On Fine-Grained Distinct Element Estimation

Ilias Diakonikolas (UW Madison)
Daniel M. Kane (UC San Diego)
Jasper C.H. Lee (UC Davis)
Thanasis Pittas (UW Wisconsin)
David P. Woodruff (Carnegie Mellon University)
Samson Zhou (Texas A&M University)

ICML International Conference On Machine Learning 2025

Distinct Elements

- α servers, server i has a set $S_i \subseteq [1,2,3,\ldots,n]$
- $S = \cup_i S_i$
- Number of distinct elements in dataset is $F_0(S) = \|S\|_0 = |S|$

Communication Model

- Each server can talk to any other server, but only on private channel
- Can designate a specific server as the *coordinator*
- Communication over channel is measured in bits
- Goal: Using minimum total communication, output $(1+\varepsilon)$-multiplicative approximation to $F_0(S)$, i.e., output some number Z such that $Z \le F_0(S) \le (1+\varepsilon)\cdot Z$

Previous Results

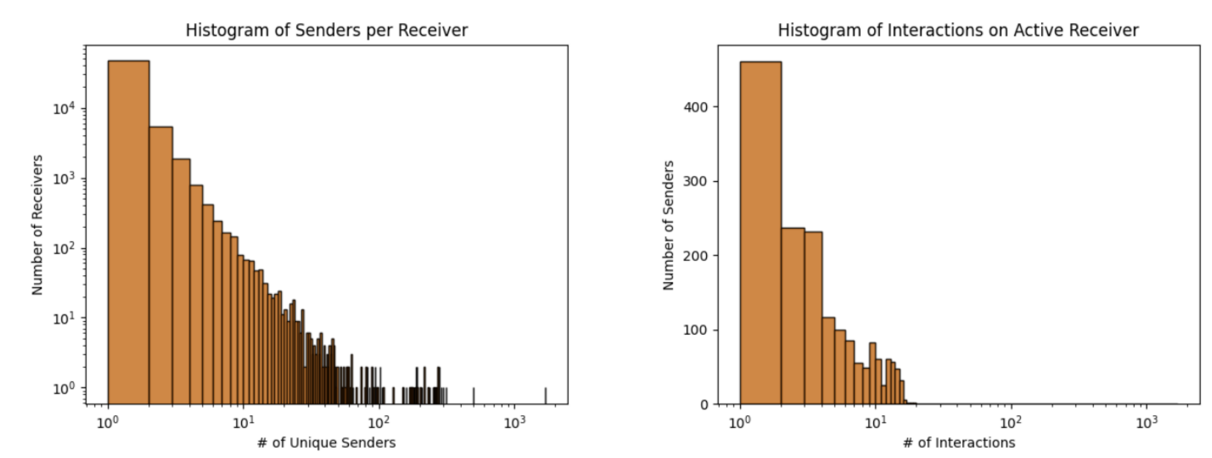
- Theorem [KNW10, Bla20]: There exists a distributed protocol that outputs a $(1+\varepsilon)$-multiplicative approximation to $F_0(S)$, using $O\left(\frac{\alpha}{\varepsilon^2} + \alpha \log n\right)$ bits of communication
- Theorem [WZ14]: Any distributed protocol that outputs a $(1+\varepsilon)$-multiplicative approximation to $F_0(S)$ requires $\Omega\left(\frac{\alpha}{\varepsilon^2} + \alpha \log n\right)$ bits of communication

Bridging Theory and Practice

- There is usually a large number of servers and we want good accuracy, so this means we should require *a lot* of communication!
- Algorithms behave well in practice with little communication
- What's going on?

Our Observations

- In the lower bound instance, all servers have roughly the same number of items in the sets S_i
- In practice, the datasets are "skewed" so a small number of servers have a large number of items, e.g., Zipf's Law, 80-20 Rule, or Pareto's Principle
- This does change the complexity of the problem? If so, how to characterize the complexity?



(a) Histogram of Receivers (b) Histogram of Activity

New Parameterization: Pairwise Collisions

- We define C to be the number of pairwise collisions, i.e., number of triplets (a,i,j) so that $a \in S_i$ and $a \in S_j$ but $i < j$
- Note that if a single server has most of the items, they will not be repeated across the other servers, so C is low and can be constant
- When all servers have similar number of items and there are many intersections, C can be as large as $\alpha^2 \cdot F_0(S)$

	$C = \beta \cdot F_0(S),\ \beta \ge 1$	
	$F_0(S) < \frac{1}{\varepsilon^2}$	$F_0(S) \ge \frac{1}{\varepsilon^2}$
Theorem 1.1	$\mathcal{O}\left(\alpha \log n + \sqrt{\beta}\cdot F_0(S)\cdot \log n\right)$	$\mathcal{O}\left(\alpha \log n + \frac{\sqrt{\beta}}{\varepsilon^2}\log n\right)$
Theorem 1.3	$\Omega(\alpha + \sqrt{\beta}\cdot F_0(S))$	$\Omega\left(\alpha + \frac{\sqrt{\beta}}{\varepsilon^2}\right)$
	$C = \beta \cdot F_0(S),\ \beta < 1,\ C > \varepsilon \cdot F_0(S)$	
	$F_0(S) < \frac{1}{\varepsilon^2}$	$F_0(S) \ge \frac{1}{\varepsilon^2}$
Theorem 1.2	$\mathcal{O}\left(\alpha \log n + \frac{\beta}{\varepsilon^2}\log n\right)$	$\mathcal{O}\left(\alpha \log n + \beta\cdot F_0(S)\cdot \log n\right)$
Theorem 1.4	$\Omega\left(\alpha + \frac{\beta}{\varepsilon^2}\right)$	$\Omega(\alpha + \beta \cdot F_0(S))$

Table 1: A summary of our results for the distributed distinct elements estimation problem on a universe of size n across α servers, parameterized by the number C of collisions across the α servers, and the accuracy parameter $\varepsilon \in (0,1)$.

Results

- Theorem: Suppose $C = \beta \cdot O\left(\min\left(F_0(S), \frac{1}{\varepsilon^2}\right)\right)$. There exists a distributed protocol that outputs a $(1+\varepsilon)$-multiplicative approximation to $F_0(S)$, using $O\left(\min\left(F_0(S), \frac{1}{\varepsilon^2}\right)\cdot \sqrt{\beta}\log n + \alpha \log n\right)$ bits of communication
- Point of comparison: Previously, first term was $\frac{\alpha}{\varepsilon^2}$, which happens if $C = \alpha^2 \cdot F_0(S)$, i.e., many items appear across many servers
- Theorem: Suppose $C = \beta \cdot O\left(\min\left(F_0(S), \frac{1}{\varepsilon^2}\right)\right)$. Any distributed protocol that outputs a $(1+\varepsilon)$-multiplicative approximation to $F_0(S)$ requires $\Omega\left(\min\left(F_0(S), \frac{1}{\varepsilon^2}\right)\cdot \sqrt{\beta} + \alpha \log n\right)$ bits of communication
- Point of comparison: Tight up to a $\log n$ factor in the first term
- Context: Shows that C is a parameter that characterizes the complexity of the problem

Algorithm 2 $(1+\varepsilon)$-approximation to F_0, given an upper bound on the number of collisions

Input: Items given to α players from a universe of size $[n]$, accuracy parameter $\varepsilon \in (0,1)$, upper bound C on the number of pair-wise collisions
Output: $(1+\varepsilon)$-approximation to the number of distinct items

```
Let X be a 4-approximation to F_0                                   ▷Lemma 2.4
Let i_0 be the largest integer such that X / 2^{i_0} > 1000 / ε^2
i ← min(0, i_0)
Let S_i be a subset of [n] where each item is subsampled with probability 1/2^i
Assume without loss of generality each player i has a binary vector v^{(i)} ∈ {0,1}^n
Each player sends their total number of items in S_i
Let Z be the sum of these numbers
η ← ε/10, p ← min(1, 100C / (η^2 X^2))
Let T be a subset of S_i where each item is subsampled with probability p
Each player sends their items in T
Let W = Σ_{j∈T} max(0, v_j − 1), where v = Σ_{i∈[α]} v^{(i)} be the excess mass in T
Return Z · 2^i − W · 1/p
```