Beyond Set Disjointness: The Communication Complexity of Finding the Intersection

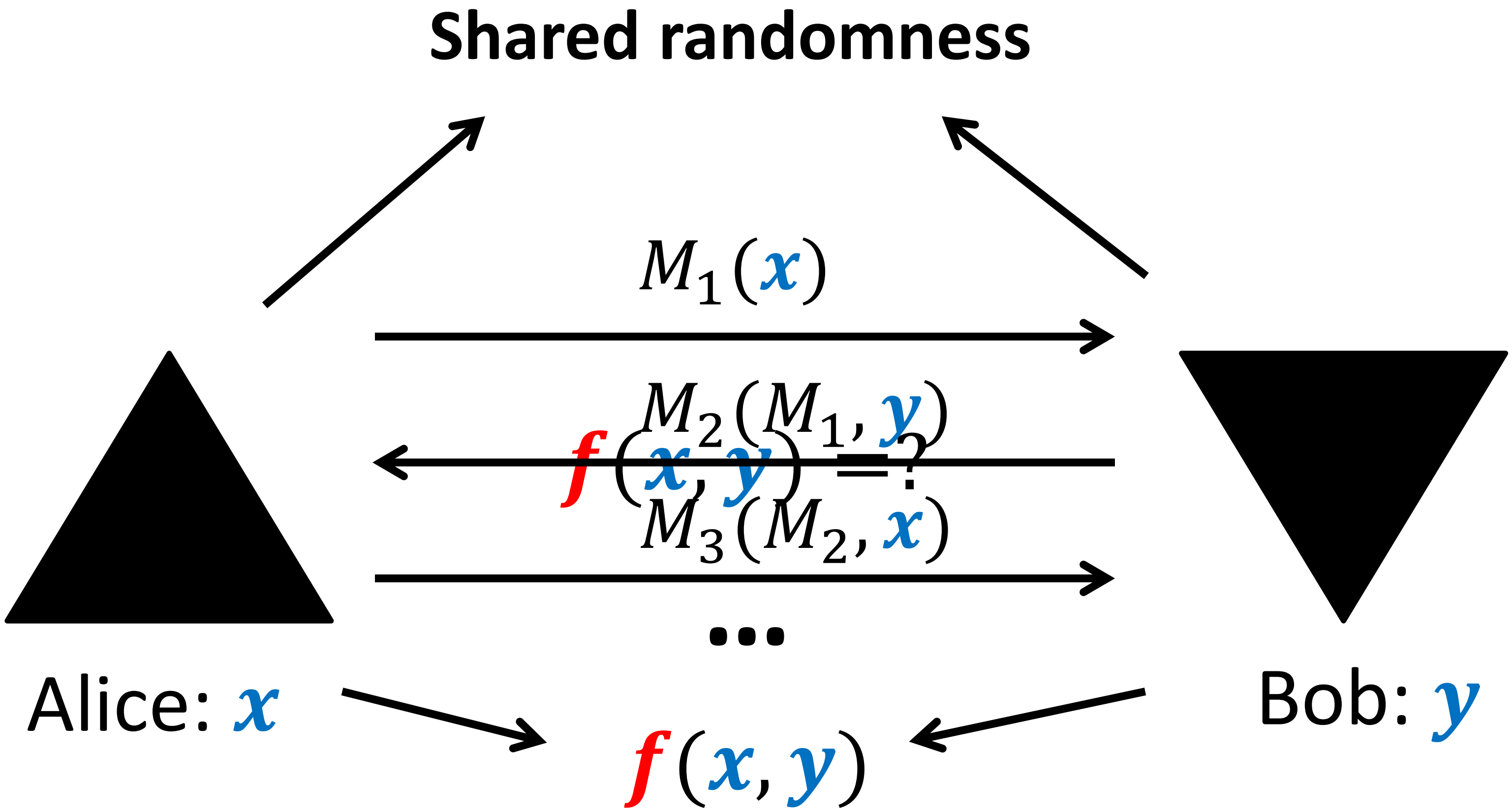
Grigory Yaroslavtsev

http://grigory.us

Penn
University of Pennsylvania

Joint with Brody, Chakrabarti, Kondapally and Woodruff

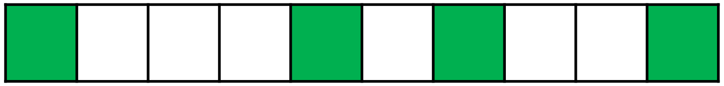
Communication Complexity [Yao'79]



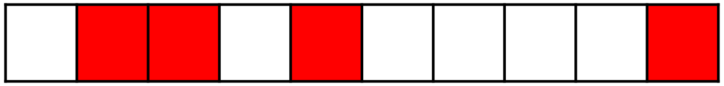
- $R(f)$ = min. communication (error 1/3)
- $R^k(f) =$ min. k-round communication (error 1/3)

Set Intersection

- $\boldsymbol{x} = \boldsymbol{S}, \boldsymbol{y} = \boldsymbol{T}, \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{S} \cap \boldsymbol{T}$

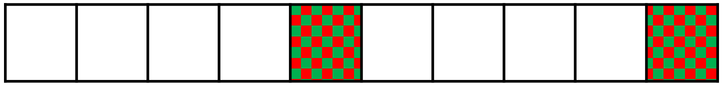


$\boldsymbol{S} \subseteq [n], |S| \leq \boldsymbol{k}$



$\boldsymbol{T} \subseteq [n], |T| \leq \boldsymbol{k}$

$\boldsymbol{S} \cap \boldsymbol{T}$ = ?



$\boldsymbol{R}^{\boldsymbol{r}}$($\boldsymbol{k}$-Intersection) = ?

$\boldsymbol{k}$ is big, n is **huge**, where **huge** $\gg$ big

Our results

Let $i\log^r k = \underbrace{\log\log\ldots\log}_{r\text{ times}} k$

- $\boldsymbol{R}^{\boldsymbol{r}}(\boldsymbol{k}$-Intersection) = $O\left(\boldsymbol{k}\ i\log^{\boldsymbol{\beta r}} \boldsymbol{k}\right)$

[Brody, Chakrabarti, Kondapally, Woodruff, Y.; PODC'14]

- $\boldsymbol{R}^{\boldsymbol{r}}(\boldsymbol{k}$-Intersection) = $\Omega(\boldsymbol{k}\ i\log^{\boldsymbol{r}} \boldsymbol{k})$

[Saglam-Tardos FOCS'13; Brody, Chakrabarti, Kondapally, Woodruff, Y.'; RANDOM'14]

$\boldsymbol{R}^{\boldsymbol{r}}(\boldsymbol{k}$-Intersection) = $\Theta(\boldsymbol{k})$ for $r = O(\log^* \boldsymbol{k})$

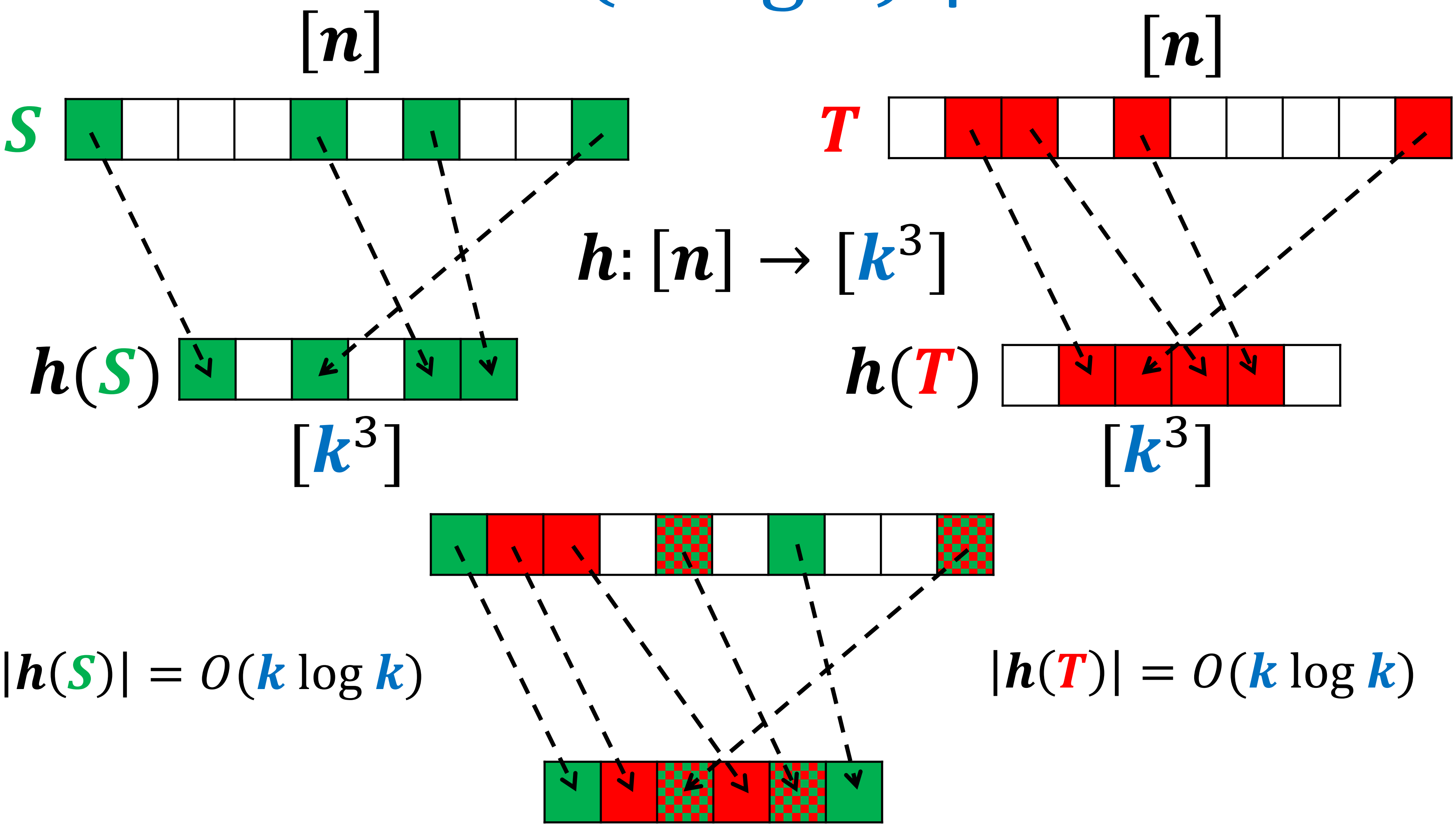
Applications

- **Exact** Jaccard index $J(\boldsymbol{S}, \boldsymbol{T}) = \frac{|\boldsymbol{S} \cap \boldsymbol{T}|}{|\boldsymbol{S} \cup \boldsymbol{T}|}$

(for $(1 \pm \epsilon)$-approximate use MinHash [Broder'98; Li-Konig'11; Path-Strokel-Woodruff'14])

- Rarity, distinct elements, joins,…
- Multi-party set intersection (later)
- Contrast: $\boldsymbol{R}(\boldsymbol{S} \cup \boldsymbol{T}) = \boldsymbol{R}(\boldsymbol{S}\, \boldsymbol{\Delta}\, \boldsymbol{T}) = \Theta\left(\boldsymbol{k} \log \frac{\boldsymbol{n}}{\boldsymbol{k}}\right)$

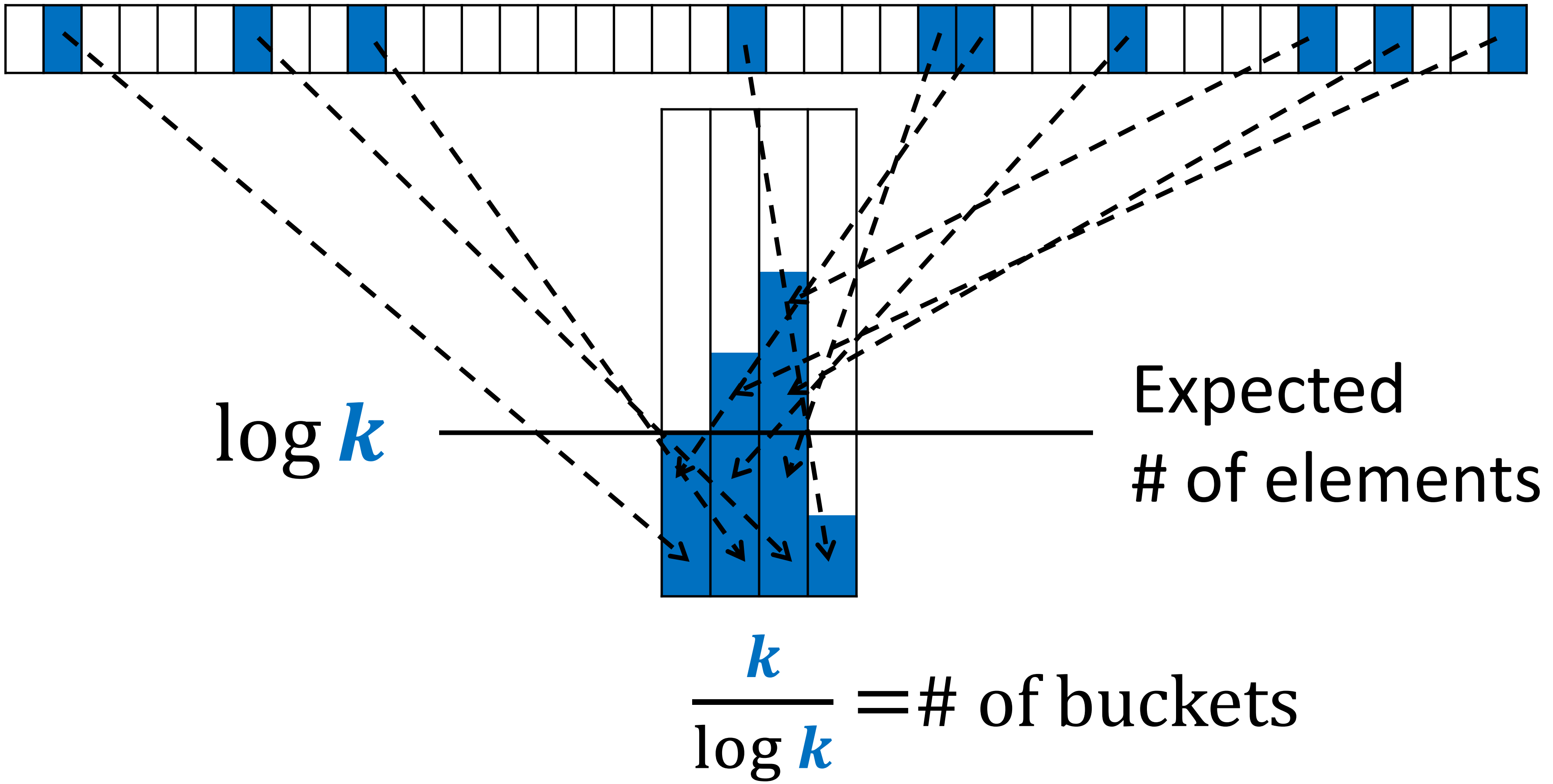
1-round $O(k \log k)$-protocol



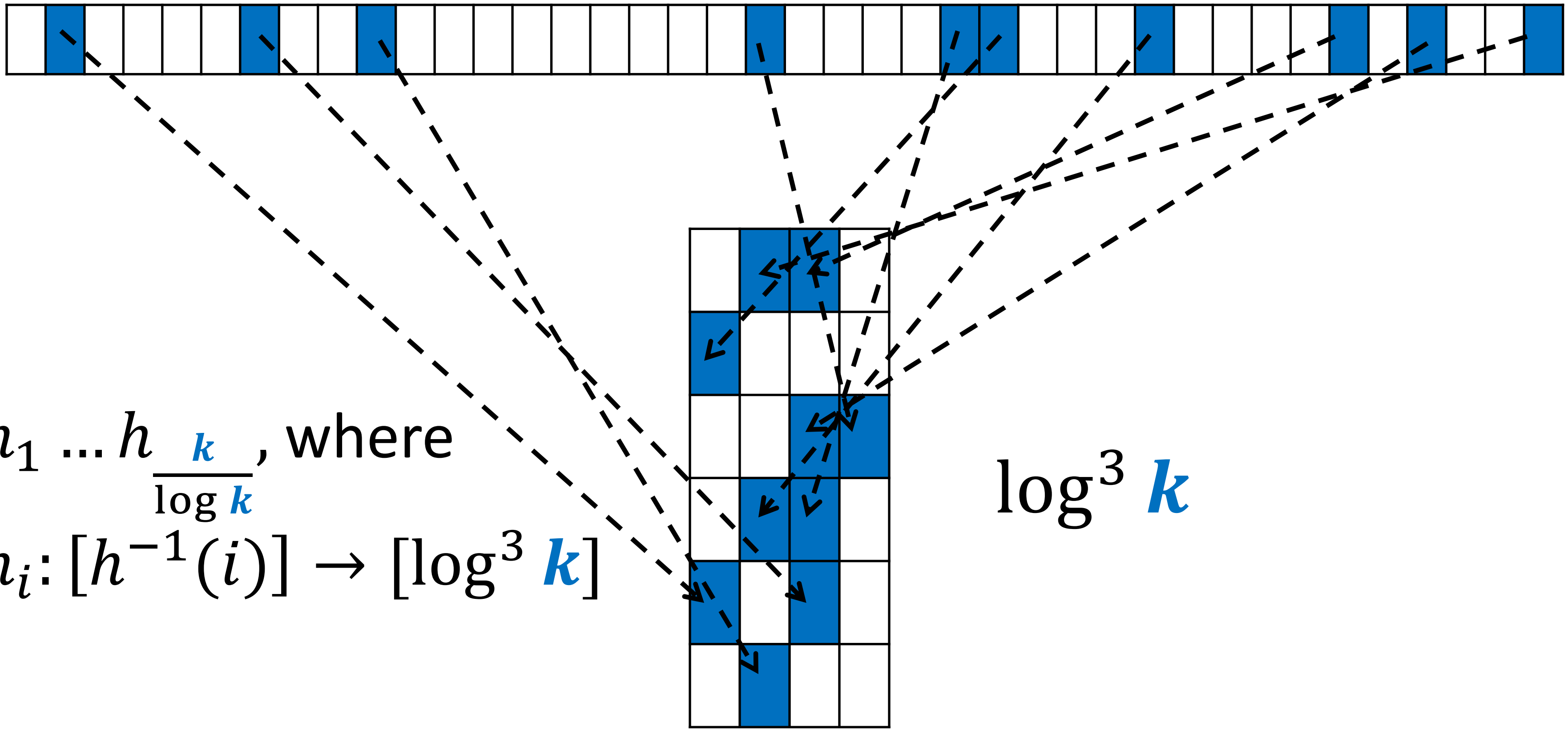
$$S \cap T = S \cap h^{-1}\big(h(T)\big) = h^{-1}\big(h(S)\big) \cap T$$

Hashing

$$h: [n] \to [k / \log k]$$



Secondary Hashing



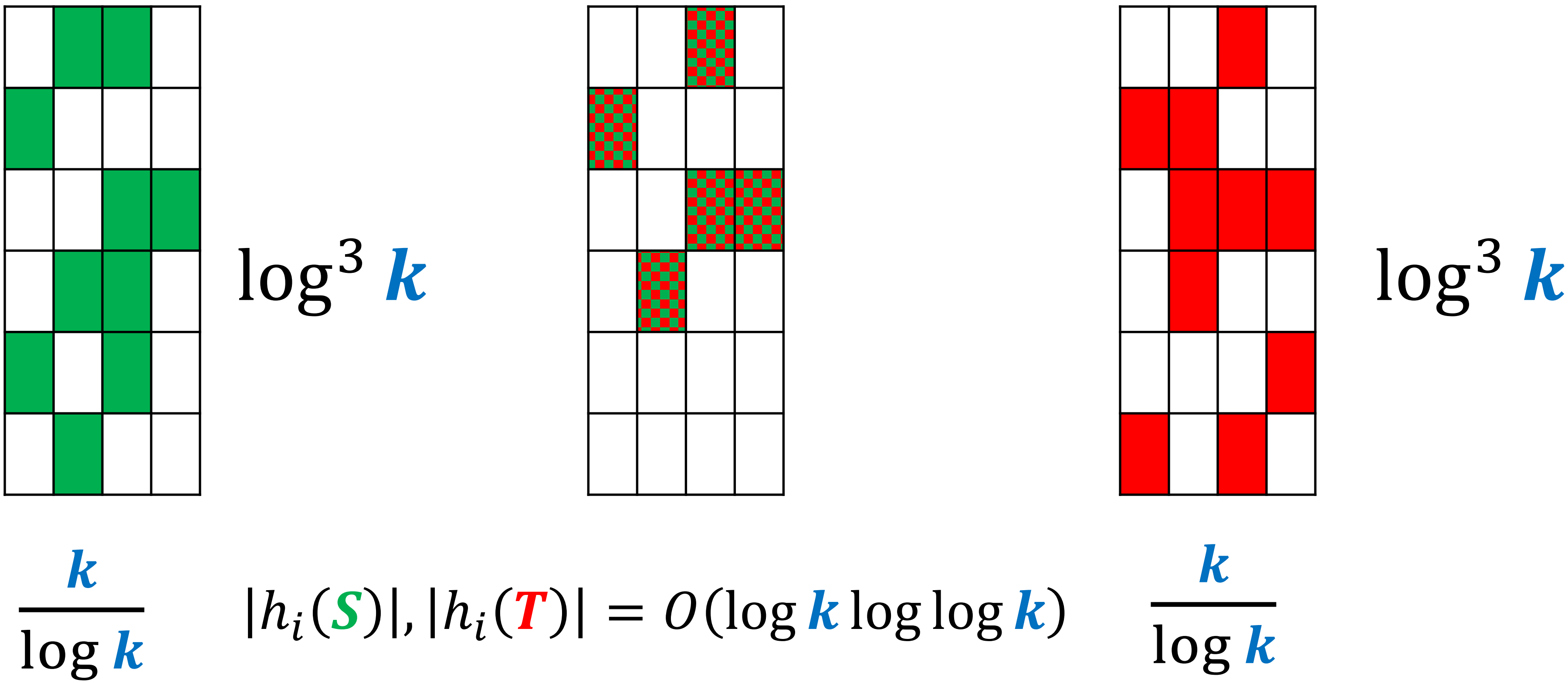
$h_1 \ldots h_{\frac{k}{\log k}}$, where

$h_i : [h^{-1}(i)] \rightarrow [\log^3 k]$

$\log^3 k$

$\frac{k}{\log k}$ = # of hash functions

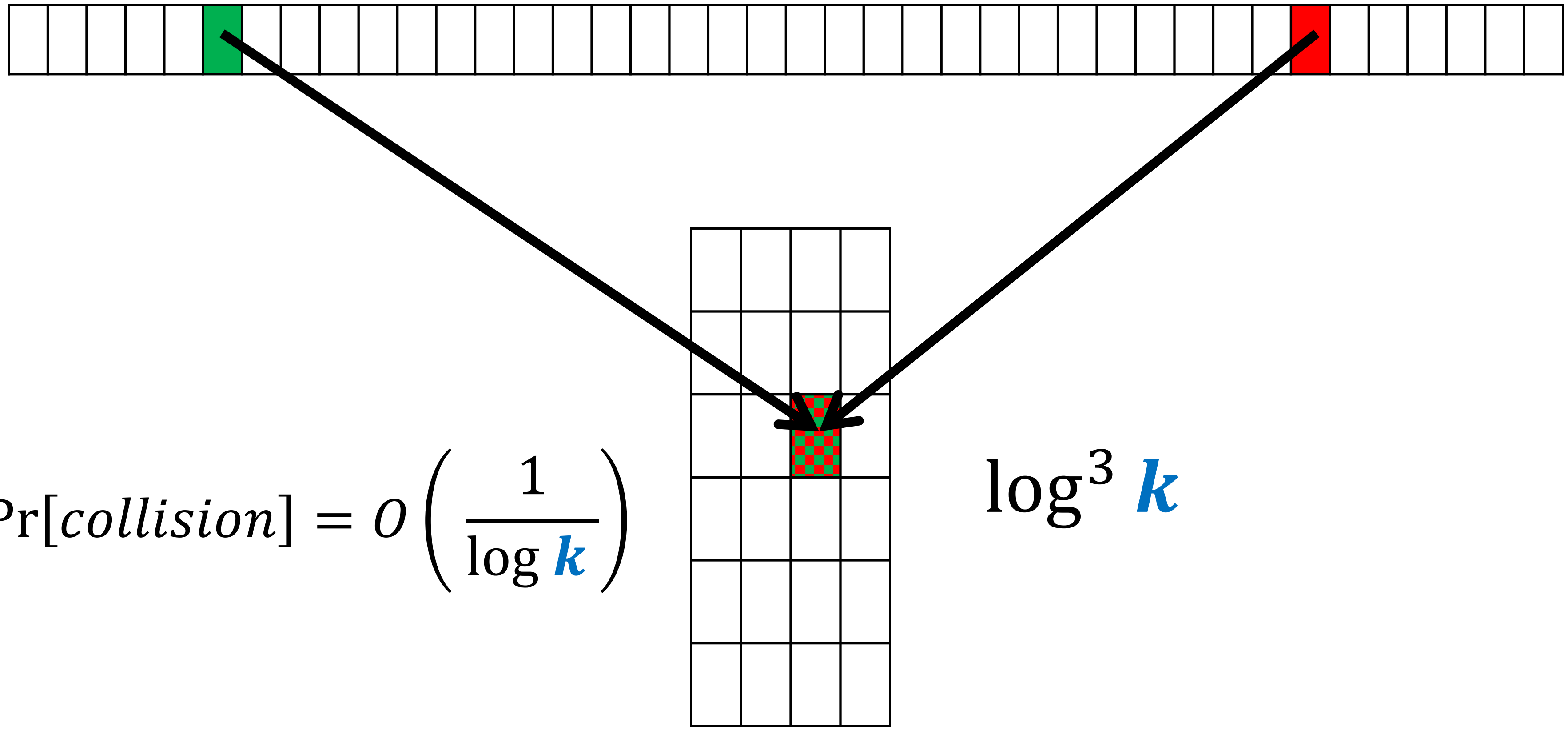
2-Round $O(k \log\log k)$-protocol



$|h_i(S)|, |h_i(T)| = O(\log k \log\log k)$

Total communication = $\frac{k}{\log k} O(\log k \log\log k)$ = $O(k \log\log k)$

Collisions

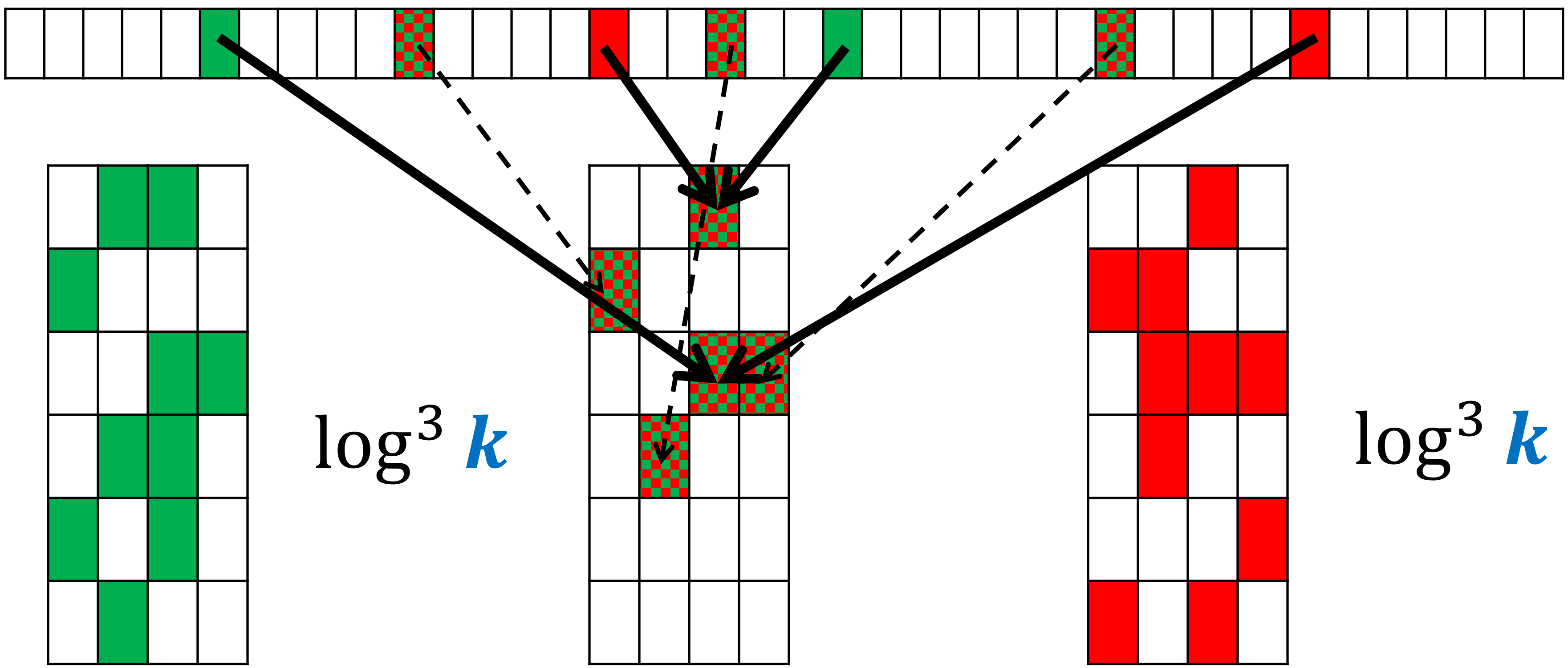


$$\Pr[collision] = O\left(\frac{1}{\log k}\right)$$

$\log^3 k$

$$\frac{k}{\log k}$$

Collisions



$\frac{k}{\log k}$ $\frac{k}{\log k}$

$S \cap T \subseteq S \cap H^{-1}(\ )$

$S \cap T \subseteq T \cap H^{-1}(\ )$

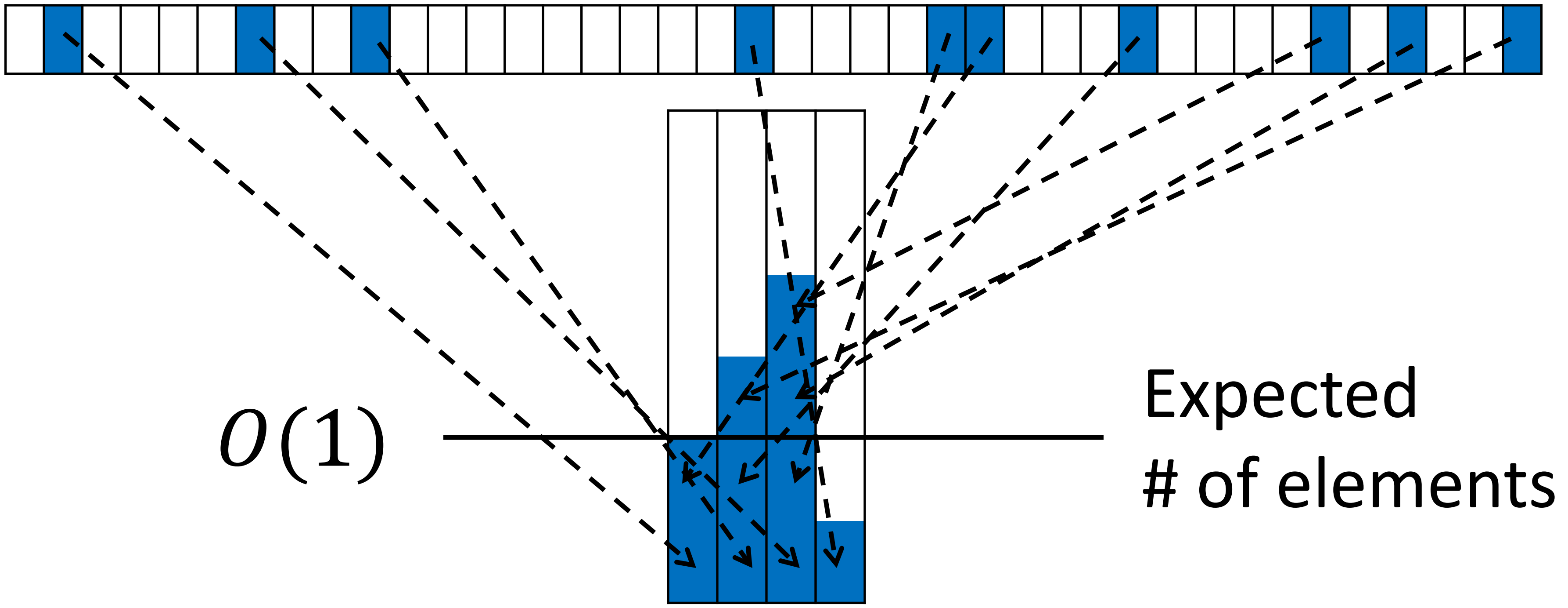
Key fact: If $S \cap H^{-1}(\ ) = T \cap H^{-1}(\ )$ then also = $S \cap T$

Collisions

- Second round:
  - For each bucket send $O(\log \boldsymbol{k})$-bit equality check (total $O(\boldsymbol{k})$-communication)
  - Correct intersection computed in buckets i where
    $$\boldsymbol{S} \cap \boldsymbol{H}_{\boldsymbol{i}}^{-\mathbf{1}}(\blacksquare) = \boldsymbol{T} \cap \boldsymbol{H}_{\boldsymbol{i}}^{-\mathbf{1}}(\blacksquare)$$
  - Expected # items in incorrect buckets $O(\boldsymbol{k} \,/ \log \boldsymbol{k})$
  - Use 1-round protocol for incorrect buckets
  - Total communication $O(\boldsymbol{k} \log \log \boldsymbol{k})$

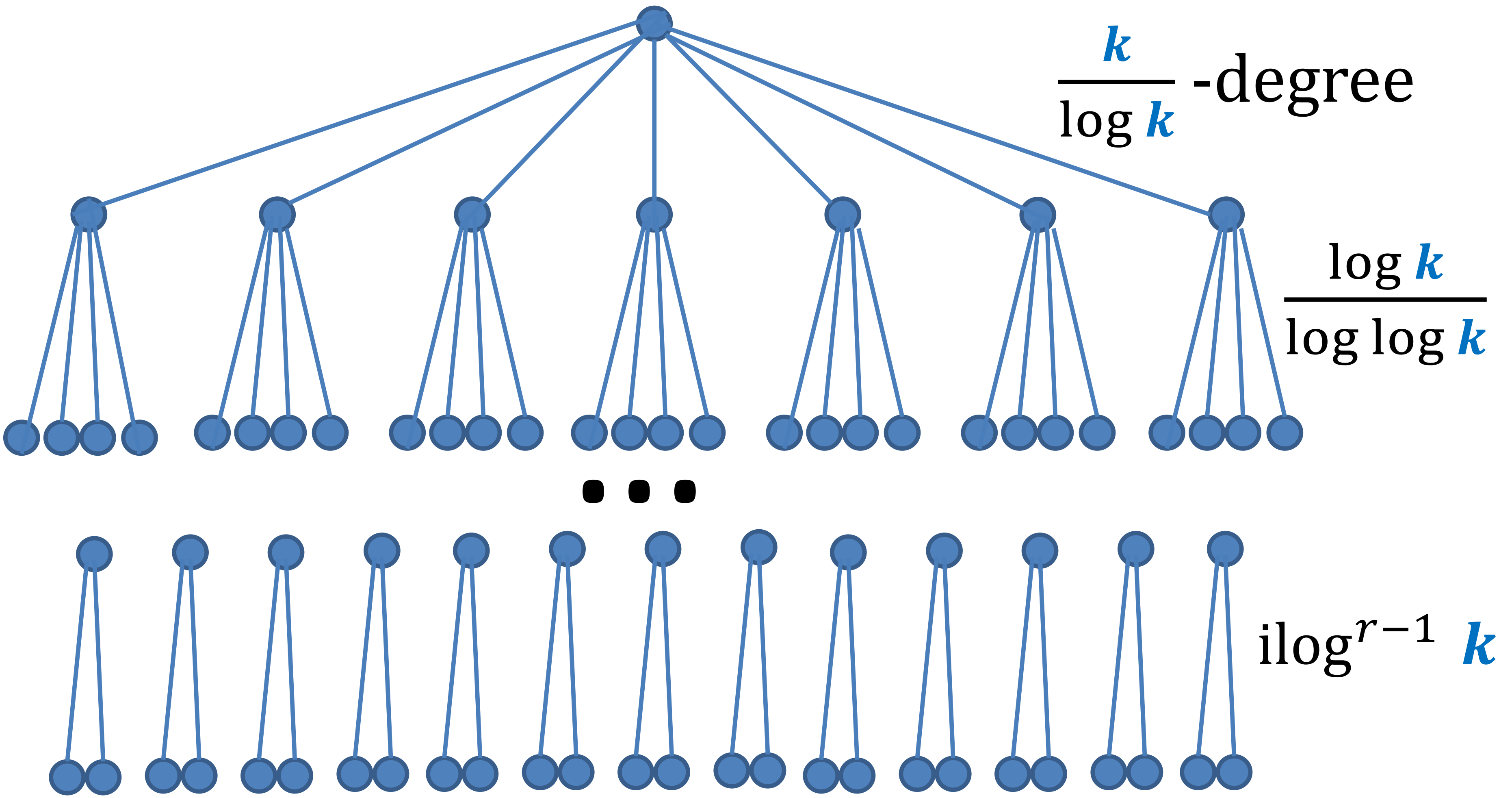
Main protocol

$h: [n] \rightarrow [k]$



k =# of buckets

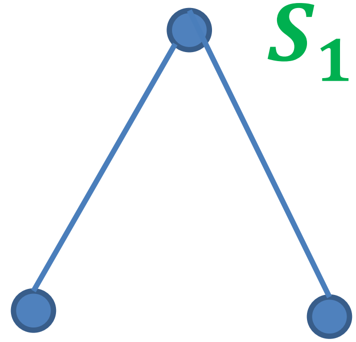
Verification tree



k buckets = leaves of the verification tree

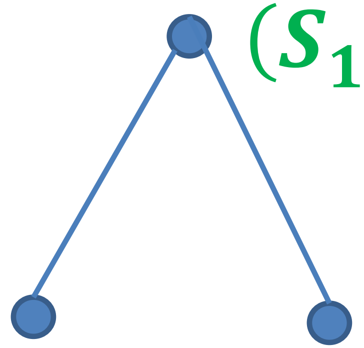
Verification bottom-up

$S_1 \cup S_2, T_1 \cup T_2$



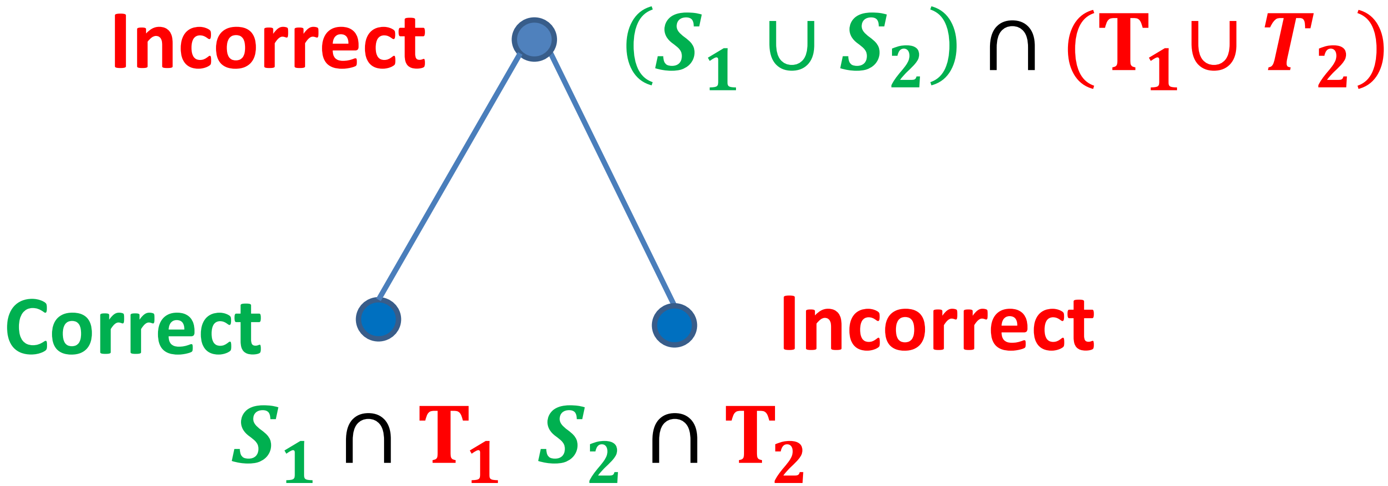
S_1, T_1 S_2, T_2

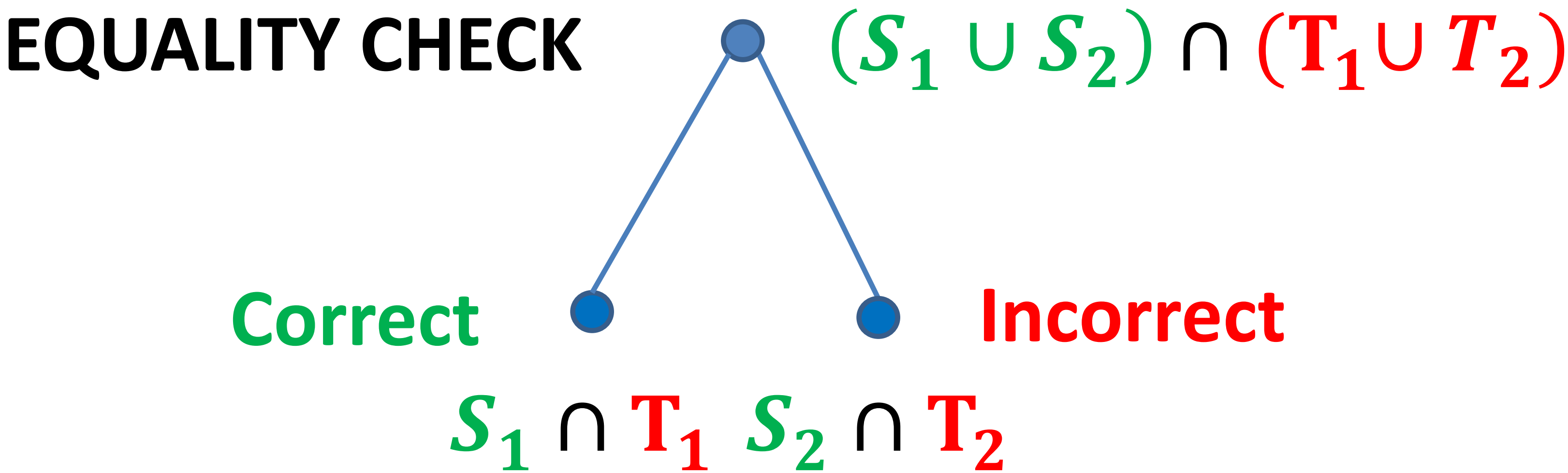
$(S_1 \cup S_2) \cap (T_1 \cup T_2)$



$S_1 \cap T_1$ $S_2 \cap T_2$

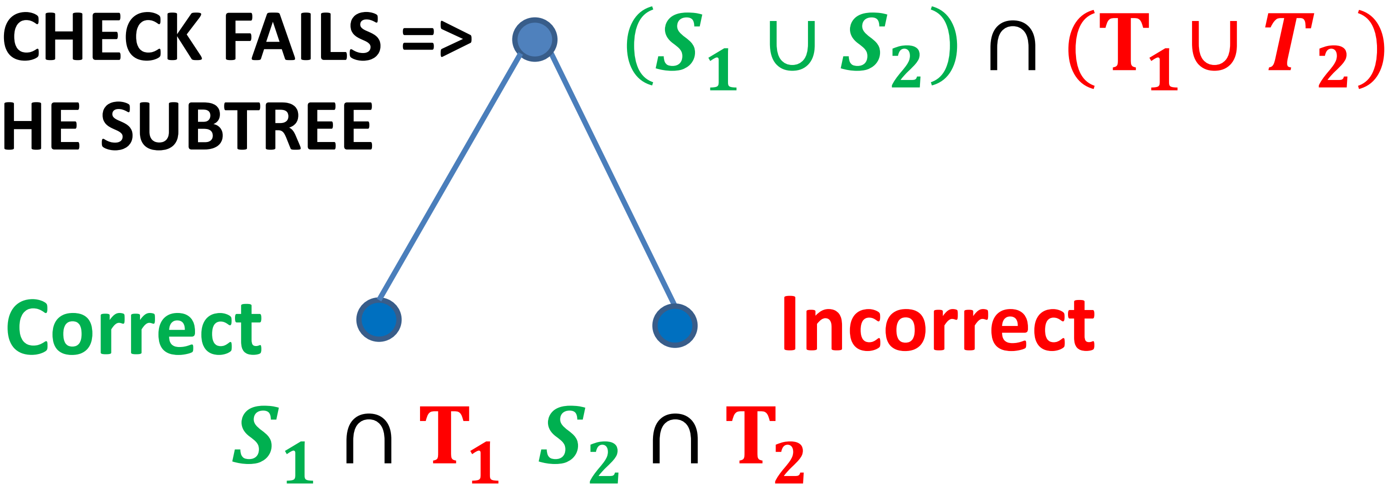
Verification bottom-up

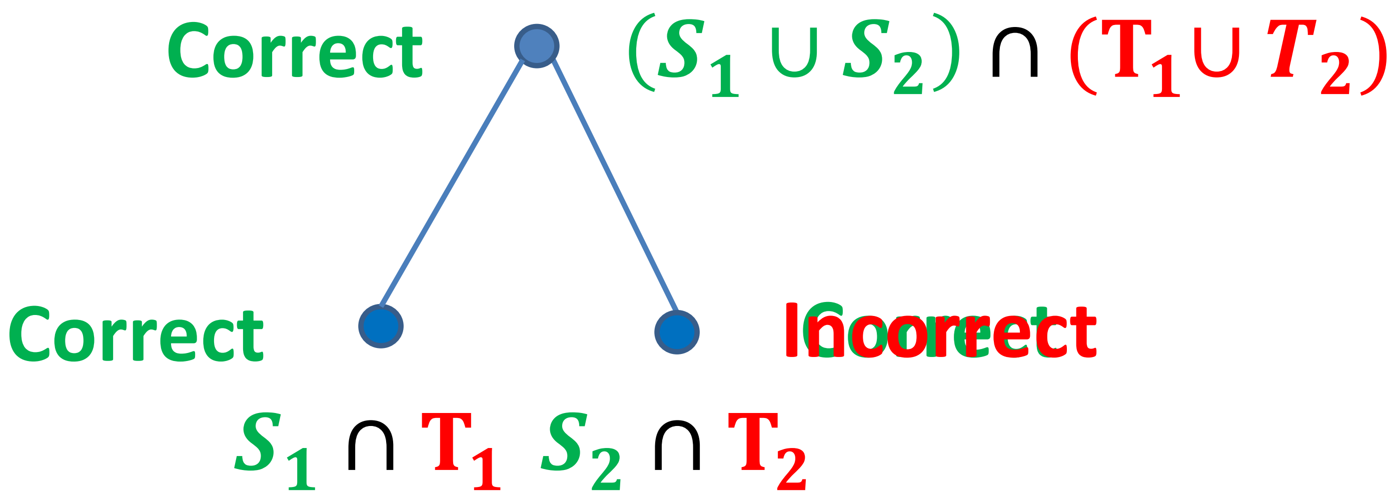




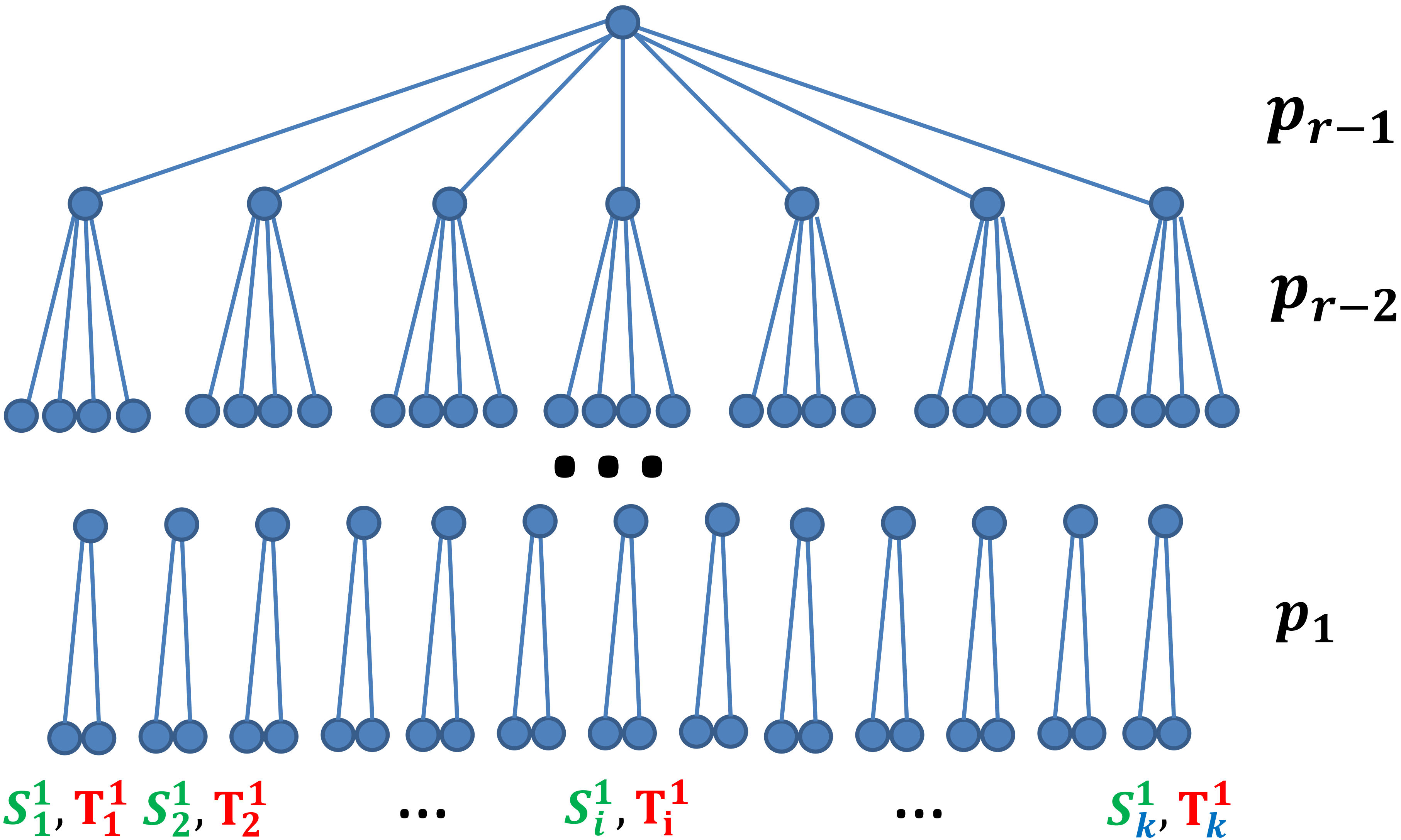
Verification bottom-up

EQUALITY CHECK FAILS =>
RESTART THE SUBTREE





Verification bottom-up



Analysis of Stage i

- p_i = Pr[node at stage i computed correctly]
- Set $p_i = 1 - \frac{1}{(i\log^{r-i-1} \boldsymbol{k})^4}$
  – Run equality checks and basic intersection protocols with success probability p_i
  – **Key lemma**: $\mathbb{E}$[# of restarts per leaf] $= O(1)$ => Cost of Intersection in leafs = $O(\boldsymbol{k})$
  – Cost of Equality = $O(\boldsymbol{k}\ ilog^r \boldsymbol{k})$
- $p_{r-1} = Pr$[protocol succeeds] = $1 - 1/\boldsymbol{k}^4$

Multi-party extensions

$\boldsymbol{m}$ players: $S_1, \ldots, S_m$, where $|S_i| \leq \boldsymbol{k}$

- $S = S_1 \cap \cdots \cap S_m = ?$
- Boost error probability of 2-player protocol to $1 - \frac{1}{2^{\boldsymbol{k}}}$
- Average per player (using coordinator):

$$O(\boldsymbol{k}\ ilog^{r} \boldsymbol{k}) \text{ in } O\left(\boldsymbol{r} \max\left(1, \frac{\log \boldsymbol{m}}{\boldsymbol{k}}\right)\right) \text{ rounds}$$

- Worst-case per player (using a tournament)

$$O\left(\boldsymbol{k}^2 \operatorname{ilog}^{r} \boldsymbol{k} \max\left(1, \frac{\log \boldsymbol{m}}{\boldsymbol{k}}\right)\right) \text{ in } O\left(\boldsymbol{r}\boldsymbol{k} \max\left(1, \frac{\log \boldsymbol{m}}{\boldsymbol{k}}\right)\right) \text{ rounds}$$

Open Problems

- $\boldsymbol{R}^{\boldsymbol{r}}(\boldsymbol{k}$-Intersection) = $O(\boldsymbol{k}\ ilog^{\boldsymbol{r}}\boldsymbol{k})$?
- Better protocols for the multi-party setting?

$\boldsymbol{k}$-Disjointness

- $\boldsymbol{f}(\boldsymbol{S}, \boldsymbol{T}) = 1$, iff $|\boldsymbol{S} \cap \boldsymbol{T}| = 0$
- $\boldsymbol{R}(\boldsymbol{k}\text{–Disjointness}) = \Theta(\boldsymbol{k})$ [Razborov'92; Hastad-Wigderson'96]
- $\boldsymbol{R}^{\mathbf{1}}(\boldsymbol{k}\text{–Disjointness}) = \Theta(\boldsymbol{k} \log \boldsymbol{k})$

[Folklore + Dasgupta, Kumar, Sivakumar; Buhrman'12, Garcia-Soriano, Matsliah, De Wolf'12]

- $\boldsymbol{R}^{r}(\boldsymbol{k}\text{–Disjointness}) = \Theta(\boldsymbol{k} \operatorname{ilog}^{r} \boldsymbol{k})$ [Saglam, Tardos'13]
- $\boldsymbol{R}(\boldsymbol{k}\text{–Disjointness}) = \boldsymbol{\alpha k} + o(\boldsymbol{k})$[Braverman, Garg, Pankratov, Weinstein'13]