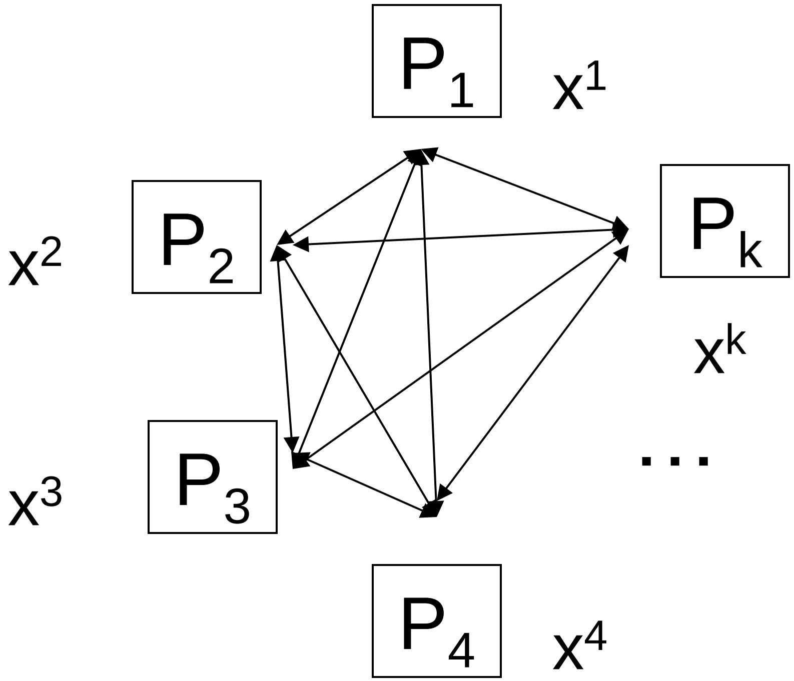
Tutorial: Message Passing Communication Model

David Woodruff

IBM Almaden

k-party Number-In-Hand Model



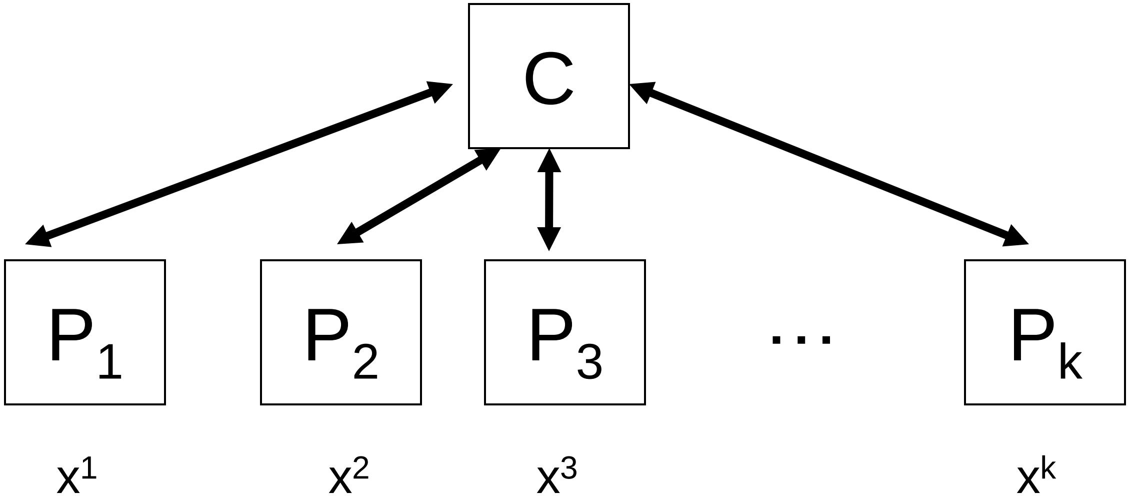
- Point-to-point communication

- Protocol transcript determines who speaks next

Goals:

- compute a function $f(x^1, \ldots, x^k)$
- minimize communication complexity

k-party Number-In-Hand Model



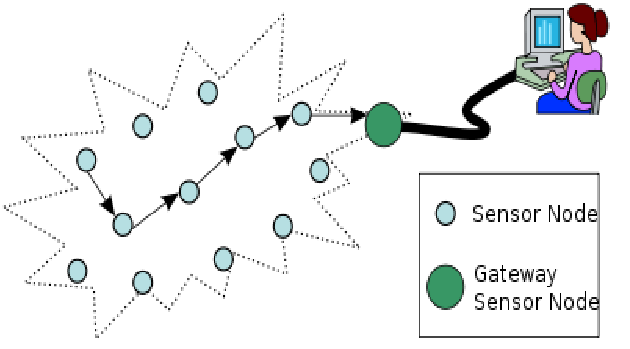
Convenient to introduce a "coordinator" C who may or may not have an input

All communication goes through the coordinator

Communication only affected by a factor of 2 (plus one word per message)

Model Motivation

- Data distributed and stored in the cloud
  - For speed
  - Just doesn't fit on one device



- Sensor networks / Network routers
  - Communication very power-intensive
  - Bandwidth limitations
- Distributed functional monitoring
  - Continuously monitor a statistic of distributed data
  - Don't want to keep sending all data to one place

Randomized Communication Complexity

- Randomized communication complexity R(f) of a function f:

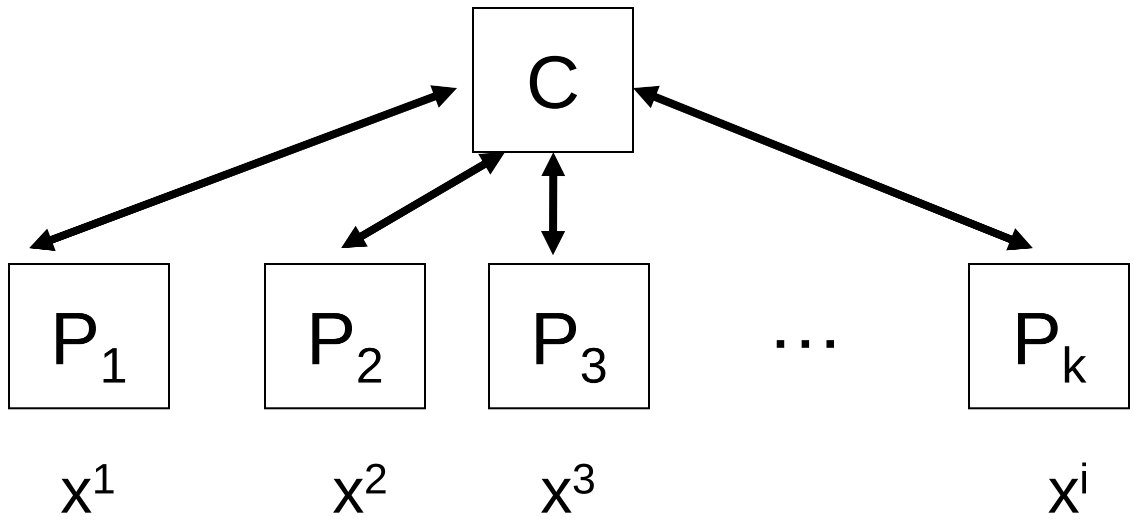
  - The communication cost of a protocol is the sum of all individual message lengths, maximized over all inputs and random coins

  - R(f) is the minimal cost of a protocol, which for every set of inputs, fails in computing f with probability < 1/3

Talk Outline

- Database Problems
- Graph Problems
- Linear-Algebra Problems
- Recent Work / Conclusions

Database Problems



Some well-studied problems

- Server i has x^i
- $x = x^1 + x^2 + \ldots + x^k$
- $f(x) = |x|_p = (\Sigma_i\, x_i^p)^{1/p}$
- for binary vectors x^i, $|x|_0$ is the number of distinct values (focus of this talk)

Exact Number of Distinct Elements

- $\Omega(n)$ randomized complexity for exact computation of $|x|_0$
- Lower bound holds already for 2 players

$S \subseteq [n]$

$T \subseteq [n]$

- Reduction from 2-Player Set-Disjointness (DISJ)
  - Either $|S \cap T| = 0$ or $|S \cap T| = 1$
  - $|S \cap T| = 1 \rightarrow$ DISJ(S,T) = 1, $|S \cap T| = 0 \rightarrow$ DISJ(S,T) = 0
  - [KS, R] $\Omega(n)$ communication
  - $|x|_0 = |S| + |T| - |S \cap T|$

Approximate Answers

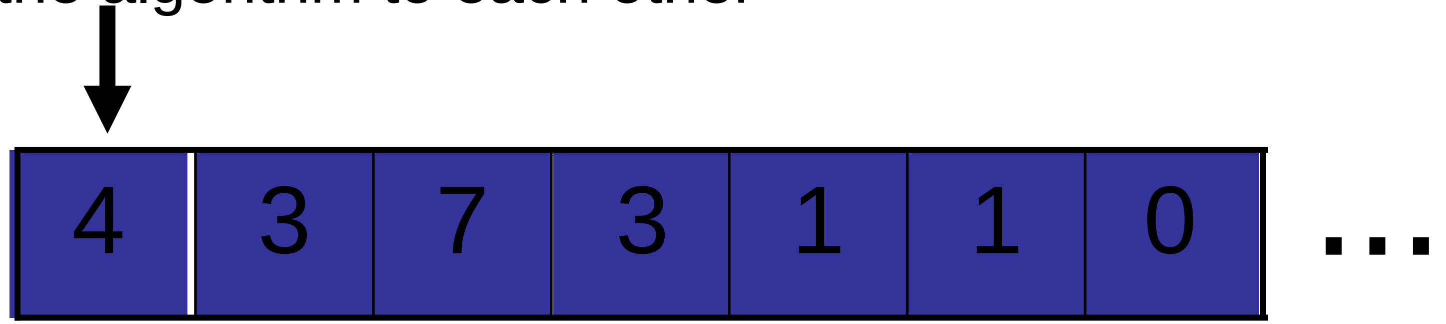
Output an estimate f(x) with $f(x) \in (1 \pm \varepsilon) |x|_0$

What is the randomized communication cost as a function of k, ε, and n?

Note that understanding the dependence on ε is critical, e.g., $\varepsilon < .01$

An Upper Bound

- Player i interprets its input as the i-th set in a data stream
- Players run a data stream algorithm, and pass the state of the algorithm to each other



- There is a data stream algorithm for estimating # of distinct elements using $O(1/\varepsilon^2 + \log n)$ bits of space
- Gives a protocol with $O(k/\varepsilon^2 + k \log n)$ communication

Lower Bound

- This approach is optimal!

- We show an $\Omega(k/\varepsilon^2 + k \log n)$ communication lower bound

- First show an $\Omega(k/\varepsilon^2)$ bound [W, Zhang 12], see also [Phillips, Verbin, Zhang 12]
  - Start with a simpler problem GAP-THRESHOLD

Lower Bound for Approximate $|x|_0$

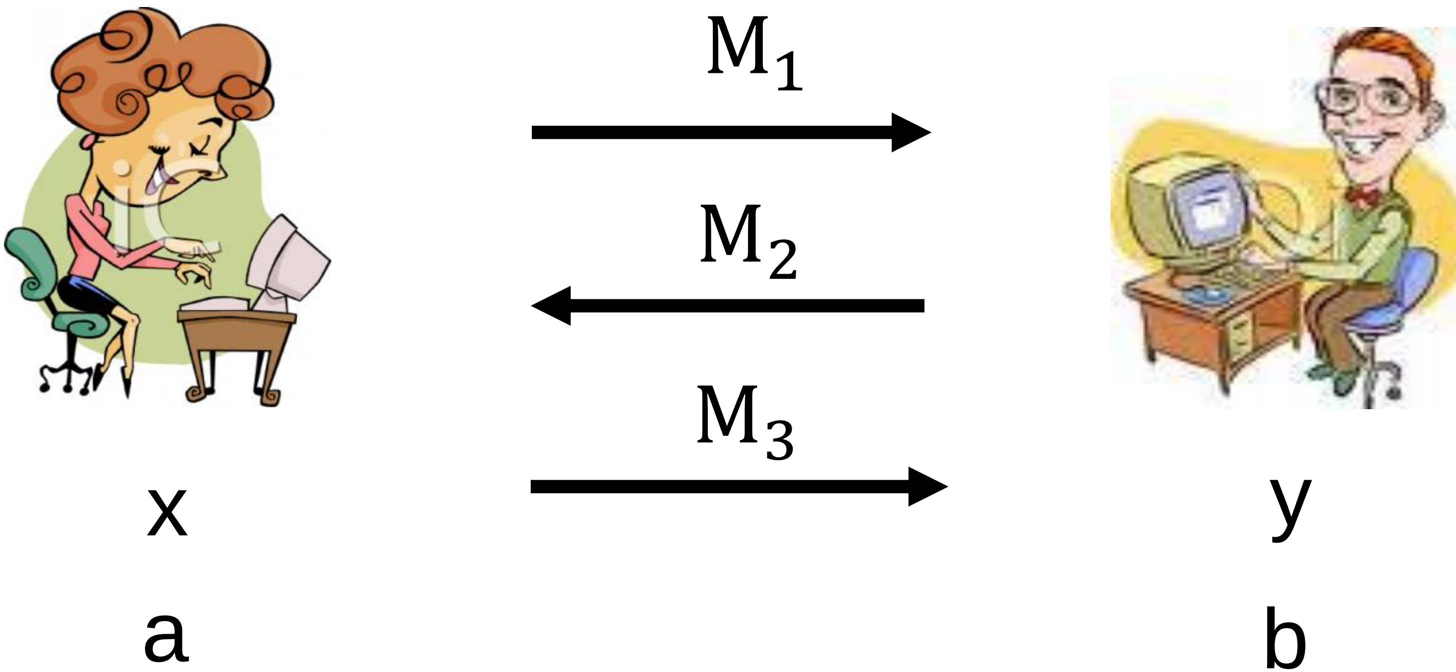
- GAP-THRESHOLD problem:
  - Player P_i holds a bit Z_i
  - Z_i are i.i.d. Bernoulli(1/2)
  - Decide if

    $\sum_{i=1}^{k} Z_i > k/2 + k^{1/2}$ or $\sum_{i=1}^{k} Z_i < k/2 - k^{1/2}$

    Otherwise don't care (distributional problem)

- Intuitively $\Omega(k)$ bits of communication is required
  - Sampling doesn't work…
  - How to prove such a statement??

Rectangle Property of Protocols



- If inputs (x,y) and (a,b) cause the same transcript, then so do (x,b) and (a,y)

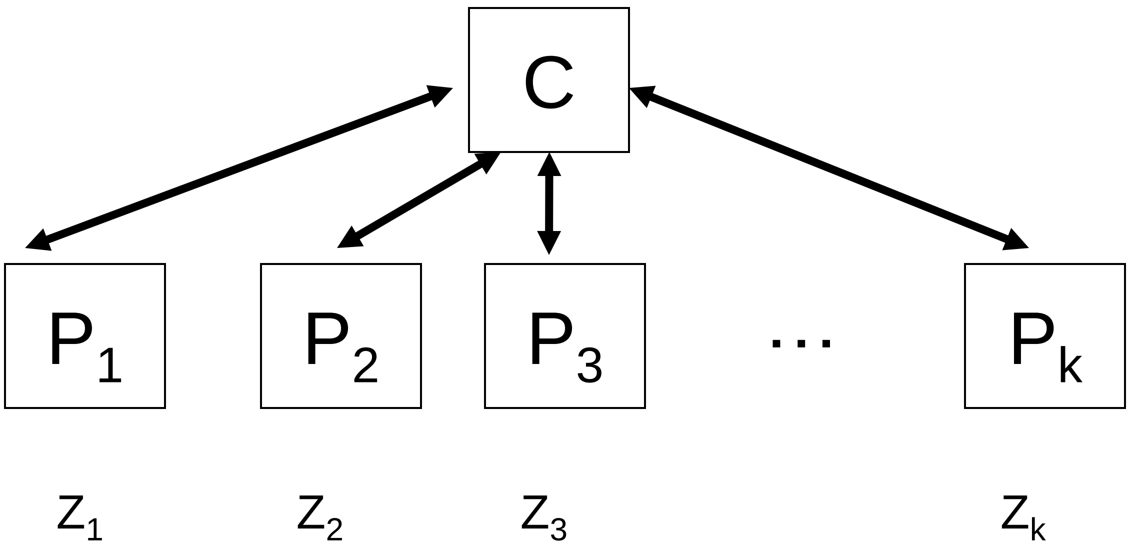
- For randomized protocols,

  Pr[seeing a transcript τ given inputs a,b] = $p(a, \tau) \cdot q(b, \tau)$

Rectangle Property

- Claim: for any protocol transcript τ, it holds that $Z_1, Z_2, \ldots, Z_k$ are independent conditioned on τ
- Can assume players are deterministic by Yao's minimax principle
- The input vector Z in $\{0,1\}^k$ giving rise to a transcript τ is a combinatorial rectangle: $S = S_1 \times S_2 \times \ldots \times S_k$ where S_i in $\{0,1\}$
- Since the Z_i are i.i.d. Bernoulli(1/2), conditioned on being in S, they are still independent!

GAP-THRESHOLD



- The Z_i are i.i.d. Bernoulli(1/2)

- Coordinator wants to decide if:

$\sum_{i=1}^{k} Z_i > k/2 + k^{1/2}$ or $\sum_{i=1}^{k} Z_i < k/2 - k^{1/2}$

- By independence of the $Z_i \mid \tau$, it is equivalent to fixing some Z_i to be 0 or 1, and the remaining Z_i to be Bernoulli(1/2)

The Proof

- Lemma [Unbiased Conditional Expectation]: W.pr. 2/3, over the transcript τ,

$$|\mathbb{E}[\sum_{i=1}^{k} Z_i \mid \tau] - k/2| < 100\, k^{1/2}$$

- Otherwise, since $\mathrm{Var}[\sum_{i=1}^{k} Z_i \mid \tau] < k$ for any τ, by Chebyshev's inequality, w.p.r. > 1/2,

$$|\sum_{i=1}^{k} Z_i - k/2| > 50k^{1/2}$$

contradicting concentration

- Lemma [Lots of Randomness After Conditioning]: If the communication is o(k), then w.pr. 1-o(1), over the transcript τ, for a 1-o(1) fraction of the indices i,

$$Z_i \mid \tau \text{ is Bernoulli}(1/2)$$

The Proof Continued

- Let's condition on a τ satisfying the previous two lemmas

- Lemma [Anti-Concentration]:

W.pr. .001, over the $Z_i \mid \tau$

$$E[\sum_{i=1}^k Z_i \mid \tau] - \sum_{i=1}^k Z_i \mid \tau > 100\, k^{1/2}$$

W.pr. .001, over the $Z_i \mid \tau$

$$\sum_{i=1}^k Z_i \mid \tau - E[\sum_{i=1}^k Z_i \mid \tau] > 100\, k^{1/2}$$

- These follow by anti-concentration

- So the protocol fails with this probability

Generalizations

- Generalizes to: Z_i are i.i.d. Bernoulli(β)

- Coordinator wants to decide if:

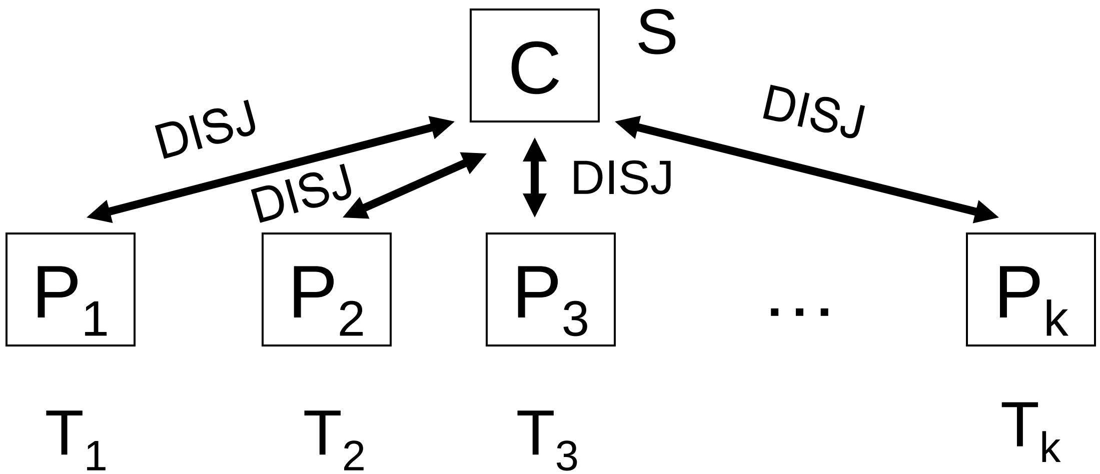
$$\sum_{i=1}^{k} Z_i > \beta k + (\beta k)^{1/2} \text{ or } \sum_{i=1}^{k} Z_i < \beta k - (\beta k)^{1/2}$$

- When the players have internal randomness, the proof generalizes: any successful protocol must satisfy:

$$\Pr_{\tau} [\text{for } 1\text{-}o(1) \text{ fraction of indices } i,\ H(Z_i \mid \tau) = o(1)] > 2/3$$

- How to get a lower bound for approximating $|x|_0$?

Composition Idea



- Give the coordinator a random set S from {1, 2, …, m}
- If $Z_i = 1$, give P_i a random set T_i so that $DISJ(S,T_i) = 1$, else give P_i a random set T_i so that $DISJ(S,T_i) = 0$
- Is $\sum_{i=1}^{k} DISJ(S,T_i) > k/2 + k^{1/2}$ or $\sum_{i=1}^{k} DISJ(S, T_i) < k/2 - k^{1/2}$?
  - Equivalently, is $\sum_{i=1}^{k} Z_i > k/2 + k^{1/2}$ or $\sum_{i=1}^{k} Z_i < k/2 - k^{1/2}$
- Our Result: total communication is $\Omega(mk)$

Composition Idea Continued

- For this composed problem, a correct protocol satisfies:
  $\Pr_{\tau}$ [for 1-o(1) fraction of indices i, $H(Z_i \mid \tau) = o(1)$] > 2/3

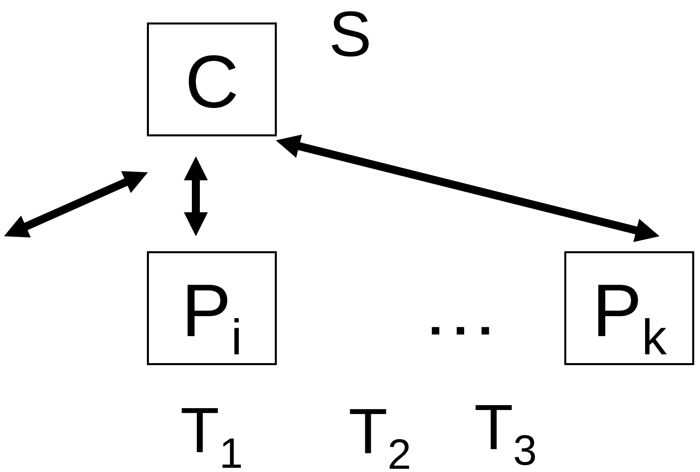
- Most DISJ instances are "solved" by the protocol

- How to formalize?

- Suppose the communication were o(km)

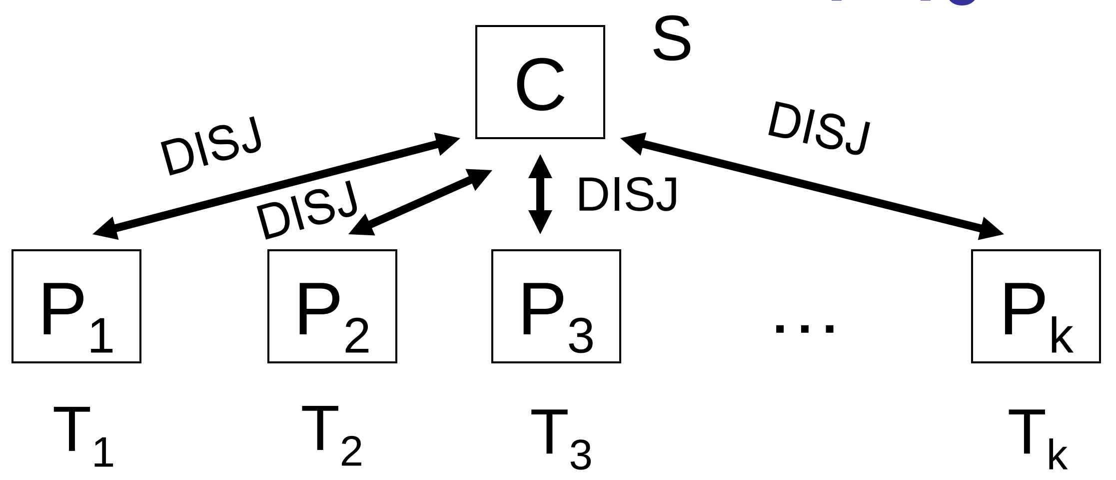
- By averaging, there is a player P_i so that
  - The communication between C and P_i is o(m)
  - $H(Z_i \mid \tau) = o(1)$ with large probability

The Punch Line



- Reduce to a 2-player problem!
- Let the two players in the 2-player DISJ problem be the coordinator C and P_i
- C can sample the inputs of all players P_j for j != i
- Run the multi-player protocol. Messages between C and P_j is sent, for j != i, can be simulated locally!
- So total communication is o(m) to solve DISJ with large probability, a contradiction!

Reduction to $|x|_0$



- $m = 1/\varepsilon^2$.

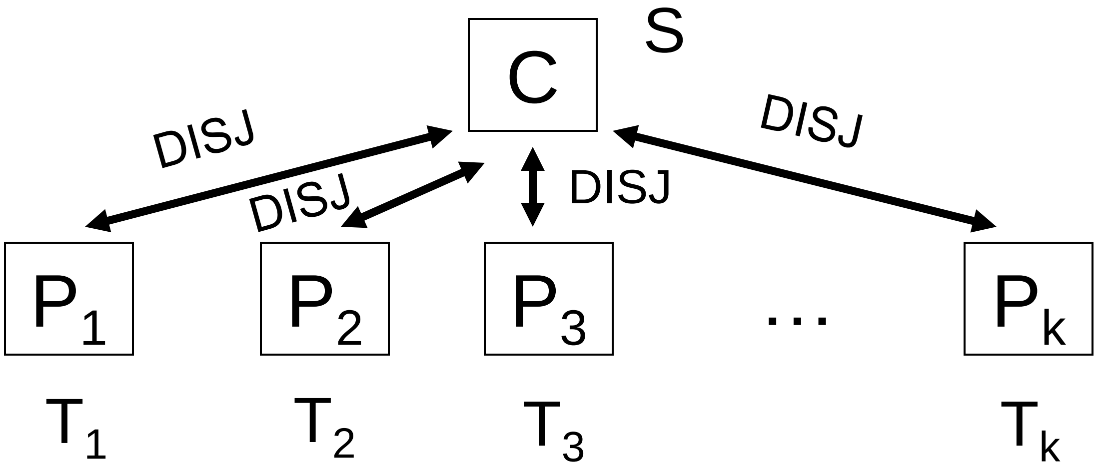
- Coordinator wants to decide if:

  $$\sum_{i=1}^{k} Z_i > \beta k + (\beta k)^{1/2} \text{ or } \sum_{i=1}^{k} Z_i < \beta k - (\beta k)^{1/2}$$

  Set probability β of intersection to be $1/(4k\varepsilon^2)$

- Approximating $|x|_0$ up to $1+\varepsilon$ solves this problem

Reduction to $|x|_0$



- Coordinator replaces its input set with $[1/\varepsilon^2] \setminus S$

- If $DISJ(S,T_i) = 0$, then T_i is contained in $[1/\varepsilon^2] \setminus S$
- If $DISJ(S,T_i) = 1$, then T_i adds a new distinct item to $[1/\varepsilon^2] \setminus S$
  - If $DISJ(S,T_i) = 1$ and $DISJ(S,T_j) = 1$, they typically add different items

- So the number of distinct items is about $1/(2\varepsilon^2) + \sum_{i=1}^{k} Z_i$

Other Lower Bound for $|x|_0$

- Overall lower bound is $\Omega(k/ \varepsilon^2 + k \log n)$

- The k log n lower bound also a reduction to a 2-player problem [W, Zhang 14]

  – This time to a 2-player Equality problem (details omitted)

Talk Outline

- Database Problems
- Graph Problems
- Linear-Algebra Problems
- Recent Work / Conclusions

Graph Problems [W,Zhang13]

- Canonical hard-multiplayer problem for graph problems:

- n x k binary matrix A
  - Each player has a column of A
  - Is the number of rows with at least one 1 larger than n/2?

- Requires Ω(kn) bits of communication to solve with probability at least 2/3

Ω(kn) lower bound for connectivity and bipartiteness without edge duplications

Talk Outline

- Database Problems
- Graph Problems
- Linear-Algebra Problems
- Recent Work / Conclusions

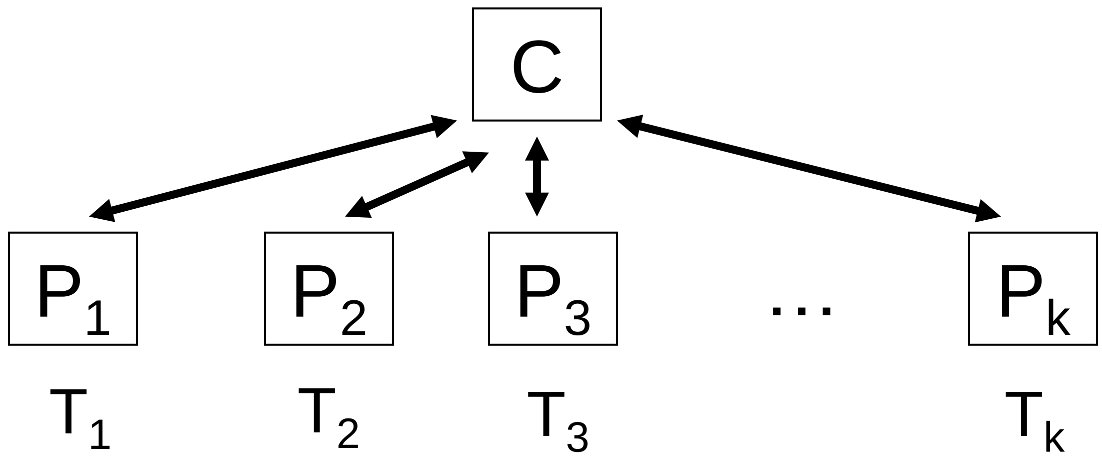
Linear Algebra [Li,Sun,Wang,W]

- k players each have an n x n matrix in a finite field of p elements
- Players want to know if the sum of their matrices is invertible
- Randomized $\Omega(kn^2 \log p)$ communication lower bound
- Same lower bound for rank, solving linear equations
- Open question: lower bound over the reals?

Talk Outline

- Database Problems
- Graph Problems
- Linear-Algebra Problems
- Recent Work / Conclusions

Recent Work: Set Disjointness



- Each set $T_i \subseteq [m]$
- k-player Disjointness: is $T_1 \cap T_2 \cap \cdots \cap T_k = \emptyset$?
- Braverman et al. obtain Ω(km) lower bound
- Input distribution
  - random half of the items appear in all sets except a random one
  - random half the items independently occur in each T_i
  - with probability 1/2, make a random item occur in each T_i

Recent Work: Set Disjointness

Random rows

1 1 1 1 0 1 1 1 1 1
0 1 1 0 1 0 1 0 0 1
1 0 1 1 1 1 1 1 1 1
1 0 0 0 1 1 0 1 1 0
1 1 1 1 1 1 1 1 1 0
0 1 0 1 1 0 0 1 0 1
...

Each player has a column

- The coordinator can figure out which rows are random, but can't easily communicate this to the players
- Each player knows which positions in its column are zero, but can't easily communicate this to the coordinator
- Direct sum theorems with mixed information cost measure

Recent Work:Topology

- Chattopadhyay, Radhakrishnan, Rudra study multiplayer communication in topologies other than star topology
  - Obtain bounds that depend on 1-median of the network
- Chattopadhyay, Rudra
  - Only players at a subset of nodes have an input
  - Communication cost depends on Steiner tree cost

Conclusion

- Illustrated techniques for lower bounds for multiplayer communication via the distinct elements problem

- Many tight lower bounds known
  - Statistical problems (lp norms)
  - Graph problems
  - Linear algebra problems

- Open Questions and Future Directions
  - Rounds vs. communication
  - Connections to other models, e.g., MapReduce
  - Topology-sensitive problems