Algorithms for Distributed Functional Monitoring

Graham Cormode

AT&T Labs

S. Muthukrishnan

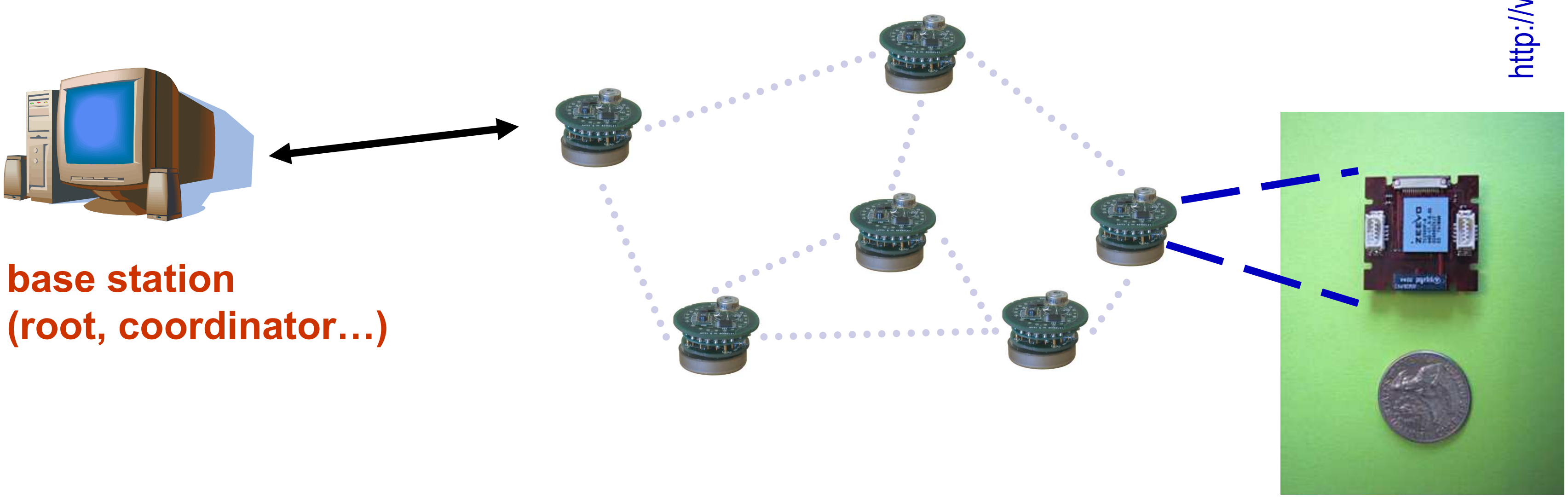
Google Research

Ke Yi

HKUST

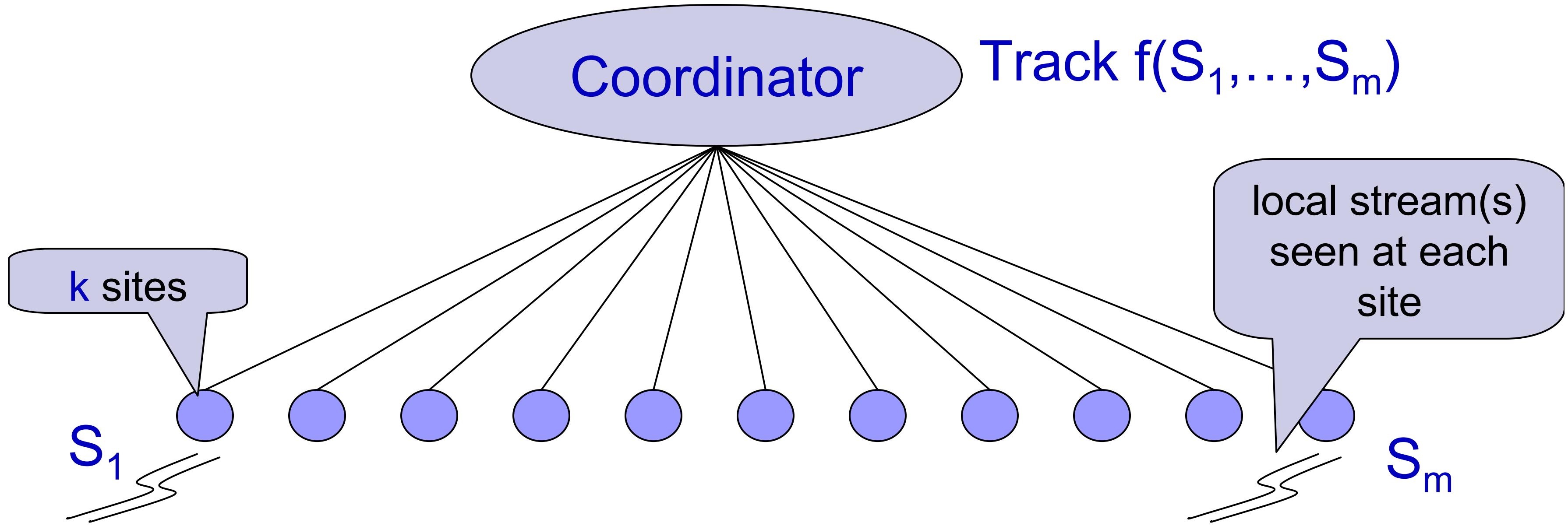
Sensor Networks

- Large number of remote, wireless sensors record environmental details, communicate back to base
- Want to monitor environment, and trigger alerts
  - Based on some complex function of *global* values
- Each sensor sees a continuous *stream* of values
- *Communication* is the major source of battery drain



http://www.intel.com/research/exploratory/motes.htm

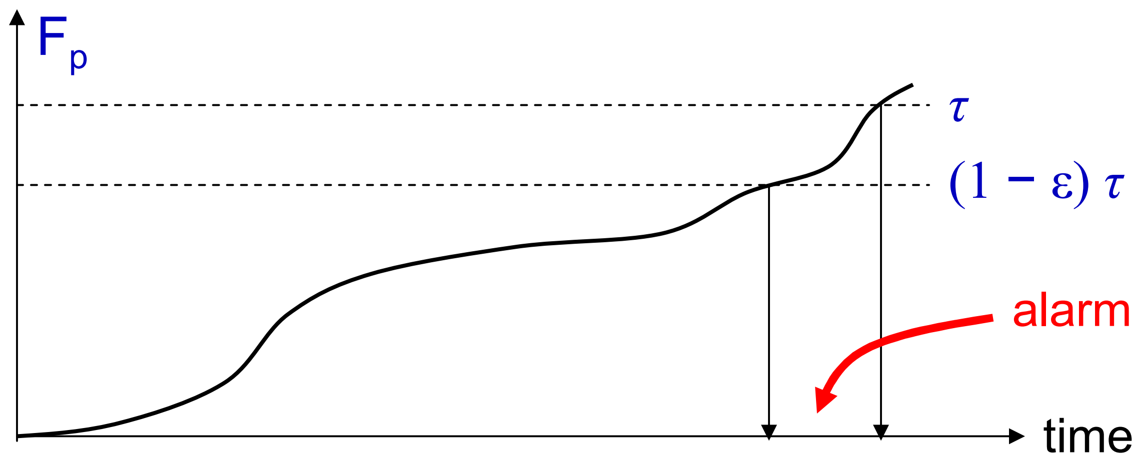
Continuous Distributed Model



- Other structures possible (e.g., hierarchical)
- Site-site communication only changes things by factor 2
- **Goal:** *Continuously track* (global) function over streams at the coordinator
- Here, study frequency moments: $F_p = \sum_i (f_i)^p$
  - f_i is the count of item i across all sites

Approximate Monitoring

- Must trigger alarm when $F_p > \tau$
- Cannot trigger alarm when $F_p < (1 - \varepsilon)\, \tau$



- Approximate is good enough for most applications.
- Contrast to “one-shot” version: coordinator initiates one-time approximate computation of F_p

General Algorithm for F_p

- Simple approach divides the current "slack" uniformly between sites
- Vector u_i represents total frequencies at round i
- Slack is $s_i = (\tau - ||u_i||_p^p)$, set threshold $t_i = s_i/2k^p$
- Each site j sees vector of updates v_{ij}, and monitors

$$|| u_i + v_{ij} ||_p^p - ||u_i||_p^p > t_i$$

  Sends a bit when threshold is exceeded
- When coordinator has received k bits, terminates round and collects u_{i+1}, computes and sends t_{i+1}.
  – $O(k)$ pieces of information sent per round
- Alert when $|| u_i ||_p^p > (1 - \varepsilon/2)\, \tau$

Analysis of General Algorithm

- By Jensen's inequality, $\|u_{i+1}\|_p^p - \| u_i \|_p^p < 2k^p t_i$
  - Since $t_i = s_i/2k^p$, we have $\| u_{i+1} \|_p^p < \tau$
- By convexity of the function $\| x + y \|_p^p - \| x \|_p^p$ for $p \geq 1$,

$$\|u_{i+1}\|_p^p - \| u_i \|_p^p \geq k\, t_i$$

- So $t_{i+1} \leq t_i (1 - k^{1-p}/2)$
  - $t_0 = \tau\, k^{-p}/2$, and halt when $t_i < \varepsilon\, \tau\, k^{-p}/2$
  - At most $O(k^{p-1} \log 1/\varepsilon)$ rounds
- Algorithm is correct (never exceeds τ without causing an alert), and has few rounds.

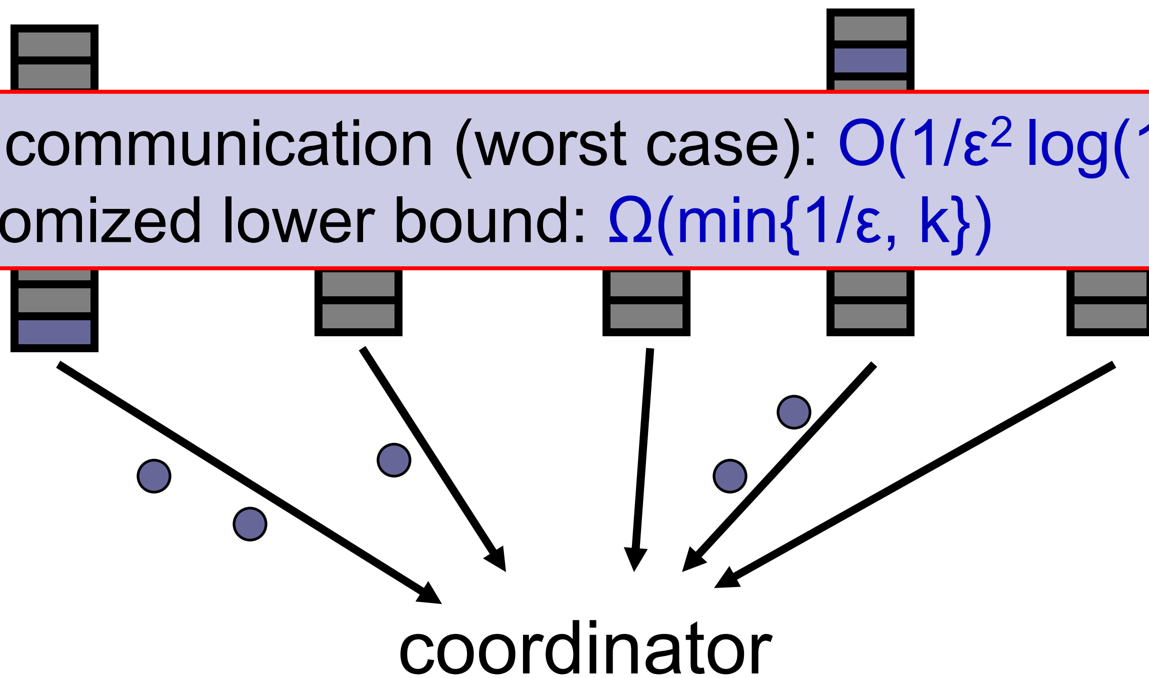
Application of General Algorithm

- F_1: $\|x\|_p^p$ is simply the sum of all updates
  - Don't even need to send $\|u_i\|_1$ or t_i values, these are implicit
  - Yields a simple, deterministic $O(k \log 1/\varepsilon)$ bits solution
- Deterministic lower bound for F_1: $\Omega(k \log 1/(\varepsilon k))$
  - Folklore lower bound for one-shot computation?
    Based on construction of sufficiently large 'fooling sets'
- F_2: use ε'-approximate sketches to communicate the vectors between sites
  - Need to set ε' so $\varepsilon' \| u_i + v_{i,j} \|_2^2 = O(t_i)$, forcing $\varepsilon' = O(\varepsilon/k^2)$
  - Gives a total cost of $\tilde{O}(k^6/\varepsilon^2)$
- F_p, $p>2$. Ganguly et al. sketches, cost $\tilde{O}(p\, \varepsilon^{-3} k^{2p+1} n^{1-2/p})$

Randomized F_1 Algorithm

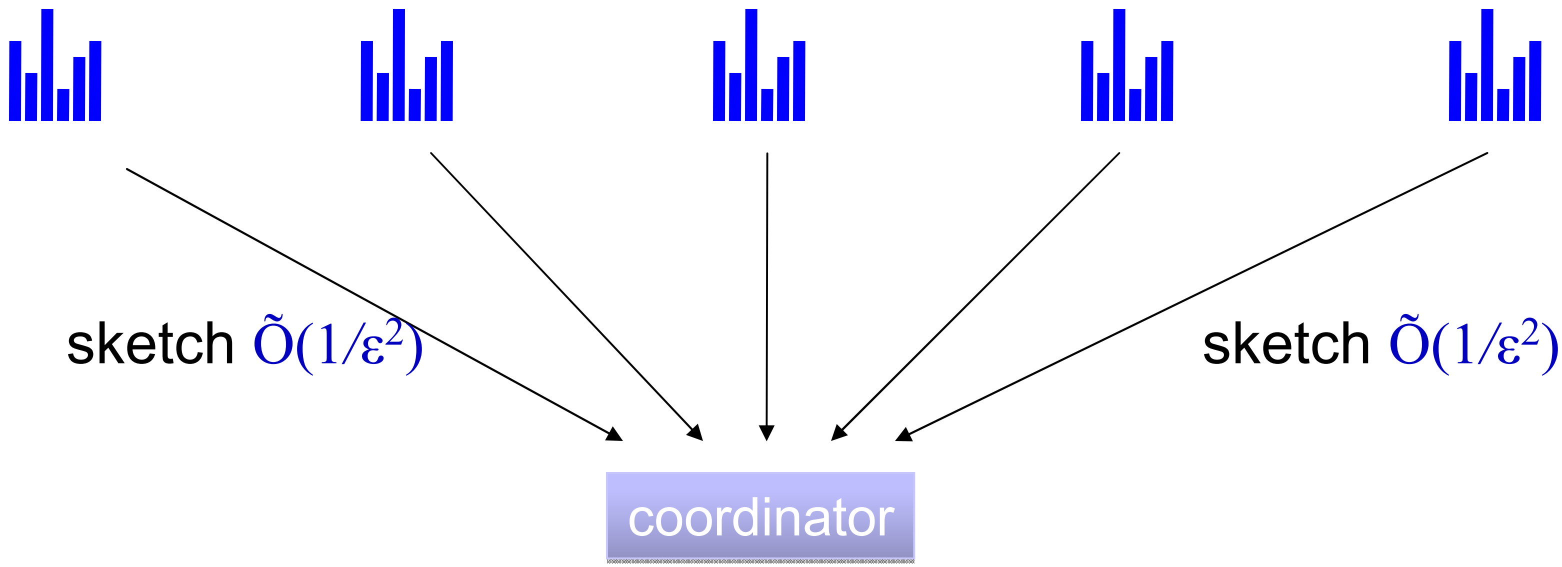
- At each site: for every $\varepsilon^2\tau/k$ items received, send a signal to coordinator with probability $1/k$
- Raise alarm when $1/\varepsilon^2$ signals received
  - By Chebyshev, constant probability of (two-sided) error
- Repeat $O(\log(1/\delta))$ times in parallel to reduce error prob

Total communication (worst case): $O(1/\varepsilon^2 \log(1/\delta))$
Randomized lower bound: $\Omega(\min\{1/\varepsilon, k\})$



F_2 Multi-Round Algorithm

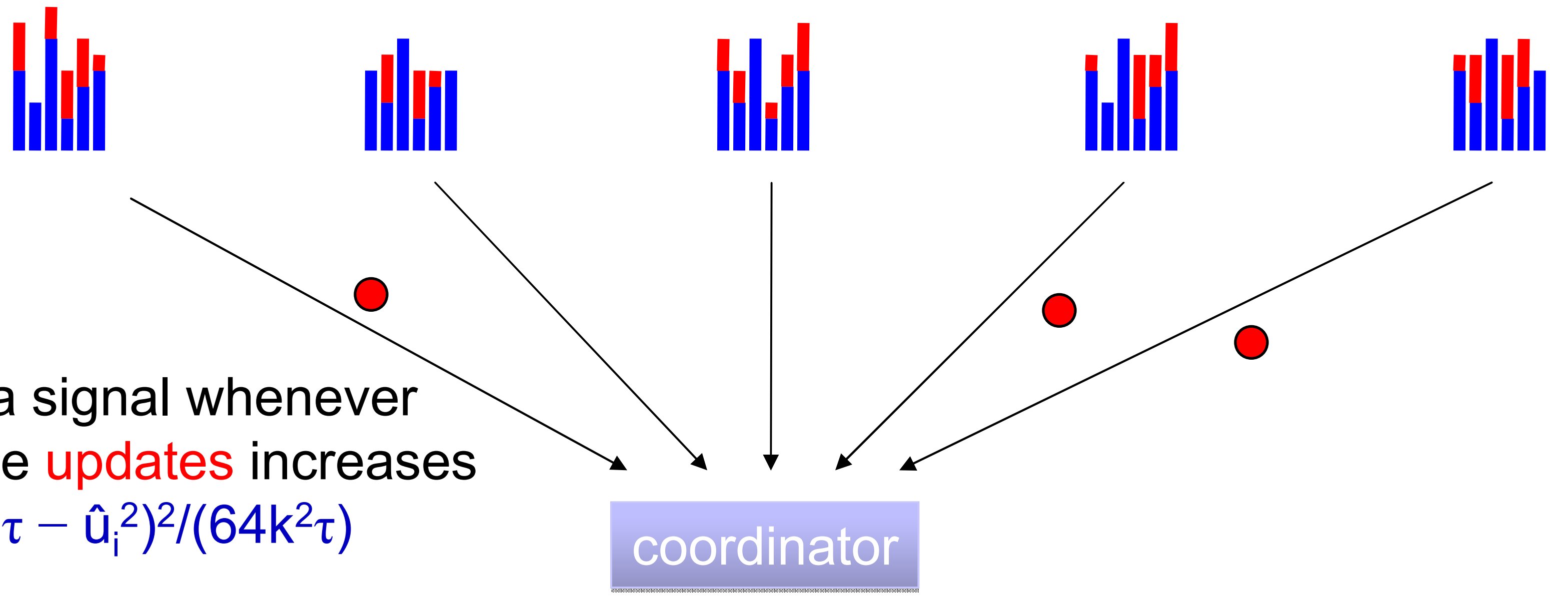
Beginning of a round: each site sends ε-accurate sketch



$\hat{u}^2$ = estimate for F_2

F_2 Multi-Round Algorithm

During a round:

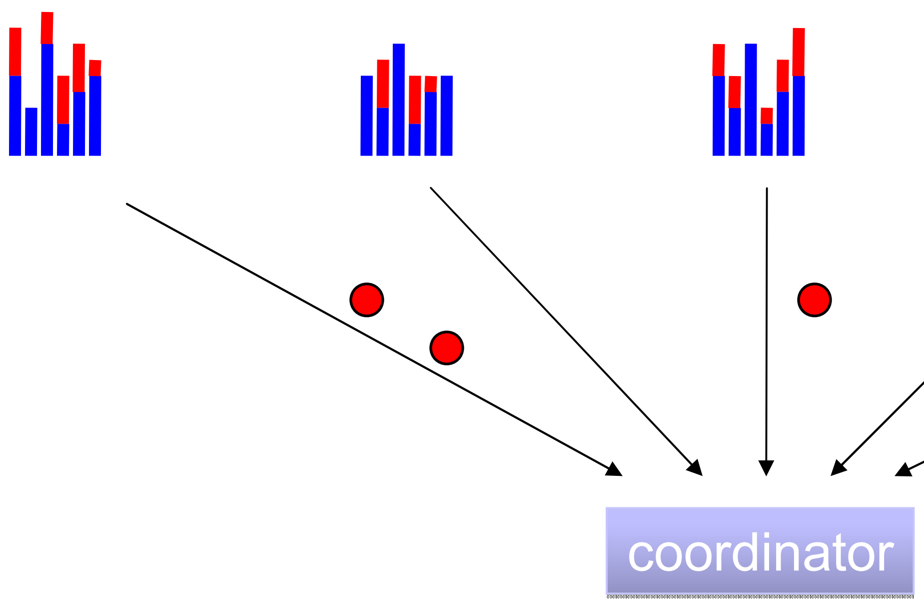


sends a signal whenever
F_2 of the updates increases
by $t_i = (\tau - \hat{u}_i^2)^2/(64k^2\tau)$

estimate for F_2

Analysis of F_2 Multi-Round Algorithm

End of a round: when k signals are received



rounds: $O(k/\varepsilon)$
Total cost: $\tilde{O}(k^2/\varepsilon^3)$

estimate for F_2

New bound on F_2 satisfies: $u_{i-1}^2 + (\tau - u_{i-1}^2) \cdot \varepsilon/k \ < \ u_i^2 < \ \tau$

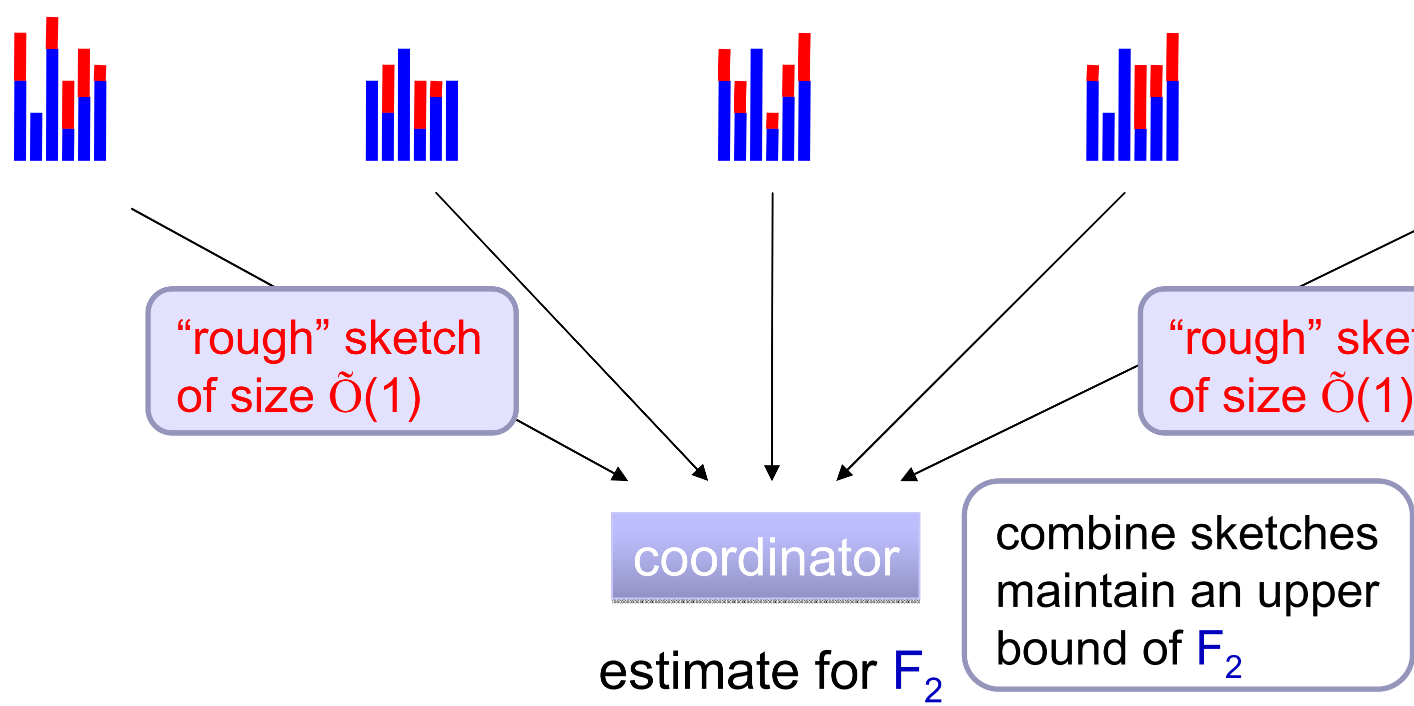
— Bound follows by using Cauchy-Shwartz inequality over the k update vectors

Modified F_2 algorithm

- Using Cauchy-Schwartz over the vectors means that we have large uncertainty in the current value (factor of k)
  - Collecting accurate sketches resolves this uncertainty, but at cost of $O(k/\varepsilon^2)$ communication
- Can improve cost by collecting less accurate sketches, and deciding whether to keep the same t_i or decrease it
  - Collect sketches with $O(1)$ accuracy in $O(k)$ communication
  - Resolves the uncertainty more cheaply
  - At most $O(\sqrt{k})$ "sub-rounds" within each round, and now at most $O(\sqrt{k}/\varepsilon)$ rounds

F_2 Round / Sub-Round Algorithm

End of a sub-round: when k signals are received



New bound on F_2 : $u_{i-1}^2 + (\tau - u_{i-1}^2) \cdot \overset{\varepsilon/\sqrt{k}}{\cancel{\varepsilon/k}} < u_i^2 < \tau$

$\boxed{\text{Total cost: } \tilde{O}(k^2/\varepsilon + k^{3/2}/\varepsilon^3)}$

$\boxed{\text{One-shot: } \tilde{O}(k/\varepsilon^2)}$

F_2 Lower Bound

- Via Minimax principle, demonstrate distribution on inputs that are hard for a deterministic algorithm (assuming compact oracle for F_2 computations)
- Proceed in rounds, in each round either send same item to all sites, or different items to each site
  - F_2 increases by either k or k^2
- If same item, $F_2 > \tau = k^2$
- Can send different items for up to $k/2$ rounds.
- All inputs look about the same to the sites, so a certain amount of communication is necessary each round
  - Implies $\Omega(k)$ bound on communication cost

Continuously Monitoring F_0

- Intuition: FM sketch for estimating F_0 is monotone
  - Site i calculates $zeros(h(x))$ for each x and maintains the maximum number Y_i of trailing zeros seen thus far.
  - Maintain $Y=\max_i Y_i$ at Coordinator so F_0 is estimated by 2^Y
  - Y_i is non-decreasing, and $Y_i < \log n$
  - Formal proof using variation of Bar-Yossef et al alg for F_0 Total communication: $\tilde{O}(k/\varepsilon^2)$

- Lower bound: $\Omega(k)$, by similar construction to F_2 bound
  - In each round updates are either all same $(\Delta F_0 = 1)$, or all different $(\Delta F_0=k)$

Summary of Results

Moment	Continuous Lower bound	Continuous Upper bound	One-shot Lower bound	One-shot Upper bound
F_0, randomized	$\Omega(k)$	$\tilde{O}(\frac{k}{\epsilon^2})$	$\Omega(k)$	$\tilde{O}(\frac{k}{\epsilon^2})$
F_1, deterministic	$\Omega(k \log \frac{1}{\epsilon k})$	$O(k \log \frac{1}{\epsilon})$	$\Omega(k \log \frac{1}{\epsilon k})$	$O(k \log \frac{1}{\epsilon})$
F_1, randomized	$\Omega(\min\{k, \frac{1}{\epsilon}\})$	$O(\min\{k \log \frac{1}{\epsilon}, \frac{1}{\epsilon^2} \log \frac{1}{\delta}\})$	$\Omega(k)$	$O(k \log \frac{1}{\epsilon \sqrt{k}})$
F_2, randomized	$\Omega(k)$	$\tilde{O}(k^2/\epsilon + (\sqrt{k}/\epsilon)^3)$	$\Omega(k)$	$\tilde{O}(\frac{k}{\epsilon^2})$

- Good news/Bad news: all continuous bounds (except F_2) are close to their one-shot counterparts
- Other problems have been studied
  - Quantiles/Heavy Hitters of a distribution
  - Tracking approximate clustering of a point set

Open Problems

- No clear separation between one-shot and continuous
  - F_2 has widest gap currently
- Many other functions f
  - Statistics: entropy, heavy hitters
  - Geometric measures: diameter, width, …
- Variations of the model
  - One-way vs two-way communication
  - Does having a broadcast channel help?
- Need for a “Continuous Communication complexity”?
  - Other formalizations: Alice must inform Bob of an (approx) value of $f(x)$. Analyze competitive ratio.