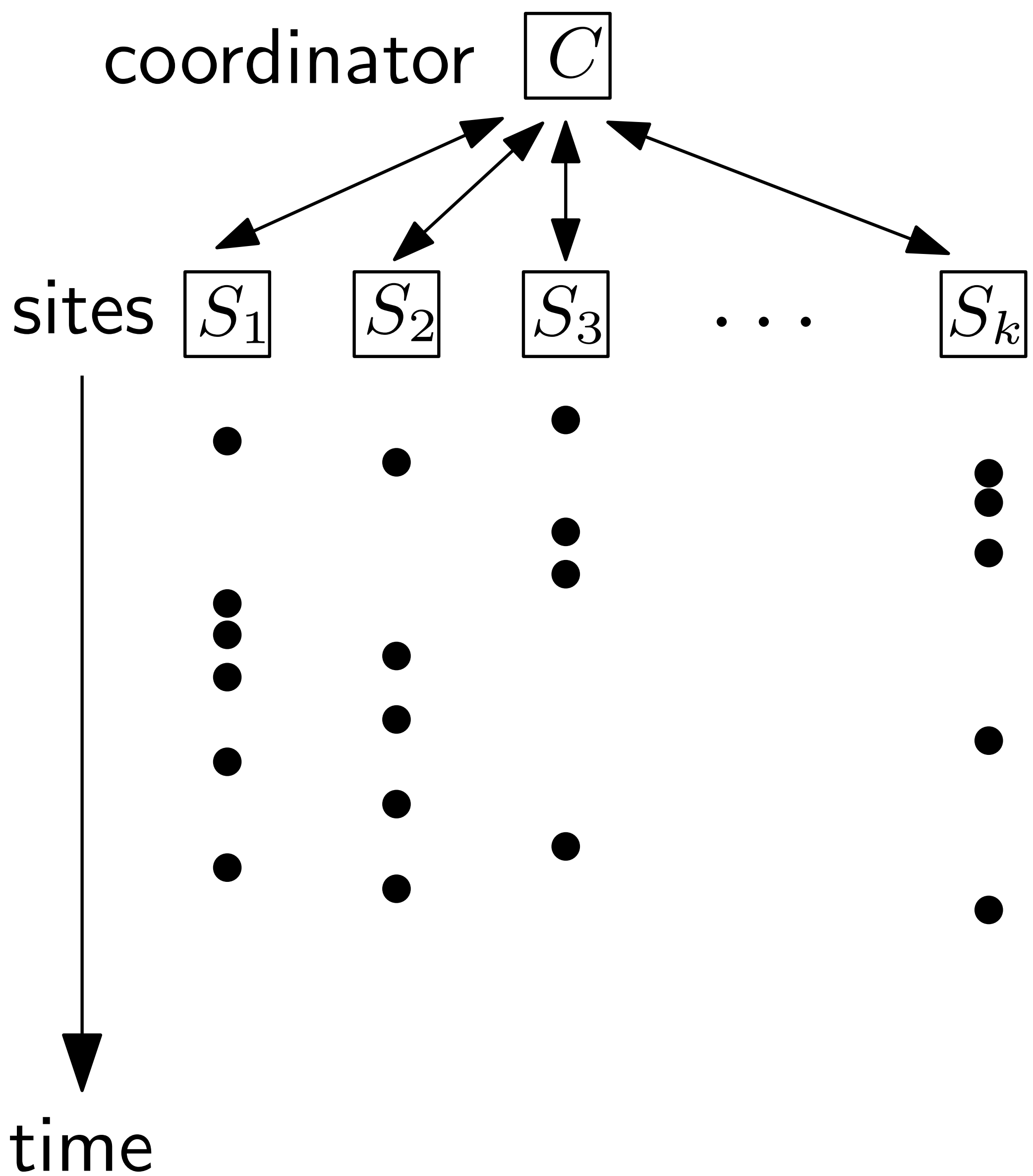
Qin Zhang
MADALGO, Aarhus University

Joint with
David Woodruff, IBM Almaden

NII Shonan meeting, Japan
Jan. 2012

The distributed streaming model

(a.k.a. distributed functional/continuous monitoring)

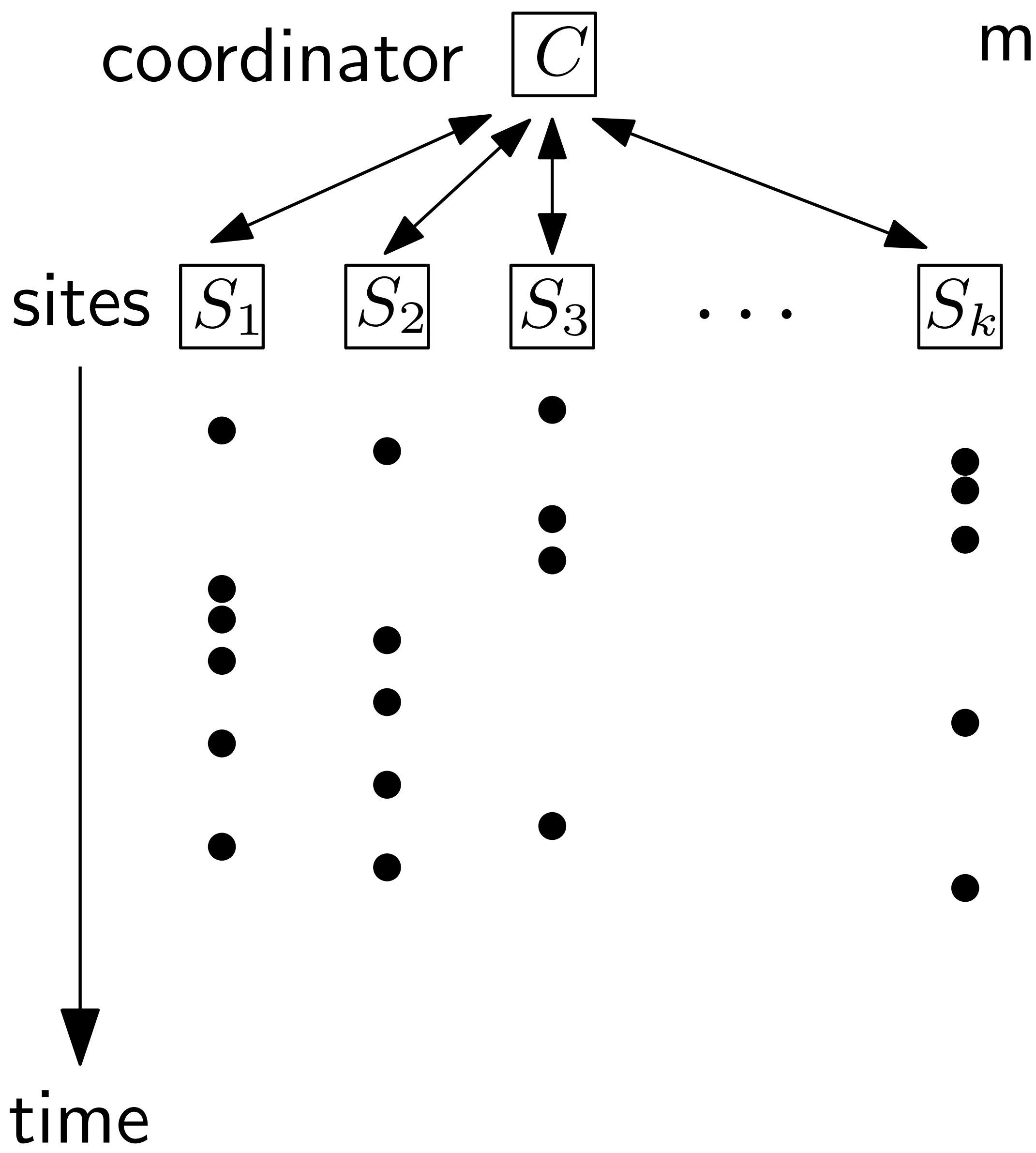


$A(t)$: set of elements received up to time t from all sites.[a]

[a] Assume ≤ 1 item comes at each time unit.

The distributed streaming model

(a.k.a. distributed functional/continuous monitoring)



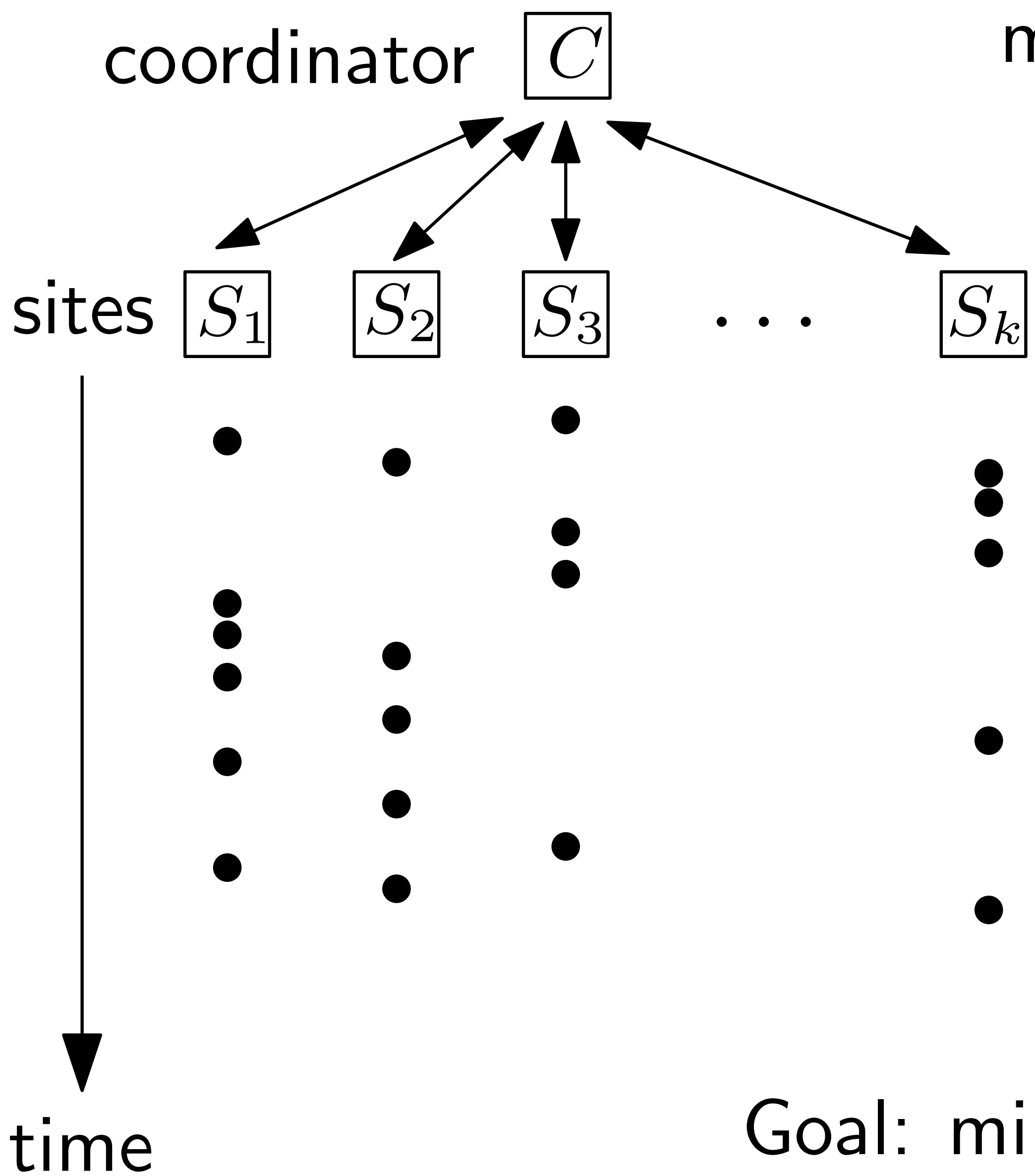
The coordinator needs to maintain $f(A(t))$ for all t.

$A(t)$: set of elements received up to time t from all sites. [a]

[a] Assume ≤ 1 item comes at each time unit.

The distributed streaming model

(a.k.a. distributed functional/continuous monitoring)



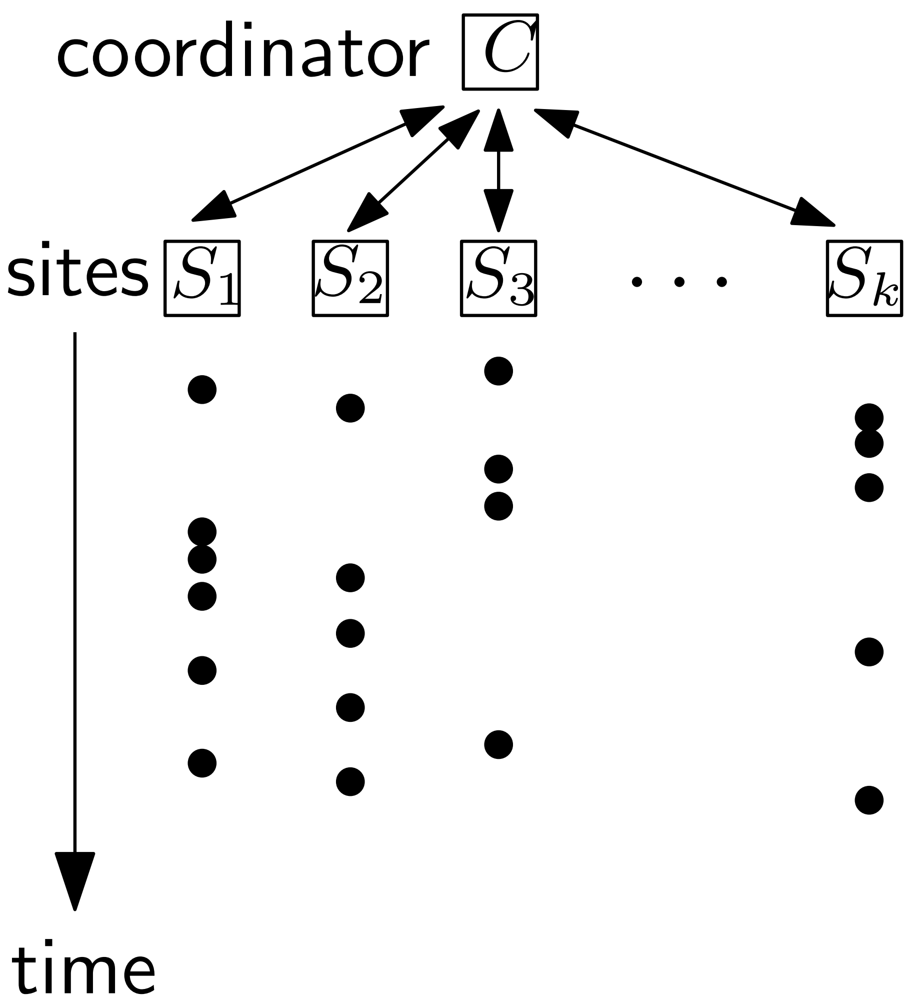
The coordinator needs to maintain $f(A(t))$ for all t.

$A(t)$: set of elements received up to time t from all sites.[a]

[a] Assume ≤ 1 item comes at each time unit.

Goal: minimize communication cost

Problems



The Distributed Streaming Model

Static case (a one-shot/static computation at the end)

- Top-k
- Heavy-hitter
- ...

Dynamic case

- Samplings
- Frequent moments $(F_0, F_1, F_2, \ldots)$
- Heavy-hitter
- Quantile
- Entropy
- Non-linear functions
- ...

This talk

What you would like to see:

- Efficient algorithms/protocols
- Practical heuristics

This talk

What you would like to see:

- Efficient algorithms/protocols
- Practical heuristics

What you (probably) do not want to see:

- "Useless" impossibility results
- Complicated proofs

This talk

What you would like to see:

- Efficient algorithms/protocols
- Practical heuristics

What you (probably) do not want to see:

- "Useless" impossibility results
- Complicated proofs

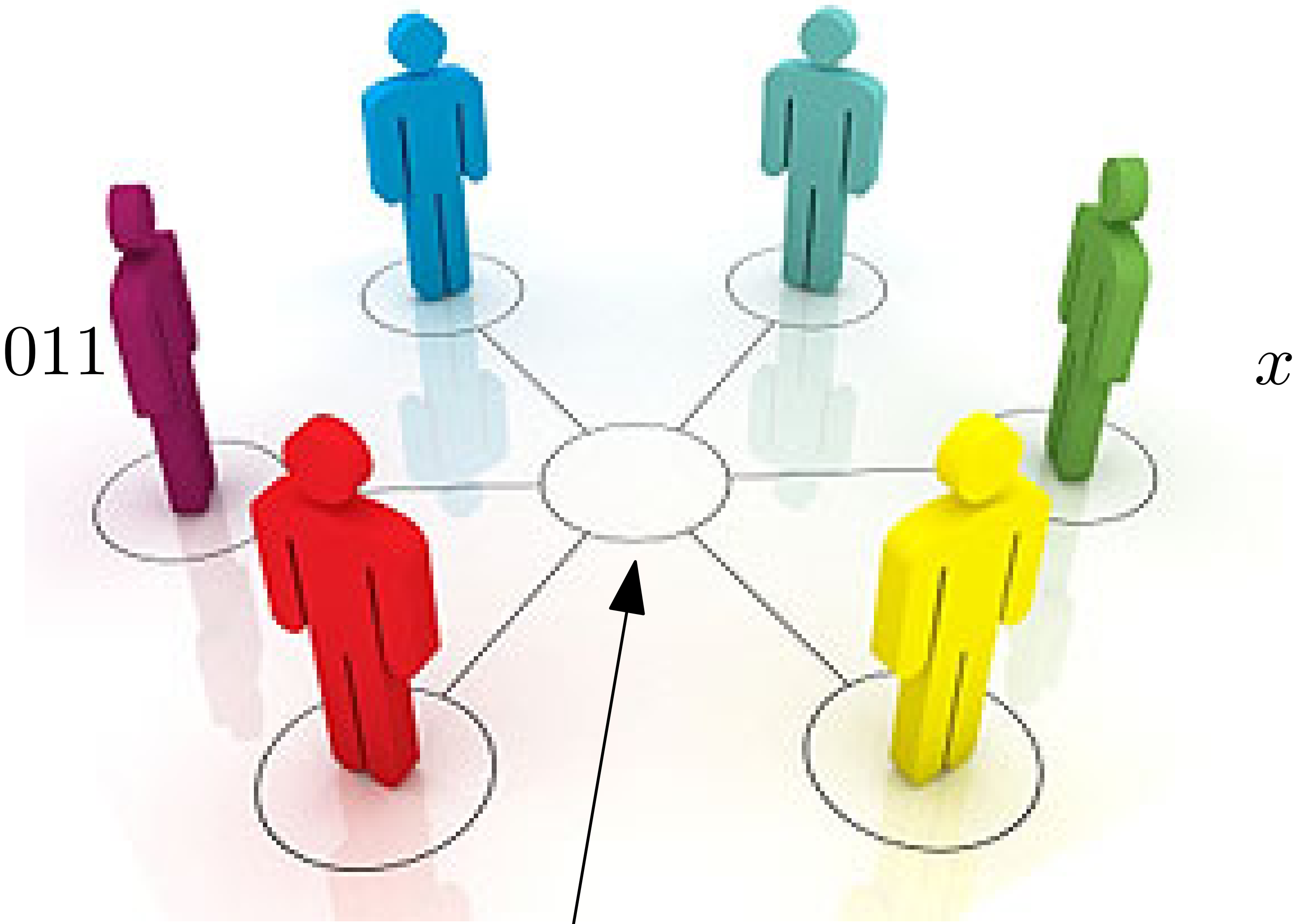
Unfortunately, in the next 30 minutes ...

The multiparty communication model

– A model for lower bounds

$x_1 = 010011$ $x_2 = 111011$

$x_k = 100011$



$x_3 = 111111$

We want to compute $f(x_1, x_2, \ldots, x_k)$

f can be bit-wise XOR, OR, AND, MAJ ...

The multiparty communication model

– A model for lower bounds

$x_1 = 010011$ $x_2 = 111011$

Blackboard: One speaks, everyone else hears.

Message passing: If x_1 talks to x_2, others cannot hear.

$x_k = 100011$

$x_3 = 111111$

We want to compute $f(x_1, x_2, \ldots, x_k)$

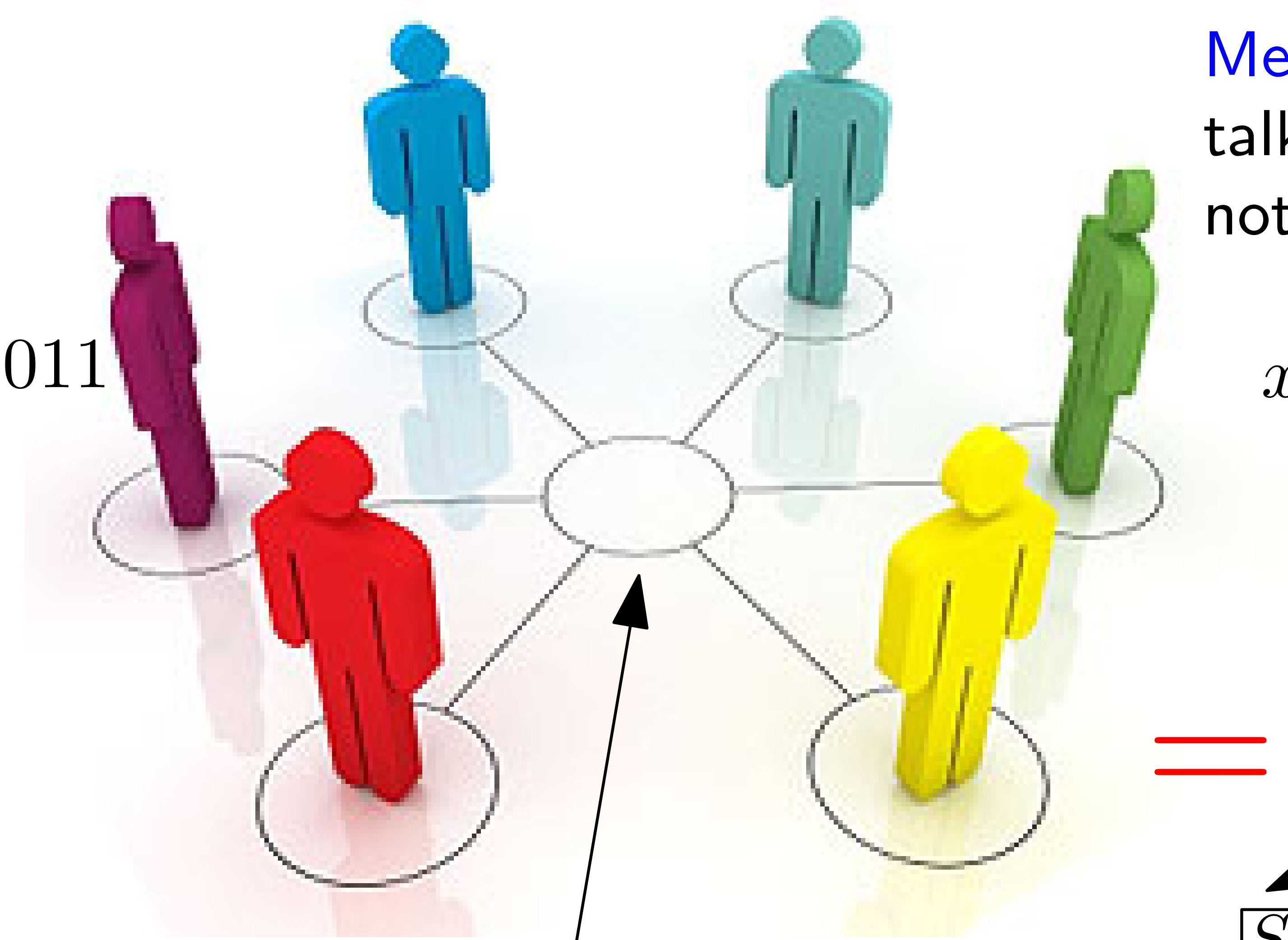
f can be bit-wise XOR, OR, AND, MAJ ...

The multiparty communication model

– A model for lower bounds

$x_1 = 010011$ $x_2 = 111011$

$x_k = 100011$



$x_3 = 111111$

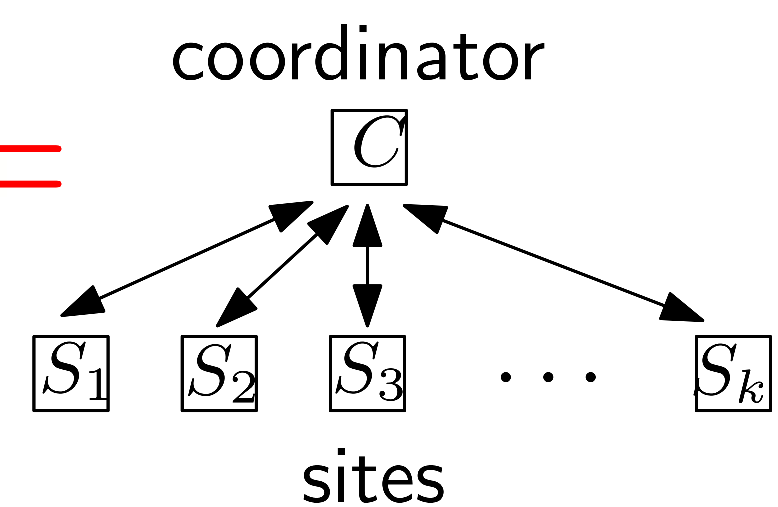
We want to compute $f(x_1, x_2, \ldots, x_k)$

f can be bit-wise XOR, OR, AND, MAJ ...

Blackboard: One speaks, everyone else hears.

Message passing: If x_1 talks to x_2, others cannot hear.

$=$



Previously

- Some works in the blackboard model. Almost nothing in the message-passing model.

Previously

- Some works in the blackboard model. Almost nothing in the message-passing model.

- This SODA, with Jeff Phillips and Elad Verbin we proposed a general and elegant technique called "symmetrization" which works in both variants. In particular, we obtained (in the message-passing model)

  1. $\Omega(nk)$ for the bitwise-XOR/OR/AND/MAJ.

  2. $\tilde{\Omega}(nk)$ for connectivity.

Previously

- Some works in the blackboard model. Almost nothing in the message-passing model.

- This SODA, with Jeff Phillips and Elad Verbin we proposed a general and elegant technique called "symmetrization" which works in both variants. In particular, we obtained (in the message-passing model)

  1. $\Omega(nk)$ for the bitwise-XOR/OR/AND/MAJ.

  2. $\tilde{\Omega}(nk)$ for connectivity.

- Artificial? Well ...

Previously

- Some works in the blackboard model. Almost nothing in the message-passing model.

- This SODA, with Jeff Phillips and Elad Verbin we proposed a general and elegant technique called "symmetrization" which works in both variants. In particular, we obtained (in the message-passing model)

  1. $\Omega(nk)$ for the bitwise-XOR/OR/AND/MAJ.

  2. $\tilde{\Omega}(nk)$ for connectivity.

- Artificial? Well ...

- In any case, let's look at real important problems.

Now, important problems

- Samplings
- Frequent moments $(F_0, F_1, F_2, \ldots)$
- Heavy-hitter
- Quantile
- Entropy
- $\ldots$

Now, important problems

- Samplings Solved
- Frequent moments $(F_0, F_1, F_2, \ldots)$
- Heavy-hitter
- Quantile
- Entropy
- $\ldots$

Now, important problems

- Samplings
- Frequent moments $(F_0, F_1, F_2, \ldots)$
- Heavy-hitter
- Quantile
- Entropy
- $\ldots$

Solved

Our work

Results

	Previous work	This paper	Previous work	This paper
Problem	LB	LB (all static)	UB	UB
F_0	$\Omega(k)$ [20]	$\mathbf{\Omega(k/\varepsilon^2)}$	$\tilde{O}(k/\varepsilon^2)$ [20]	–
F_2	$\Omega(k)$ [20]	$\mathbf{\tilde{\Omega}(k/\varepsilon^2)}$ (BB)	$\tilde{O}(k^2/\varepsilon + k^{1.5}/\varepsilon^3)$ [20]	$\mathbf{\tilde{O}(\frac{k}{\mathrm{poly}(\varepsilon)})}$
F_p $(p > 1)$	$\Omega(k + 1/\varepsilon^2)$ [5, 16]	$\mathbf{\tilde{\Omega}(k^{p-1}/\varepsilon^2)}$ (BB)	$\tilde{O}(\frac{p}{\varepsilon^{1+2/p}}k^{2p+1}N^{1-2/p})$ [20]	$\mathbf{\tilde{O}(\frac{k^{p-1}}{\mathrm{poly}(\varepsilon)})}$
All-quantile	$\tilde{\Omega}(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})$ [32]	$\mathbf{\Omega(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})}$ (BB)	$\tilde{O}(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})$ [32]	–
Heavy Hitters	$\tilde{\Omega}(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})$ [32]	$\mathbf{\Omega(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})}$ (BB)	$\tilde{O}(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})$ [32]	–
Entropy	$\tilde{\Omega}(1/\sqrt{\varepsilon})$ [5]	$\mathbf{\tilde{\Omega}(k/\varepsilon^2)}$ (BB)	$\tilde{O}(\frac{k}{\varepsilon^3})$ [5], $\tilde{O}(\frac{k}{\varepsilon^2})$ (static) [31]	–
ℓ_p $(p \in (0, 2])$	–	$\mathbf{\tilde{\Omega}(k/\varepsilon^2)}$ (BB)	$\tilde{O}(k/\varepsilon^2)$ (static) [38]	–

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

Results

	Previous work	This paper	Previous work	This paper
Problem	LB	LB (all static)	UB	UB
F_0	$\Omega(k)$ [20]	$\mathbf{\Omega(k/\varepsilon^2)}$	$\tilde{O}(k/\varepsilon^2)$ [20]	–
F_2	$\Omega(k)$ [20]	$\mathbf{\tilde{\Omega}(k/\varepsilon^2)}$ (BB)	$\tilde{O}(k^2/\varepsilon + k^{1.5}/\varepsilon^3)$ [20]	$\mathbf{\tilde{O}(\frac{k}{poly(\varepsilon)})}$
F_p $(p > 1)$	$\Omega(k + 1/\varepsilon^2)$ [5, 16]	$\mathbf{\tilde{\Omega}(k^{p-1}/\varepsilon^2)}$ (BB)	$\tilde{O}(\frac{p}{\varepsilon^{1+2/p}} k^{2p+1} N^{1-2/p})$ [20]	$\mathbf{\tilde{O}(\frac{k^{p-1}}{poly(\varepsilon)})}$
All-quantile	$\tilde{\Omega}(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})$ [32]	$\mathbf{\Omega(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})}$ (BB)	$\tilde{O}(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})$ [32]	–
Heavy Hitters	$\tilde{\Omega}(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})$ [32]	$\mathbf{\Omega(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})}$ (BB)	$\tilde{O}(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})$ [32]	–
Entropy	$\tilde{\Omega}(1/\sqrt{\varepsilon})$ [5]	$\mathbf{\tilde{\Omega}(k/\varepsilon^2)}$ (BB)	$\tilde{O}(\frac{k}{\varepsilon^3})$ [5], $\tilde{O}(\frac{k}{\varepsilon^2})$ (static) [31]	–
ℓ_p $(p \in (0, 2])$	–	$\mathbf{\tilde{\Omega}(k/\varepsilon^2)}$ (BB)	$\tilde{O}(k/\varepsilon^2)$ (static) [38]	–

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

- (Almost) tight bounds for all these questions
- Static lower bounds (almost) match dynamic upper bounds. (up to polylog factors)

Results

	Previous work	This paper	Previous work	This paper
Problem	LB	LB (all static)	UB	UB
F_0	$\Omega(k)$ [20]	$\mathbf{\Omega(k/\varepsilon^2)}$ Today	$\tilde{O}(k/\varepsilon^2)$ [20]	–
F_2	$\Omega(k)$ [20]	$\mathbf{\tilde{\Omega}(k/\varepsilon^2)}$ (BB)	$\tilde{O}(k^2/\varepsilon + k^{1.5}/\varepsilon^3)$ [20]	$\mathbf{\tilde{O}(\frac{k}{\mathrm{poly}(\varepsilon)})}$
$F_p\ (p > 1)$	$\Omega(k + 1/\varepsilon^2)$ [5, 16]	$\mathbf{\tilde{\Omega}(k^{p-1}/\varepsilon^2)}$ (BB)	$\tilde{O}(\frac{p}{\varepsilon^{1+2/p}} k^{2p+1} N^{1-2/p})$ [20]	$\mathbf{\tilde{O}(\frac{k^{p-1}}{\mathrm{poly}(\varepsilon)})}$
All-quantile	$\tilde{\Omega}(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})$ [32]	$\mathbf{\Omega(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})}$ (BB)	$\tilde{O}(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})$ [32]	–
Heavy Hitters	$\tilde{\Omega}(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})$ [32]	$\mathbf{\Omega(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})}$ (BB)	$\tilde{O}(\min\{\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\})$ [32]	–
Entropy	$\tilde{\Omega}(1/\sqrt{\varepsilon})$ [5]	$\mathbf{\tilde{\Omega}(k/\varepsilon^2)}$ (BB)	$\tilde{O}(\frac{k}{\varepsilon^3})$ [5], $\tilde{O}(\frac{k}{\varepsilon^2})$ (static) [31]	–
$\ell_p\ (p \in (0, 2])$	–	$\mathbf{\tilde{\Omega}(k/\varepsilon^2)}$ (BB)	$\tilde{O}(k/\varepsilon^2)$ (static) [38]	–

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

- (Almost) tight bounds for all these questions
- Static lower bounds (almost) match dynamic upper bounds. (up to polylog factors)

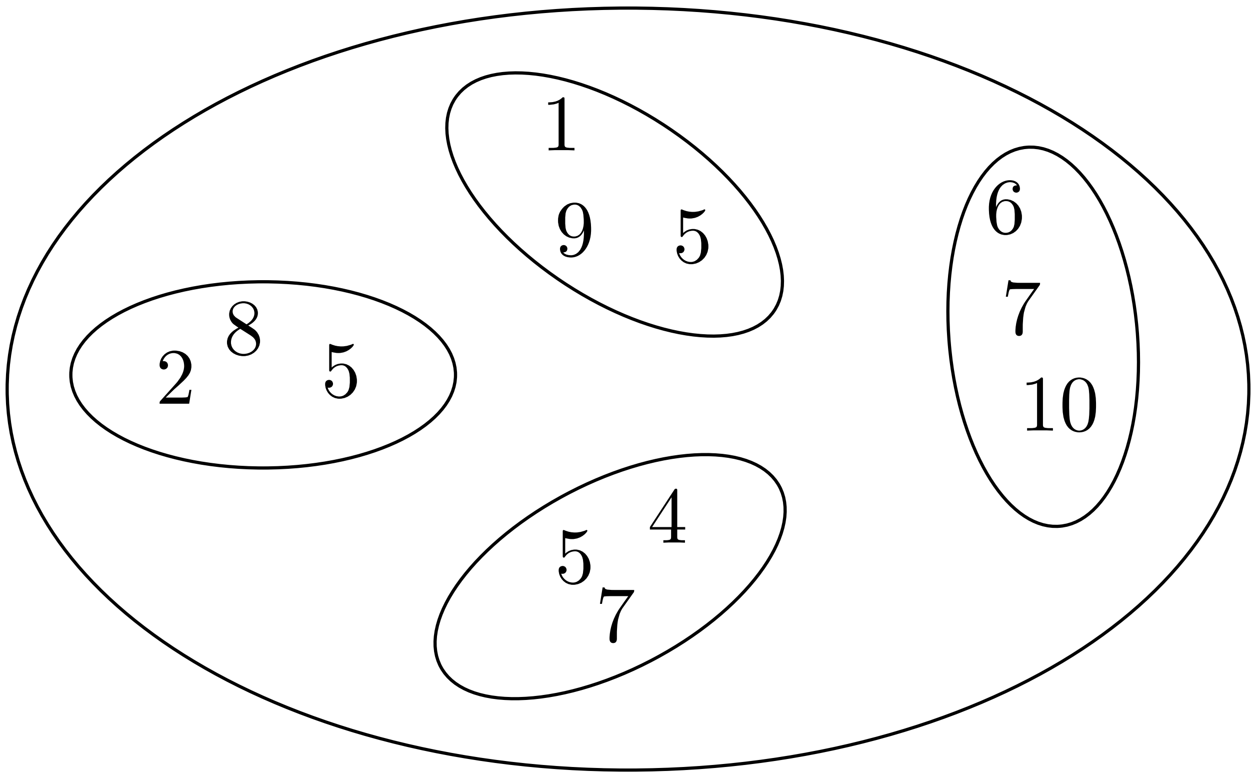
F_0 upper bound

(Cormode, Muthu and Yi 2008)

The $(1+\varepsilon)$-approximation F_0 problem

We have k sites $S_1, S_2, \ldots, S_k$. S_i holds a set X_i.

Our goal: compute $F_0(\cup_{i \in k} X_i)$ up to $(1+\varepsilon)$-approximation.



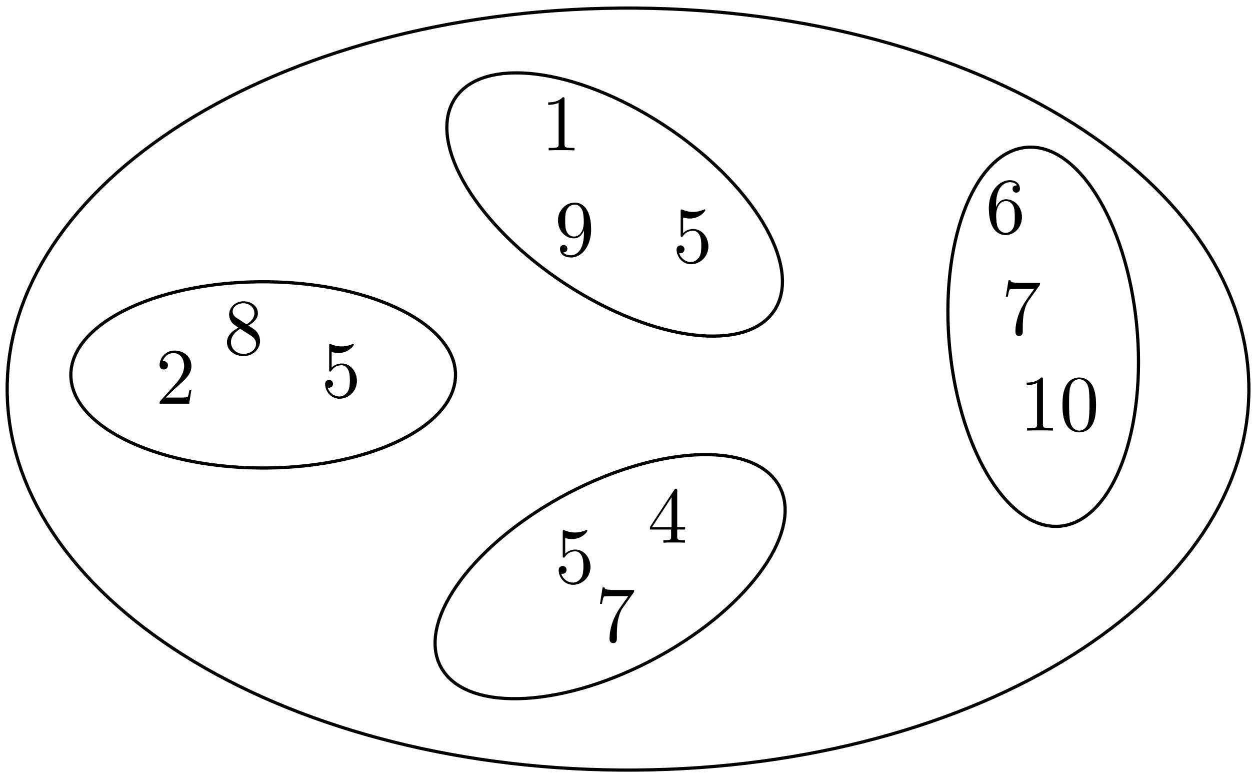
How many distinct items?

A fundamental problem in data analysis.

The $(1+\varepsilon)$-approximation F_0 problem

We have k sites $S_1, S_2, \ldots, S_k$. S_i holds a set X_i.

Our goal: compute $F_0(\cup_{i \in k} X_i)$ up to $(1+\varepsilon)$-approximation.



How many distinct items?

A fundamental problem in data analysis.

Current best UB: $\tilde{O}(k/\varepsilon^2)$ (Cormode, Muthu, Yi 2008)
Holds in the dynamic case.

General idea for the one-shot computation

- Each site generates a “sketch” via small-space streaming algorithms.

- The coordinator combines (via communication) the sketches from the k sites to obtain a global sketch, from which we can extract the answer.

The FM sketch

- Take a pair-wise independent random hash function $h : \{1, \ldots, n\} \rightarrow \{1, \ldots, 2^d\}$, where $2^d > n$

The FM sketch

- Take a pair-wise independent random hash function $h : \{1, \ldots, n\} \to \{1, \ldots, 2^d\}$, where $2^d > n$
- For each incoming element x, compute $h(x)$
  - e.g., $h(5) = 10101100010000$
  - Count how many trailing zeros
  - Remember the max $\#$ trailing zeroes in any $h(x)$

The FM sketch

- Take a pair-wise independent random hash function $h : \{1, \ldots, n\} \to \{1, \ldots, 2^d\}$, where $2^d > n$
- For each incoming element x, compute $h(x)$
  - e.g., $h(5) = 10101100010000$
  - Count how many trailing zeros
  - Remember the max # trailing zeroes in any $h(x)$
- Let Y be the max # trailing zeroes
  - Can show $E[2^Y] = \#$distinct elements

One-shot case, the FM sketch (cont.)

- So 2^Y is an unbiased estimator for $\#$ distinct elements

One-shot case, the FM sketch (cont.)

- So 2^Y is an unbiased estimator for $\#$ distinct elements
- However, has a large variance
  - Some techniques [Bar-Yossef et. al. 2002] can produce a good estimator that has probability $1-\delta$ to be within relative error ε.
  - Space increased to $\tilde{O}(1/\varepsilon^2)$

One-shot case, the FM sketch (cont.)

- So 2^Y is an unbiased estimator for $\#$ distinct elements
- However, has a large variance
  - Some techniques [Bar-Yossef et. al. 2002] can produce a good estimator that has probability $1-\delta$ to be within relative error ε.
  - Space increased to $\tilde{O}(1/\varepsilon^2)$
- FM sketch has linearity
  - Y_1 from A, Y_2 from B, then $2^{\max\{Y_1,Y_2\}}$ estimates $\#$ distinct items in $A \cup B$.

One-shot case, the FM sketch (cont.)

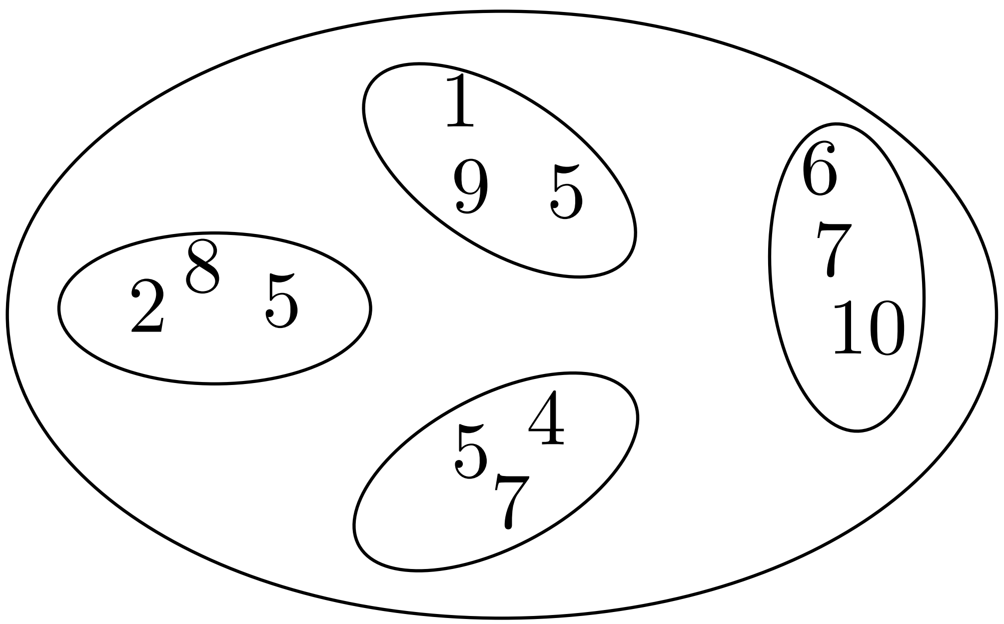
- So 2^Y is an unbiased estimator for $\#$ distinct elements
- However, has a large variance
  - Some techniques [Bar-Yossef et. al. 2002] can produce a good estimator that has probability $1-\delta$ to be within relative error ε.
  - Space increased to $\tilde{O}(1/\varepsilon^2)$
- FM sketch has linearity
  - Y_1 from A, Y_2 from B, then $2^{\max\{Y_1, Y_2\}}$ estimates $\#$ distinct items in $A \cup B$.
- Thus, we can use it to design a one-shot algorithm with communication $\tilde{O}(k/\varepsilon^2)$

F_0 lower bound

The F_0 problem

We have k sites $S_1, S_2, \ldots, S_k$. S_i holds a set X_i.

Our goal: compute $F_0(\cup_{i \in k} X_i)$ up to $(1+\varepsilon)$-approximation.



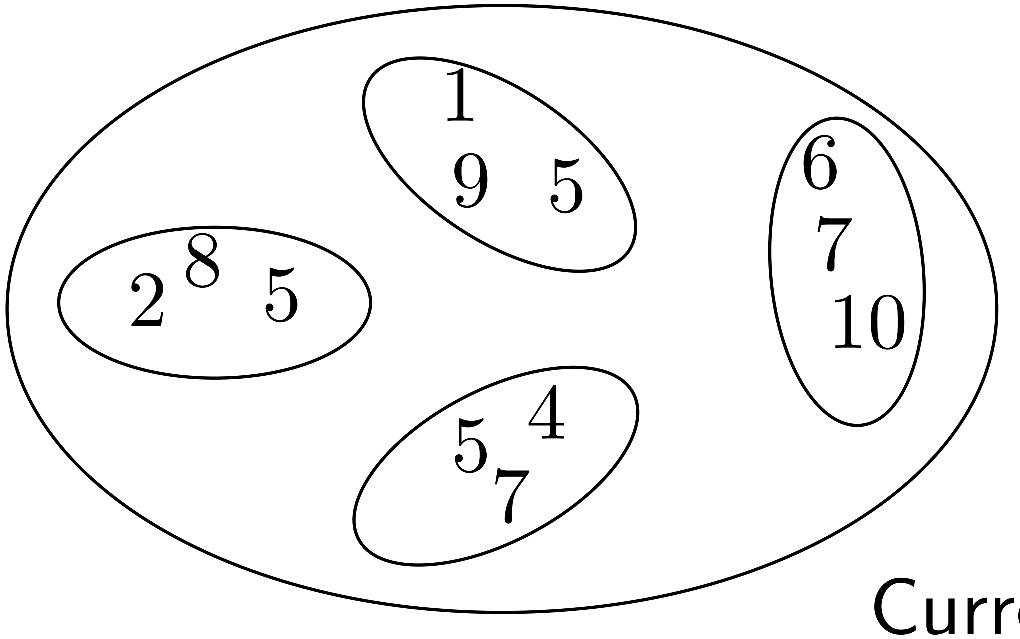
How many distinct items?

A fundamental problem in data analysis.

The F_0 problem

We have k sites $S_1, S_2, \ldots, S_k$. S_i holds a set X_i.

Our goal: compute $F_0(\cup_{i \in k} X_i)$ up to $(1+\varepsilon)$-approximation.



How many distinct items?

A fundamental problem in data analysis.

Current best UB: $\tilde{O}(k/\varepsilon^2)$

(Cormode, Muthu, Yi, 2008)

Holds in the dynamic case.

Previous LB: $\Omega(k)$ (Cormode, Muthu, Yi, 2008)

$\Omega(1/\varepsilon^2)$ (reduction from Gap-Hamming)

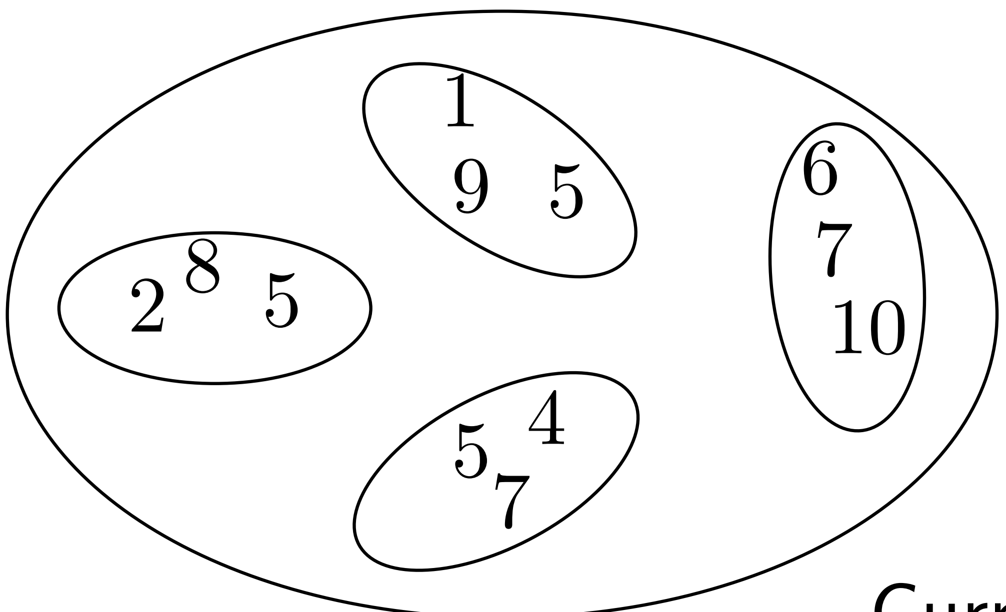
Our LB: $\Omega(k/\varepsilon^2)$.

Holds in the static and message-passing case.

The F_0 problem

We have k sites $S_1, S_2, \ldots, S_k$. S_i holds a set X_i.

Our goal: compute $F_0(\cup_{i \in k} X_i)$ up to $(1+\varepsilon)$-approximation.



How many distinct items?

A fundamental problem in data analysis.

Current best UB: $\tilde{O}(k/\varepsilon^2)$

(Cormode, Muthu, Yi, 2008)

Holds in the dynamic case.

Previous LB: $\Omega(k)$ (Cormode, Muthu, Yi, 2008)

$\Omega(1/\varepsilon^2)$ (reduction from Gap-Hamming)

Our LB: $\Omega(k/\varepsilon^2)$.

Holds in the static and message-passing case.

Tight!

The proof framework

Step 1: We first introduce a simpler problem called k-GAP-MAJ

Step 2: We compose k-GAP-MAJ with the Set Disjointness problem using information cost to prove a lower bound for F_0

k-GAP-MAJ

We have k sites $S_1, S_2, \ldots, S_k$. S_i holds a bit Z_i which is 1 w.p. β and 0 w.p. $1-\beta$ where $\omega(1/k) \leq \beta \leq 1/2$ is a prefixed value.

Our goal: compute the following function.

$$\mathsf{GM}(Z_1, Z_2, \ldots, Z_k) = \begin{cases} 0, & \text{if } \sum_{i\in[k]} Z_i \leq \beta k - \sqrt{\beta k}, \\ 1, & \text{if } \sum_{i\in[k]} Z_i \geq \beta k + \sqrt{\beta k}, \\ *, & \text{otherwise}, \end{cases}$$

where "$*$" means that the answer can be arbitrary.

k-GAP-MAJ

We have k sites $S_1, S_2, \ldots, S_k$. S_i holds a bit Z_i which is 1 w.p. β and 0 w.p. $1-\beta$ where $\omega(1/k) \leq \beta \leq 1/2$ is a prefixed value.

Our goal: compute the following function.

$$\mathrm{GM}(Z_1, Z_2, \ldots, Z_k) = \begin{cases} 0, & \text{if } \sum_{i \in [k]} Z_i \leq \beta k - \sqrt{\beta k}, \\ 1, & \text{if } \sum_{i \in [k]} Z_i \geq \beta k + \sqrt{\beta k}, \\ *, & \text{otherwise,} \end{cases}$$

where "$*$" means that the answer can be arbitrary.

- **Lemma 1**: If a protocol $\mathcal{P}$ computes k-GAP-MAJ correctly w.p. 0.9999, then w.p. $\Omega(1)$, the protocol has to learn at least $\Omega(k)$ of Z_i each with $\Omega(1)$ bit (that is, $H(Z_i \mid \Pi) \leq H_b(0.01\beta)$).

k-GAP-MAJ

We have k sites $S_1, S_2, \ldots, S_k$. S_i holds a bit Z_i which is 1 w.p. β and 0 w.p. $1-\beta$ where $\omega(1/k) \le \beta \le 1/2$ is a prefixed value.

Our goal: compute the following function.

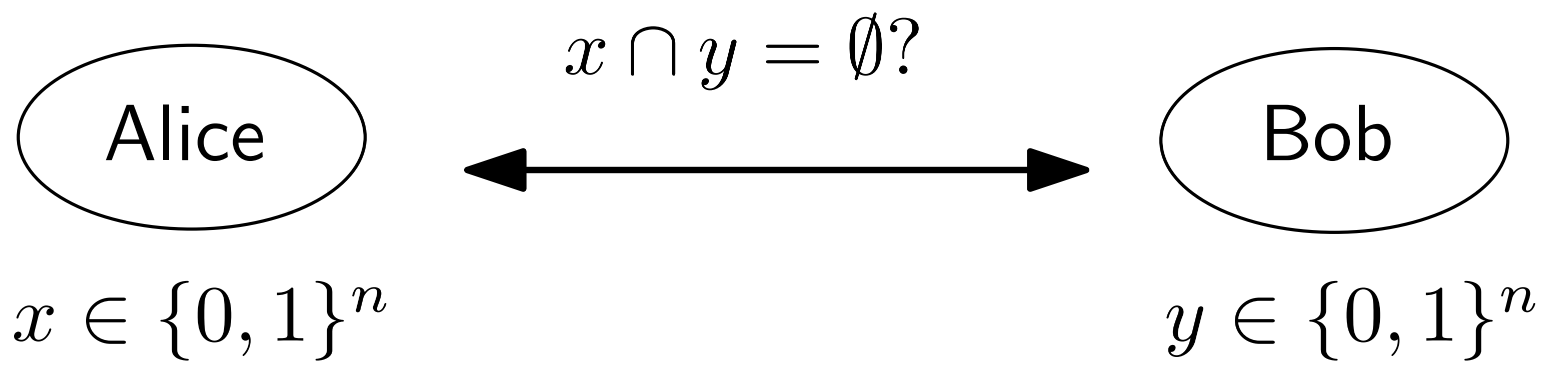
$$\text{GM}(Z_1, Z_2, \ldots, Z_k) = \begin{cases} 0, & \text{if } \sum_{i\in[k]} Z_i \le \beta k - \sqrt{\beta k}, \\ 1, & \text{if } \sum_{i\in[k]} Z_i \ge \beta k + \sqrt{\beta k}, \\ *, & \text{otherwise,} \end{cases}$$

where "$*$" means that the answer can be arbitrary.

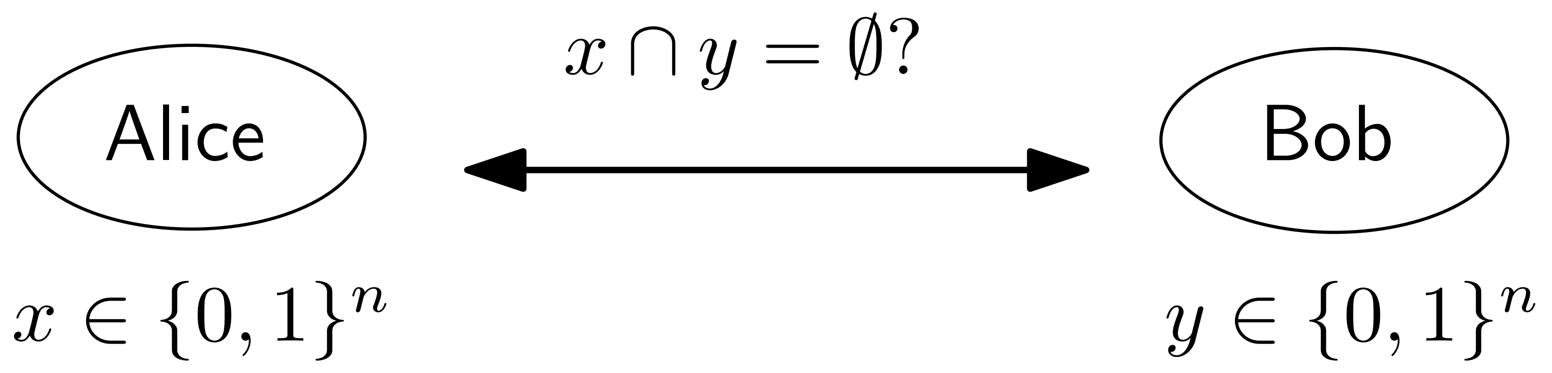
- **Lemma 1**: If a protocol $\mathcal{P}$ computes k-GAP-MAJ correctly w.p. 0.9999, then w.p. $\Omega(1)$, the protocol has to learn at least $\Omega(k)$ of Z_i each with $\Omega(1)$ bit (that is, $H(Z_i \mid \Pi) \le H_b(0.01\beta)$).

- Alternatively: $I(Z_1, Z_2, \ldots, Z_k; \Pi) = \Omega(k)$

Set disjointness (2-DISJ)



Set disjointness (2-DISJ)



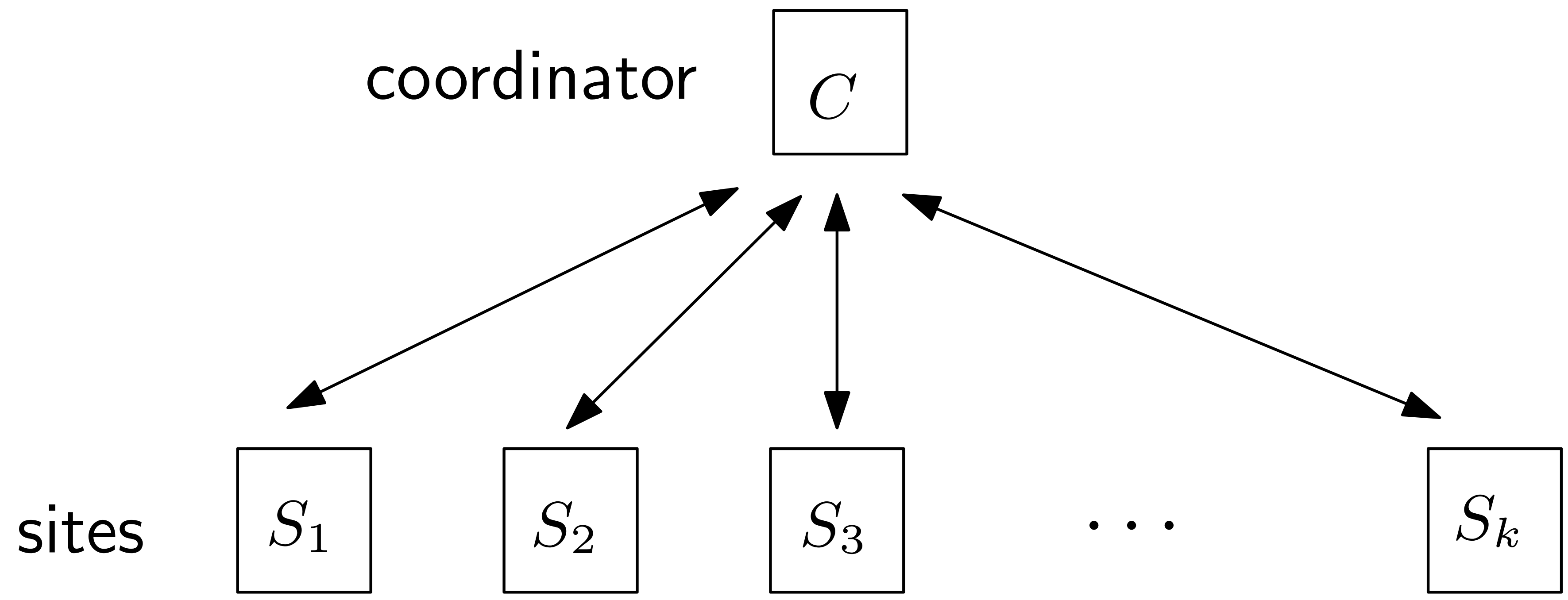
- A classical hard instance:

  Distribution μ: X and Y are both random subsets of size $\ell = (n+1)/4$ from $[n]$ such that $|X \cap Y| = 1$ w.p. β and $|X \cap Y| = 0$ w.p. $1 - \beta$.

  Razborov [1990] shows an $\Omega(n)$ for this hard distribution and error $\beta/100$.

Next step: Compose k-GAP-MAJ with 2-DISJ

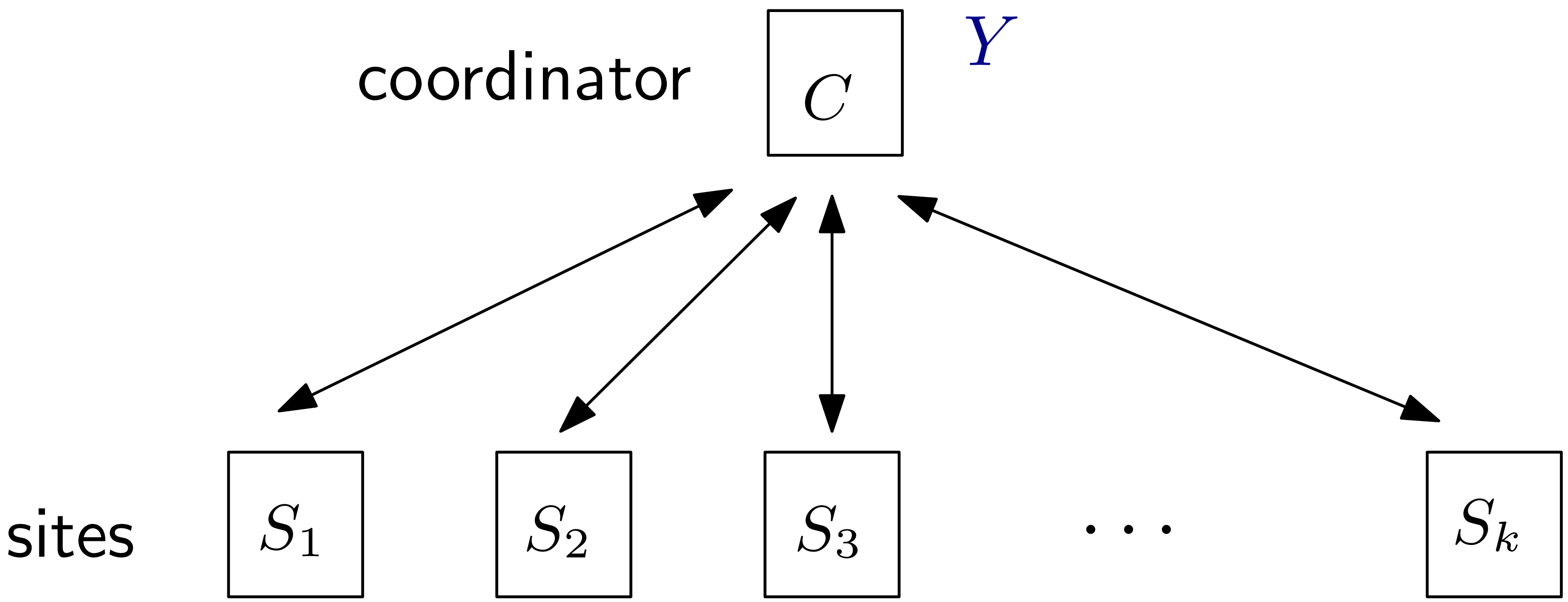
$n = \Theta(1/\varepsilon^2)$
$\ell = (n+1)/4$
$\beta = 1/k\varepsilon^2$



Next step: Compose k-GAP-MAJ with 2-DISJ

$n = \Theta(1/\varepsilon^2)$
$\ell = (n+1)/4$
$\beta = 1/k\varepsilon^2$

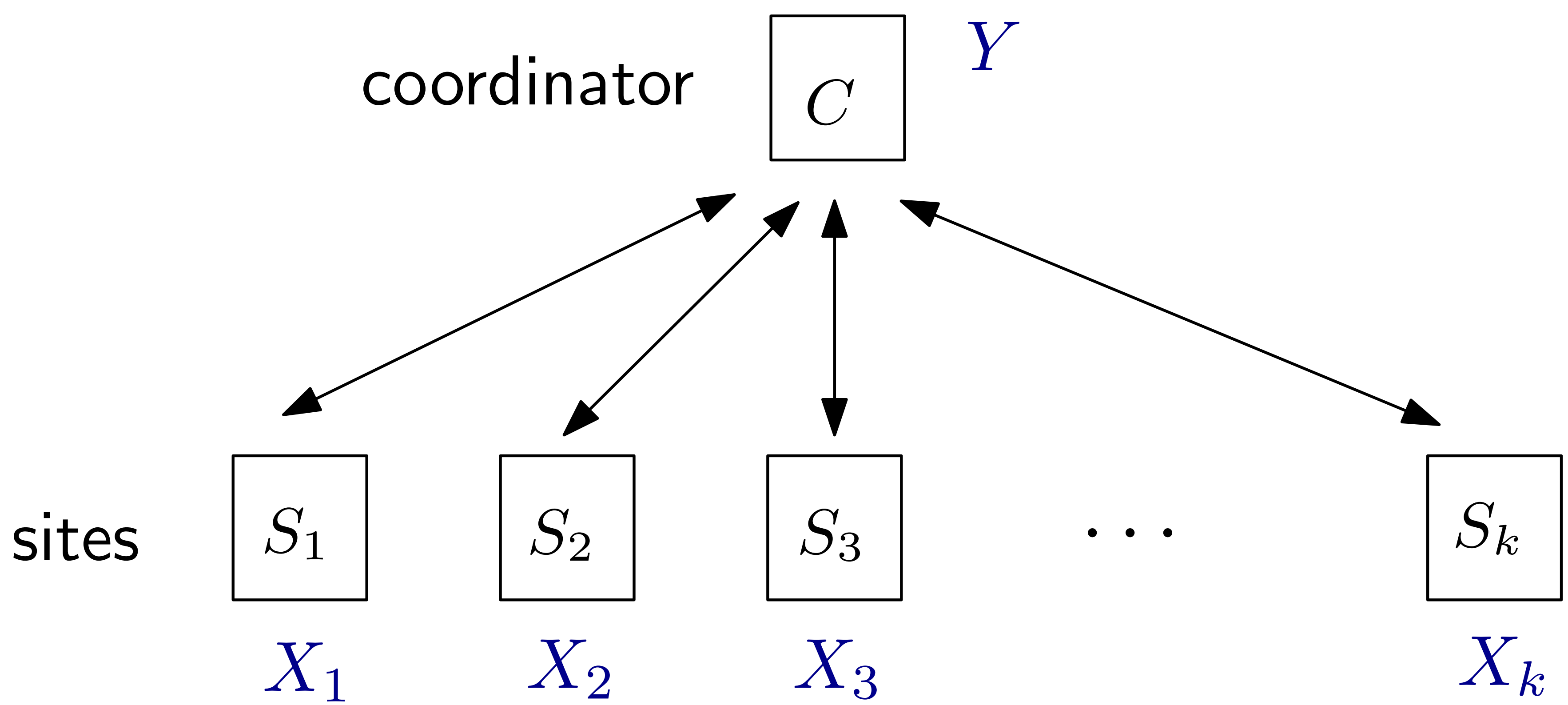
Step 1: Pick $Y = y \subset [n]$ of size ℓ uniformly at random



Next step: Compose k-GAP-MAJ with 2-DISJ

$n = \Theta(1/\varepsilon^2)$
$\ell = (n+1)/4$
$\beta = 1/k\varepsilon^2$

Step 1: Pick $Y = y \subset [n]$ of size ℓ uniformly at random

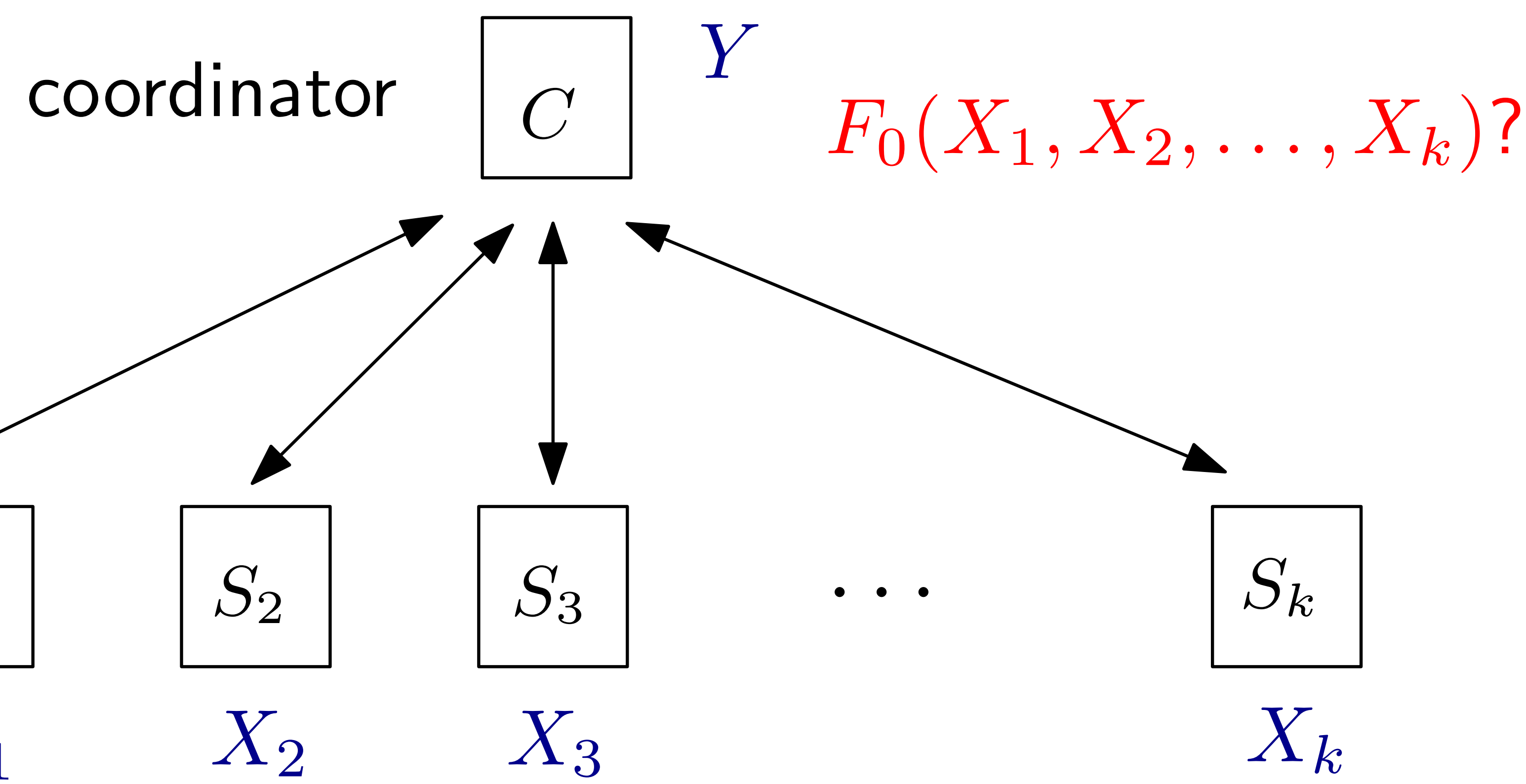


Step 2: Pick $X_1, \ldots, X_k \subset [n]$ indepedently and randomly from $\mu|_{Y=y}$

Next step: Compose k-GAP-MAJ with 2-DISJ

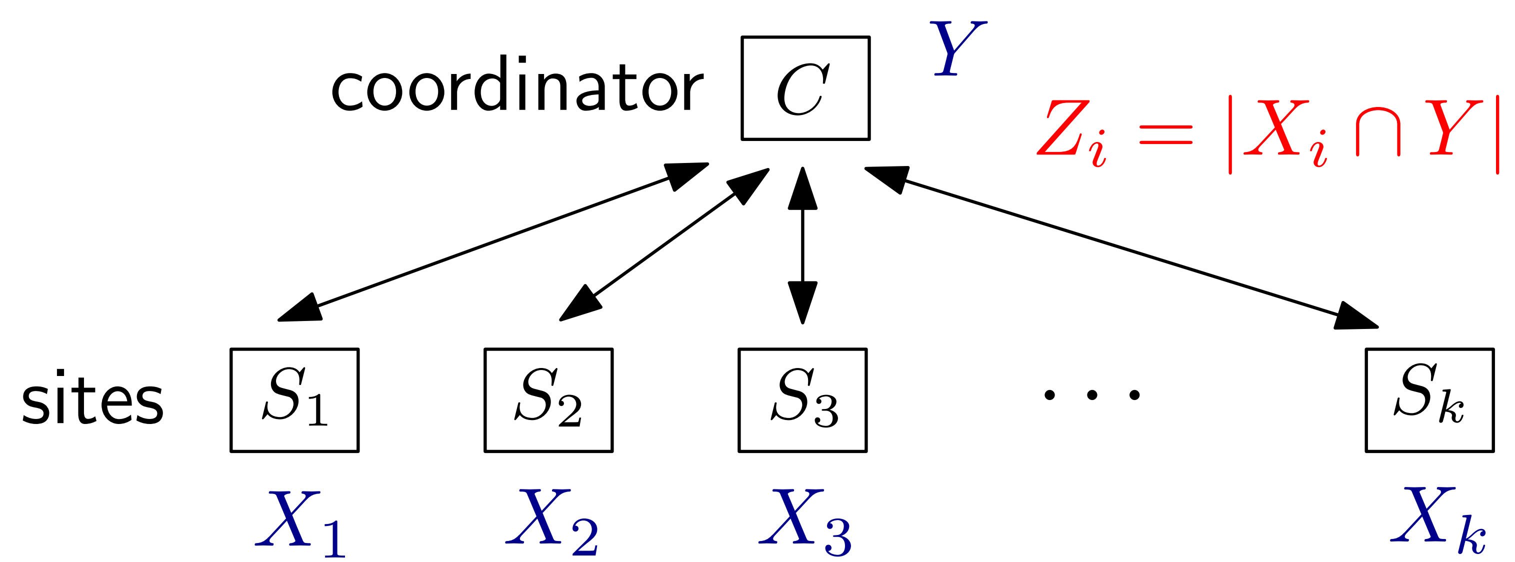
$n = \Theta(1/\varepsilon^2)$
$\ell = (n+1)/4$
$\beta = 1/k\varepsilon^2$

Step 1: Pick $Y = y \subset [n]$ of size ℓ uniformly at random



Step 2: Pick $X_1, \ldots, X_k \subset [n]$ indepedently and randomly from $\mu|_{Y=y}$

The proof



$$Z_i = |X_i \cap Y| \begin{cases} 1 & \text{w.p. } \beta \\ 0 & \text{w.p. } 1-\beta \end{cases}$$

$F_0(X_1, X_2, \ldots, X_k) \iff k$-GAP-MAJ$(Z_1, Z_2, \ldots, Z_k)$

$(Z_i = |X_i \cap Y|)$

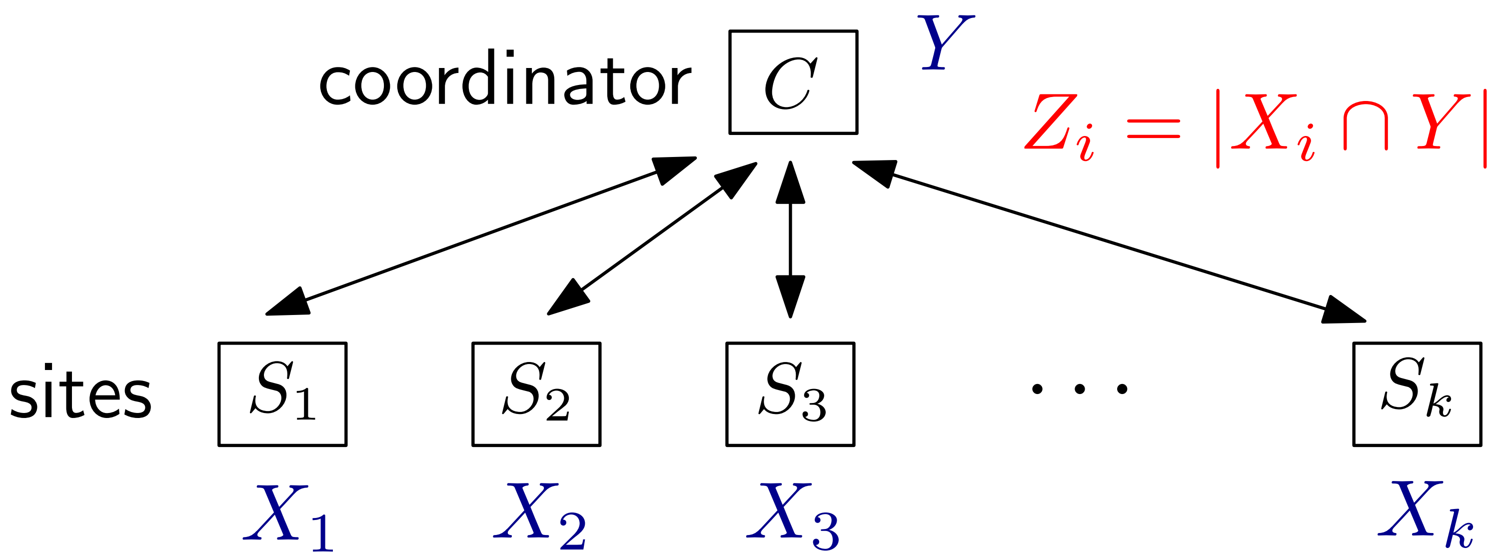
$\iff$ learn $\Omega(k)$ Z_i's well

(by Lemma 1)

$\iff$ need $\Omega(k/\varepsilon^2)$ bits

(learning each $Z_i = |X_i \cap Y|$ well needs $\Omega(n) = \Omega(1/\varepsilon^2)$ bits, by 2-DISJ)

The proof



$Z_i = |X_i \cap Y| \begin{cases} 1 & \text{w.p. } \beta \\ 0 & \text{w.p. } 1-\beta \end{cases}$

$F_0(X_1, X_2, \ldots, X_k) \iff k\text{-GAP-MAJ}(Z_1, Z_2, \ldots, Z_k)$

$(Z_i = |X_i \cap Y|)$

$\iff$ learn $\Omega(k)$ Z_i's well

(by Lemma 1)

$\iff$ need $\Omega(k/\varepsilon^2)$ bits

(learning each $Z_i = |X_i \cap Y|$ well needs $\Omega(n) = \Omega(1/\varepsilon^2)$ bits, by 2-DISJ)

Q.E.D.

Proof sketch of Lemma 1

Lemma 1: If a protocol $\mathcal{P}$ computes k-GAP-MAJ correctly w.p. 0.9999, then w.p. $\Omega(1)$, for $\Omega(k)$ Z_i's, we have $H(Z_i \mid \Pi) \leq H_b(0.01\beta)$.

Proof:

1. Suppose Π does not satisfy this.
2. Since the Z_i are independent given Π, $\sum_{i=1}^{k} Z_i \mid \Pi$ is a sum of independent Bernoulli random variables.
3. Since most $H(Z_i \mid \Pi)$ are large, by anti-concentration, both of the following events occur with constant probability:
   - $\sum_{i=1}^{k} Z_i \mid \Pi > \beta k + \sqrt{\beta k}$,
   - $\sum_{i=1}^{k} Z_i \mid \Pi < \beta k - \sqrt{\beta k}$.
4. So $\mathcal{P}$ can't succeed with large probability.

F_2 lower bound

The F_2 problem

We have k sites $S_1, S_2, \ldots, S_k$. S_i holds a set X_i.

Our goal: compute $F_2(\cup_{i \in k} X_i)$ up to $(1+\varepsilon)$-approximation.

2		2
9		9
7		7
2	Join	2
4		4
2		2

What's the size of self-join?

Another fundamental problem in data analysis.

The F_2 problem

We have k sites $S_1, S_2, \ldots, S_k$. S_i holds a set X_i.

Our goal: compute $F_2(\cup_{i \in k} X_i)$ up to $(1+\varepsilon)$-approximation.

$\begin{pmatrix} 2 \\ 9 \\ 7 \\ 2 \\ 4 \\ 2 \end{pmatrix}$ Join $\begin{pmatrix} 2 \\ 9 \\ 7 \\ 2 \\ 4 \\ 2 \end{pmatrix}$

What's the size of self-join?

Another fundamental problem in data analysis.

Previous UB: $\tilde{O}(k^2/\varepsilon + k^{1.5}/\varepsilon^3)$
(Cormode, Muthu, Yi 2008)

Our UB: $\tilde{O}(k/\text{poly}(\varepsilon))$, one way protocol
Holds in the dynamic case.

Previous LB: $\Omega(k)$ (Cormode, Muthu, Yi, 2008)
Our LB: $\tilde{\Omega}(k/\varepsilon^2)$.
Holds in the static and blackboard case.

The F_2 problem

We have k sites $S_1, S_2, \ldots, S_k$. S_i holds a set X_i.

Our goal: compute $F_2(\cup_{i \in k} X_i)$ up to $(1+\varepsilon)$-approximation.

$2, 9, 7, 2, 4, 2$ Join $2, 9, 7, 2, 4, 2$

What's the size of self-join?

Another fundamental problem in data analysis.

Previous UB: $\tilde{O}(k^2/\varepsilon + k^{1.5}/\varepsilon^3)$ (Cormode, Muthu, Yi 2008)

Our UB: $\tilde{O}(k/\mathrm{poly}(\varepsilon))$, one way protocol
Holds in the dynamic case.

Almost Tight!

Previous LB: $\Omega(k)$ (Cormode, Muthu, Yi, 2008)
Our LB: $\tilde{\Omega}(k/\varepsilon^2)$.
Holds in the static and blackboard case.

A quick glance: $(1+\varepsilon)$-approximation F_2

- 2-**party gap-hamming**: Alice has $X = \{X_1, X_2, \ldots, X_{1/\varepsilon^2}\}$, Bob has $Y = \{Y_1, Y_2, \ldots, Y_{1/\varepsilon^2}\}$. They want to compute:

$$\mathsf{GHD}(X,Y) = \begin{cases} 0, & \text{if } \sum_{i \in [1/\varepsilon^2]} X_i \oplus Y_i \le 1/2\varepsilon^2 - 1/\varepsilon, \\ 1, & \text{if } \sum_{i \in [1/\varepsilon^2]} X_i \oplus Y_i \ge 1/2\varepsilon^2 + 1/\varepsilon, \\ *, & \text{otherwise,} \end{cases}$$

where "$*$" means that the answer can be arbitrary.

A quick glance: $(1+\varepsilon)$-approximation F_2

- 2-**party gap-hamming**: Alice has $X = \{X_1, X_2, \ldots, X_{1/\varepsilon^2}\}$, Bob has $Y = \{Y_1, Y_2, \ldots, Y_{1/\varepsilon^2}\}$. They want to compute:

$$\mathsf{GHD}(X, Y) = \begin{cases} 0, & \text{if } \sum_{i \in [1/\varepsilon^2]} X_i \oplus Y_i \le 1/2\varepsilon^2 - 1/\varepsilon, \\ 1, & \text{if } \sum_{i \in [1/\varepsilon^2]} X_i \oplus Y_i \ge 1/2\varepsilon^2 + 1/\varepsilon, \\ *, & \text{otherwise,} \end{cases}$$

  where "$*$" means that the answer can be arbitrary.

- k-**DISJ**: We have k sites $S_1, S_2, \ldots, S_k$. S_i holds a set Z_i. We promise that either Z_i $(i = 1, \ldots, k)$ are all disjoint, or they intersect on one element and the rest are all disjoint (sun-flower).

  The goal is to find out which is the case.

A quick glance: $(1+\varepsilon)$-approximation F_2

- 2-**party gap-hamming**: Alice has $X = \{X_1, X_2, \ldots, X_{1/\varepsilon^2}\}$, Bob has $Y = \{Y_1, Y_2, \ldots, Y_{1/\varepsilon^2}\}$. They want to compute:

$$\mathsf{GHD}(X,Y) = \begin{cases} 0, & \text{if } \sum_{i \in [1/\varepsilon^2]} X_i \oplus Y_i \le 1/2\varepsilon^2 - 1/\varepsilon, \\ 1, & \text{if } \sum_{i \in [1/\varepsilon^2]} X_i \oplus Y_i \ge 1/2\varepsilon^2 + 1/\varepsilon, \\ *, & \text{otherwise,} \end{cases}$$

where "$*$" means that the answer can be arbitrary.

k-**XOR**

2 copies

- k-**DISJ**: We have k sites $S_1, S_2, \ldots, S_k$. S_i holds a set Z_i. We promise that either Z_i $(i = 1, \ldots, k)$ are all disjoint, or they intersect on one element and the rest are all disjoint (sun-flower).

The goal is to find out which is the case.

A quick glance: $(1+\varepsilon)$-approximation F_2

- 2-**party gap-hamming**: Alice has $X = \{X_1, X_2, \ldots, X_{1/\varepsilon^2}\}$, Bob has $Y = \{Y_1, Y_2, \ldots, Y_{1/\varepsilon^2}\}$. They want to compute:

$$\mathrm{GHD}(X,Y) = \begin{cases} 0, & \text{if } \sum_{i \in [1/\varepsilon^2]} X_i \oplus Y_i \le 1/2\varepsilon^2 - 1/\varepsilon, \\ 1, & \text{if } \sum_{i \in [1/\varepsilon^2]} X_i \oplus Y_i \ge 1/2\varepsilon^2 + 1/\varepsilon, \\ *, & \text{otherwise,} \end{cases}$$

where "$*$" means that the answer can be arbitrary.

compose via information cost $\longrightarrow$ k-**BTA**

k-**XOR**

2 copies

- k-**DISJ**: We have k sites $S_1, S_2, \ldots, S_k$. S_i holds a set Z_i. We promise that either Z_i $(i = 1, \ldots, k)$ are all disjoint, or they intersect on one element and the rest are all disjoint (sun-flower).

The goal is to find out which is the case.

A quick glance: $(1+\varepsilon)$-approximation F_2

- 2-**party gap-hamming**: Alice has $X = \{X_1, X_2, \ldots, X_{1/\varepsilon^2}\}$, Bob has $Y = \{Y_1, Y_2, \ldots, Y_{1/\varepsilon^2}\}$. They want to compute:

$$\mathsf{GHD}(X,Y) = \begin{cases} 0, & \text{if } \sum_{i\in[1/\varepsilon^2]} X_i \oplus Y_i \le 1/2\varepsilon^2 - 1/\varepsilon, \\ 1, & \text{if } \sum_{i\in[1/\varepsilon^2]} X_i \oplus Y_i \ge 1/2\varepsilon^2 + 1/\varepsilon, \\ *, & \text{otherwise,} \end{cases}$$

where "$*$" means that the answer can be arbitrary.

compose via information cost $\longrightarrow$ k-**BTA** $\quad \mathsf{CC}(k\text{-BTA}) = \tilde{\Omega}(k/\varepsilon^2)$

k-**XOR**

2 copies

- k-**DISJ**: We have k sites $S_1, S_2, \ldots, S_k$. S_i holds a set Z_i. We promise that either Z_i $(i = 1, \ldots, k)$ are all disjoint, or they intersect on one element and the rest are all disjoint (sun-flower).

  The goal is to find out which is the case.

A quick glance: $(1+\varepsilon)$-approximation F_2

- 2-**party gap-hamming**: Alice has $X = \{X_1, X_2, \ldots, X_{1/\varepsilon^2}\}$, Bob has $Y = \{Y_1, Y_2, \ldots, Y_{1/\varepsilon^2}\}$. They want to compute:

$$\text{GHD}(X,Y) = \begin{cases} 0, & \text{if } \sum_{i \in [1/\varepsilon^2]} X_i \oplus Y_i \leq 1/2\varepsilon^2 - 1/\varepsilon, \\ 1, & \text{if } \sum_{i \in [1/\varepsilon^2]} X_i \oplus Y_i \geq 1/2\varepsilon^2 + 1/\varepsilon, \\ *, & \text{otherwise,} \end{cases}$$

where "$*$" means that the answer can be arbitrary.

compose via information cost $\longrightarrow$ k-**BTA** CC(k-BTA) $= \tilde{\Omega}(k/\varepsilon^2)$

k-**XOR**

2 copies

- k-**DISJ**: We have k sites $S_1, S_2, \ldots, S_k$. S_i holds a set Z_i. We promise that either Z_i $(i = 1, \ldots, k)$ are all disjoint, or they intersect on one element and the rest are all disjoint (sun-flower).

The goal is to find out which is the case.

- Finally, we reduce F_2 to k-BTA.

The end

THANK YOU

Q and A