Optimal Gossip Algorithms for Exact and Approximate Quantile Computations

Hsin-Hao Su
(UNC Charlotte => Boston College)

joint work with Bernhard Haeupler (CMU) and Jeet Mohapatra (MIT)
WOLA ‘18
Shifting in Computing Paradigm

- Reduced costs on devices
- Increase in the amount of data
- Advances in connectivity between computers
Basic Aggregation Problems

- Basic Problems
  - Sum
  - Average
  - Min, Max
  - Quantiles: Median, 90%, 10%..
Basic Aggregation Problems

• Basic Problems
  – Sum
  – Average
  – Min, Max
  – Quantiles: Median, 90%, 10%..
Gossip Algorithms

• Gossip Algorithms / Epidemic Algorithms/ Population Protocols
  – Each node interacts with another node \( t(v) \), chosen uniformly at random
  – **PUSH** or **PULL** \( O(\log n) \) bits

– Nice properties:
  • Scalable
  • Low overhead (i.e. \( O(n) \) messages per round)
  • Fast convergence
  • Fault-tolerant

– Captures interaction patterns in Nature:
  • Molecules interactions in chemical reactions
  • Rumor spreading
Previous Results

- **Max, Min**
  - Folklore: $O(\log n)$ rounds w.h.p.

- **Sum, Average**
  - [Kempe, Dobra, Gehrke ’03]
    - PUSH-SUM: approximate within $(1 \pm \epsilon)$ in $O(\log n + \log(\frac{1}{\epsilon}))$ rounds w.h.p.
Previous Results

• Quantile Computation
  – $\phi$-quantile: Given $0 < \phi < 1$, every node outputs a value whose rank is $\lfloor \phi n \rfloor$
    • [Kempe, Dobra, Gehrke ’03]: $O(\log^2 n)$
  – $\epsilon$-approximate $\phi$-quantile: Given $0 < \phi, \epsilon < 1$, every node outputs a value whose rank is $(\phi \pm \epsilon)n$
    • [Doerr et al. ‘11]: $O(\log n)$ rounds alg. for $\epsilon = O\left(\sqrt{\frac{\log n}{n}}\right)$
      and $\phi = 0.5$ (median)
New results

- **φ-quantile:** Given $0 < \phi < 1$, every node outputs a value whose rank is $\lfloor \phi n \rfloor$

  $O(\log n)$ rounds  Optimal

- **ε-approximate φ-quantile:** Given $0 < \phi, \epsilon < 1$, every node outputs a value whose rank is $(\phi \pm \epsilon)n$

  $O \left( \log \log n + \log \left( \frac{1}{\epsilon} \right) \right)$ rounds  Optimal
New results

- **$\phi$-quantile**: Given $0 < \phi < 1$, every node outputs a value whose rank is $\lfloor \phi n \rfloor$

  \[
  O(\log n) \text{ rounds} \quad \text{Optimal}
  \]

- **$\epsilon$-approximate $\phi$-quantile**: Given $0 < \phi, \epsilon < 1$, every node outputs a value whose rank is $(\phi \pm \epsilon) n$

  \[
  O\left(\log \log n + \log \left(\frac{1}{\epsilon}\right)\right) \text{ rounds} \quad \text{Optimal}
  \]
Suppose each node randomly samples $\Theta \left( \frac{\log n}{\epsilon^2} \right)$ values and outputs the $\phi$-quantile of the sampled values.

Then, w.h.p. the quantile of the value is $(\phi \pm \epsilon)$.

\{39, 50, 77, 37, 43\}

$O \left( \frac{\log n}{\epsilon^2} \right)$ rounds

$O \left( \log \log n + \log \left( \frac{1}{\epsilon} \right) \right)$ rounds?
Second Attempt (Doubling)

The sampled values can be doubled in every round.

Doubling

\[ O \left( \log \log n + \log \left( \frac{1}{\epsilon} \right) \right) \text{ rounds} \]

Message size \( \Theta \left( \frac{\log n}{\epsilon^2} \cdot \log n \right) \) bits
Third Attempt (Quantile Sketch)

Instead of storing the whole set of sampled values, only keep a sketch of it.

Quantile Sketch: [Munro and Patterson ‘80, Manku et al. ‘99, Greenwald and Khanna ‘01]

\[
\{25, 35, 39, 41, 41, 50, 77, 103\} \rightarrow \{25, 35, 39, 41, 41, 50, 77, 103\}
\]

Message size:
\[
\Theta((\log \log n + (\log 1/\epsilon)) \cdot \log n) \text{ bits}
\]

\(O(\log n)\) bits possible?
Our Approach

• Phase I: Shift the $\phi$-quantile to the approximate median

• Phase II: Compute the approximate median
The TOURNAMENT Algorithm

3-Tournament
For each iteration:

Each node \( v \) randomly samples 3 values and sets itself to the middle one

After \( O \left( \log \log n + \log \left( \frac{1}{\epsilon} \right) \right) \) iterations, the rank of every value is in \( (0.5 \pm \epsilon)n \)
Running Time

Nodes whose values are in \((0.5 \pm \epsilon)\) quantile

Nodes whose values are not in \((0.5 \pm \epsilon)\) quantile

\[
\begin{align*}
H_{i+1} &= 3H_i^2 - 2H_i^3 \\
L_{i+1} &= 3L_i^2 - 2L_i^3
\end{align*}
\]

Expected Behavior

\[
O\left(\log\left(\frac{1}{\epsilon}\right)\right)
\]

\[
O(\log \log n)
\]
Approximate Median

• Compute the approximate median ($\phi = 0.5$)

\[ O \left( \log \log n + \log \left( \frac{1}{\epsilon} \right) \right) \text{ rounds} \]

• $\phi$-quantiles for other $\phi$?
φ-quantiles for other φ?

- Phase I: Shift the φ-quantile to the approximate median

- Phase II: Compute the approximate median

\[ O \left( \log \log n + \log \left( \frac{1}{\epsilon} \right) \right) \text{ rounds} \]
Phase I: Shifting

2-Tournament ($\phi > 1/2$)
For each iteration:
Each node $v$ randomly samples 2 values and sets itself to the higher one

After at most $O\left(\log\left(\frac{1}{\epsilon}\right)\right)$ rounds, the $(\phi \pm \epsilon)$-quantiles become the current approximate median.
Running Time

Expected Behavior

- $L_{i+1} = L_i^2$
- $M_{i+1} \geq M_i$

$$O \left( \log \left( \frac{1}{\epsilon} \right) \right)$$
Our Approach

• Phase I: Shift the $\phi$-quantile to approximate median

\[ O \left( \log \left( \frac{1}{\epsilon} \right) \right) \text{ rounds} \]

• Phase II: Compute the approximate median

\[ O \left( \log \log n + \log \left( \frac{1}{\epsilon} \right) \right) \text{ rounds} \]
Caveat

• Given $0 \leq \phi \leq 1$ and $0 < \epsilon < 1$, every node outputs a value whose rank is $(\phi \pm \epsilon)n$

\[
O \left( \log \log n + \log \left( \frac{1}{\epsilon} \right) \right) \text{ rounds w.h.p., but only for } \epsilon \geq \frac{1}{n^{0.01}}
\]

• Example: $\phi = 0.5$ (median), $\epsilon = 1/(2n)$ (exact quantile computation)
  – After the first round of the 3-Tournament algorithm, with a constant probability, the answer is erased.
Quantile Computation for Small $\epsilon$

- **$\phi$-quantile**: Given $0 < \phi < 1$, every node outputs a value whose rank is $\lfloor \phi n \rfloor$
  
  $O(\log n)$ rounds

- **$\epsilon$-approximate $\phi$-quantile**: Given $0 < \phi < 1$ and $1/n^{0.01} < \epsilon < 1$, every node outputs a value whose rank is $(\phi \pm \epsilon)n$
  
  $O\left(\log \log n + \log \left(\frac{1}{\epsilon}\right)\right)$ rounds
Bootstrapping

• Given $0 \leq \phi \leq 1$, every node outputs a value whose rank is $\lceil \phi n \rceil$
• Bootstrap the approximation algorithm:
  1. Run the algorithm for $\epsilon = 1/n^{0.01}$
  2. Discard values lying outside $\phi \pm \epsilon$ quantiles
  3. Duplicate the remaining values
  4. REPEAT for $O(1)$ rounds
Quantile Computation

• Given $0 \leq \phi \leq 1$ and $0 < \epsilon < 1$, every node outputs a value whose rank is $(\phi \pm \epsilon)n$

$O\left(\log \log n + \log\left(\frac{1}{\epsilon}\right)\right)$ rounds, but only for $\epsilon \geq \frac{1}{n^{0.01}}$
Robustness

- Our algorithm also tolerates constant probability node failure in each round with the same asymptotic running time.
Theorem: Given \( \frac{\log n}{n} < \epsilon < 1 \), any gossip algorithms that uses \( o\left(\log \log n + \log \left(\frac{1}{\epsilon}\right)\right) \) rounds fail to \( \epsilon \)-approximate the median with prob. at least \( \frac{1}{3} \).
Lower Bound

• Indistinguishable Arguments
  – Scenario 1: the values are \{1, 2, \ldots, n\}
  – Scenario 2: the values are \{1 + \epsilon n, 2 + \epsilon n, \ldots, n + \epsilon n\}
Lower Bound

- Indistinguishable Arguments
  - The median in the first scenario and the second differ by $\epsilon n$
  - In Scenario 1, each node must receive a value in $S = \{1, 2, \ldots, \epsilon n\}$ to ensure it is not in Scenario 2
Lower Bound

• Lower Bound
  – $|S| = \epsilon n$
  – It takes $\Omega \left( \log \log n + \log \left( \frac{1}{\epsilon} \right) \right)$ to spread messages from S to every node using both PUSH and PULL.
Open Problems

• **For all** $\epsilon$-quantile computation problem
  – Each node $v$ outputs $Q(v)$ such that $|Q(v) - \phi(v)| \leq \epsilon$, where $\phi(v)$ is the quantile of $v$
  
  – **Approach 1:**
    
    • Use our algorithm to select the $\frac{\epsilon}{2}, \frac{2\epsilon}{2}, \frac{3\epsilon}{2}, ...$ quantiles
    • Each node outputs the nearest quantile
    
    • $O\left(\frac{1}{\epsilon} \left(\log \log n + \log \left(\frac{1}{\epsilon}\right)\right)\right)$ rounds

  – **Approach 2:**
    
    • Broadcast every value to every node in $O(n + \log n)$ rounds using Network Coding [Haeupler ‘11]
    
    • $O(n)$ rounds for exact answers
Open Problems

- Pipeline problems
  - [Haeupler ‘11] Broadcast $k$ messages in $O(k + \log n)$ rounds by gossiping.
  - [Kempe, Dobra, Gehrke ’03] Compute the sum in $O(\log n)$ rounds
  - Compute $k$ sums in $O(k + \log n)$ rounds?

Thank you