Beyond Set Disjointness: The Communication Complexity of Finding the Intersection

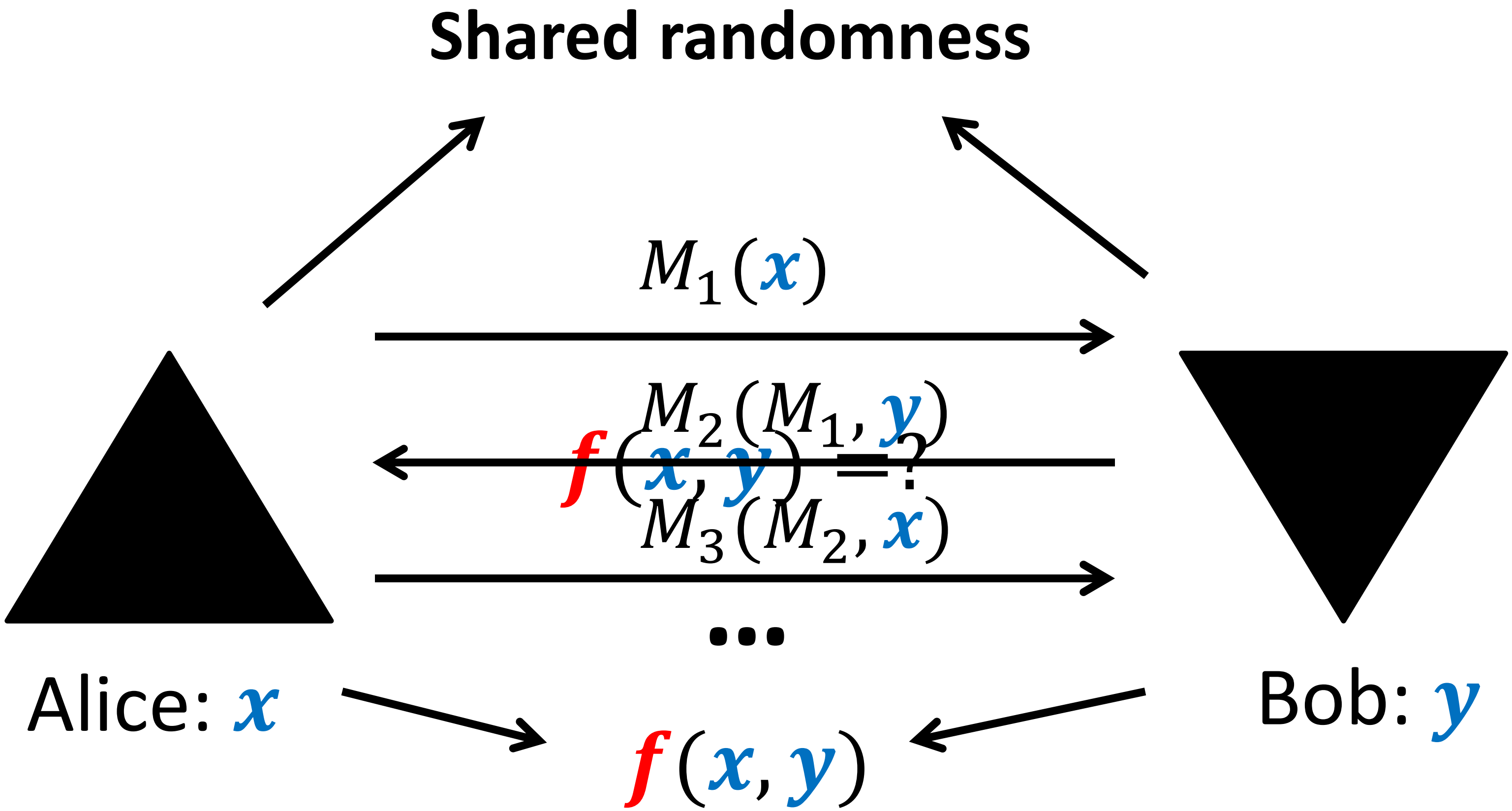
Grigory Yaroslavtsev

http://grigory.us

BROWN ICERM

Joint with Brody, Chakrabarti, Kondapally and Woodruff

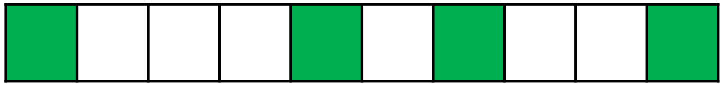
Communication Complexity [Yao'79]

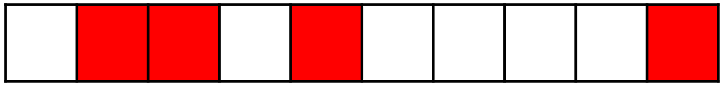


- $R(f)$ = min. communication (error 1/3)
- $R^k(f) =$ min. k-round communication (error 1/3)

Set Intersection

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

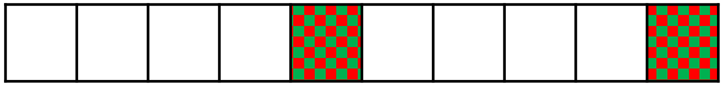




$\boldsymbol{S} \subseteq [n], |S| \leq \boldsymbol{k}$

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

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



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

This talk

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

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

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

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

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

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

k-Disjointness

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

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

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

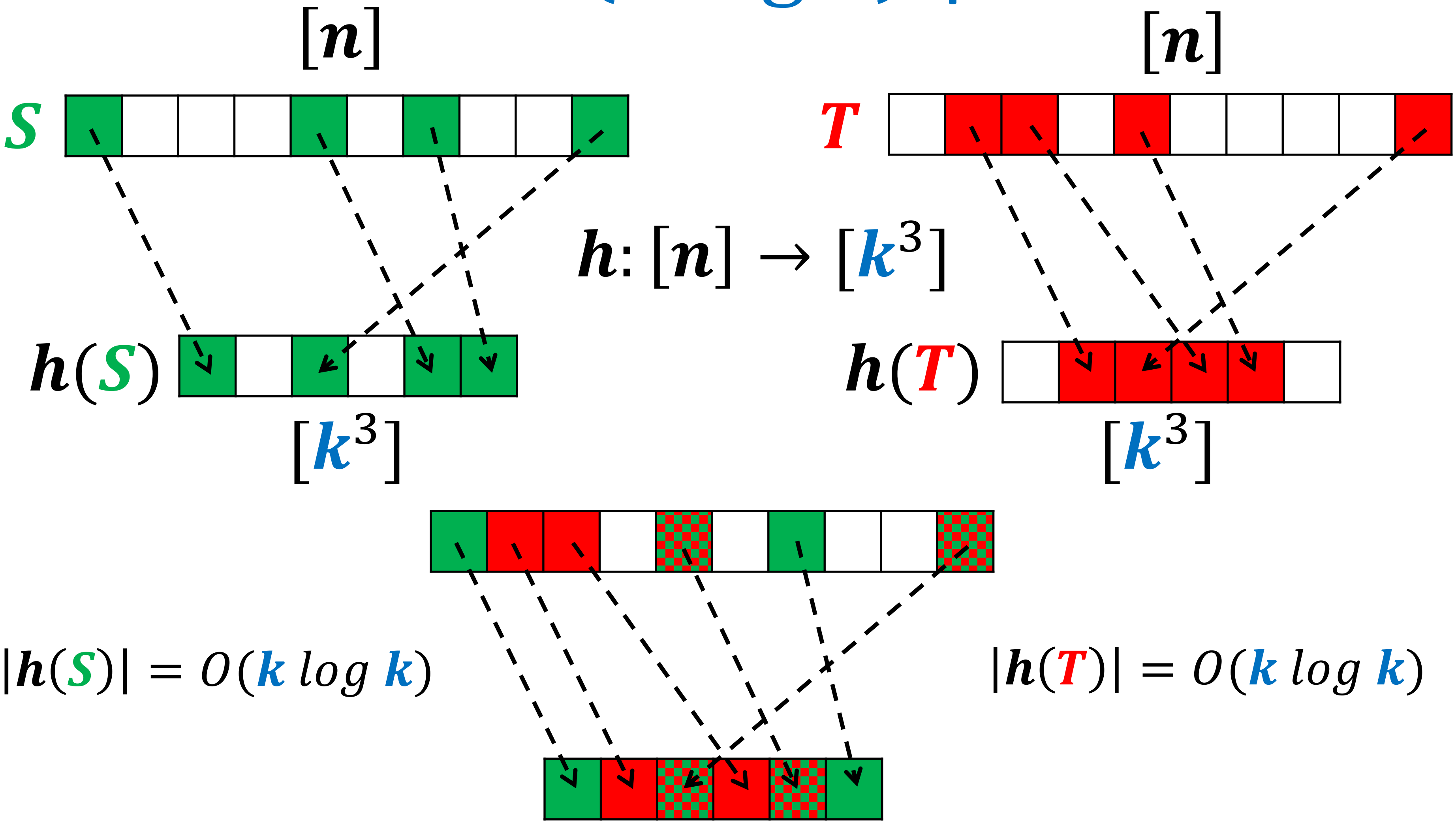
Applications

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

(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: $R(\boldsymbol{S} \cup \boldsymbol{T}) = R(\boldsymbol{S}\, \boldsymbol{\Delta}\, \boldsymbol{T}) = \Theta(\mathrm{k} \log \frac{\mathrm{n}}{k})$

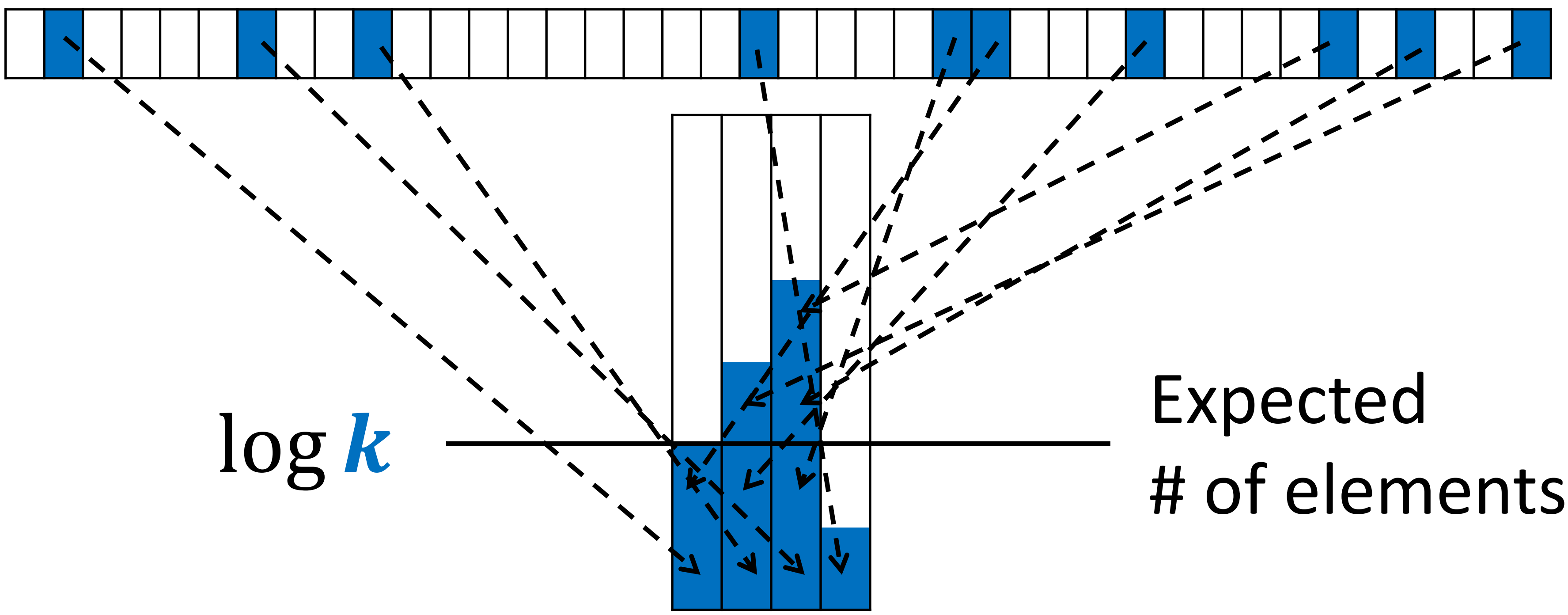
1-round $O(k \log k)$-protocol



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

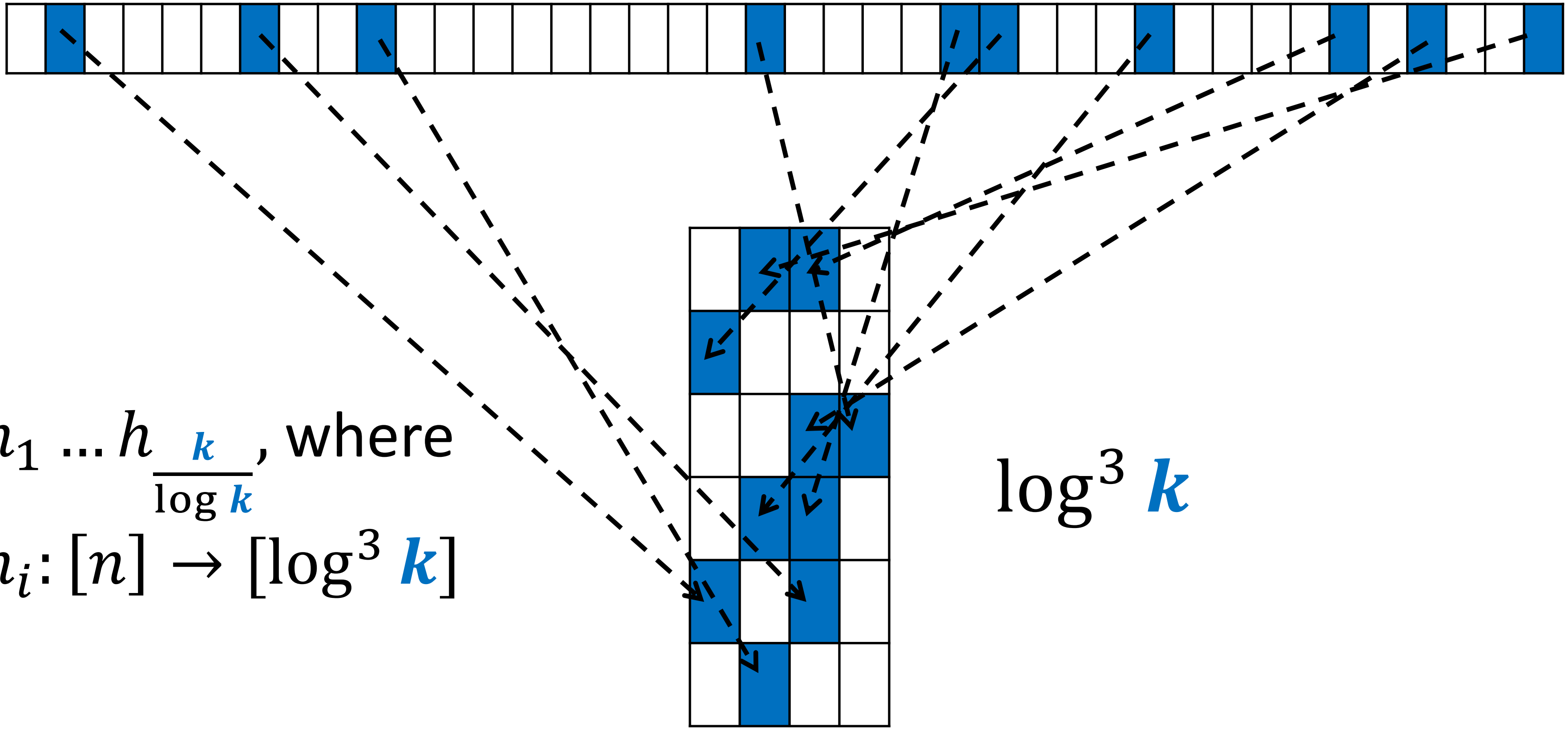
Hashing

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



$$\frac{k}{\log k} = \text{\# of buckets}$$

Secondary Hashing



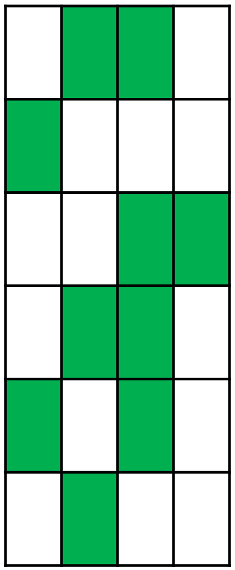
$h_1 \ldots h_{\frac{k}{\log k}}$, where

$h_i : [n] \rightarrow [\log^3 \boldsymbol{k}]$

$\log^3 \boldsymbol{k}$

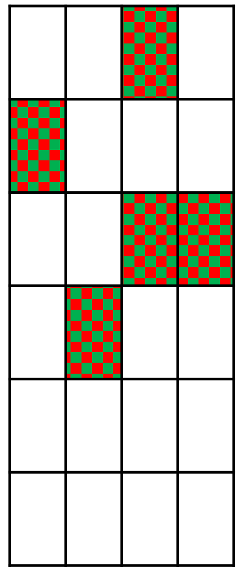
$\frac{\boldsymbol{k}}{\log \boldsymbol{k}}$ = # of hash functions

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

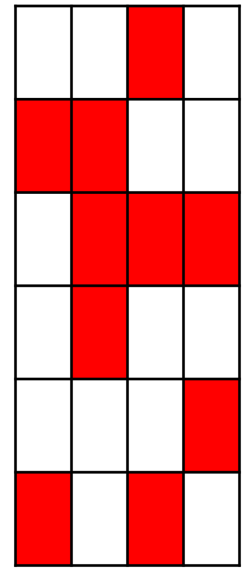


$\log^3 k$

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



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

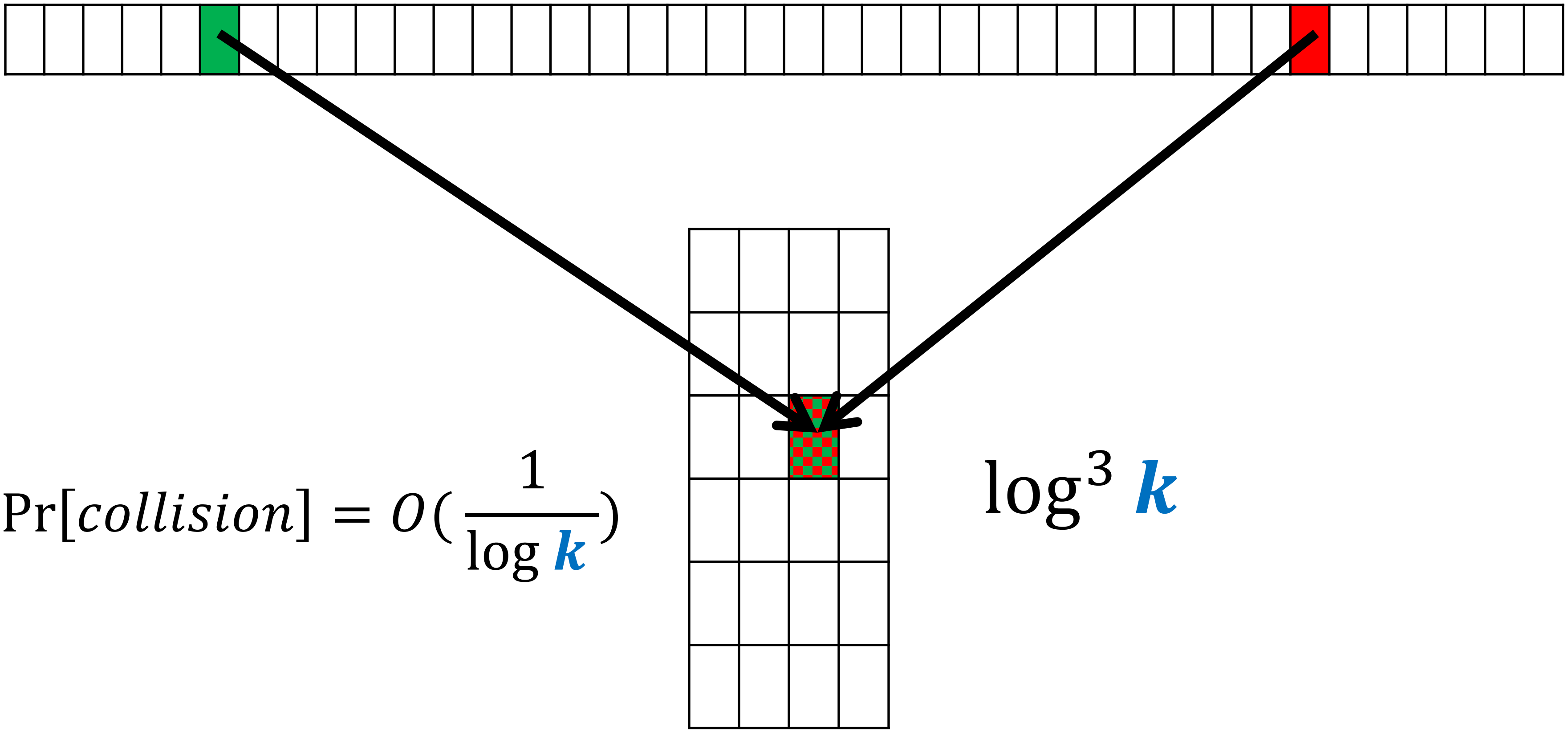


$\log^3 k$

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

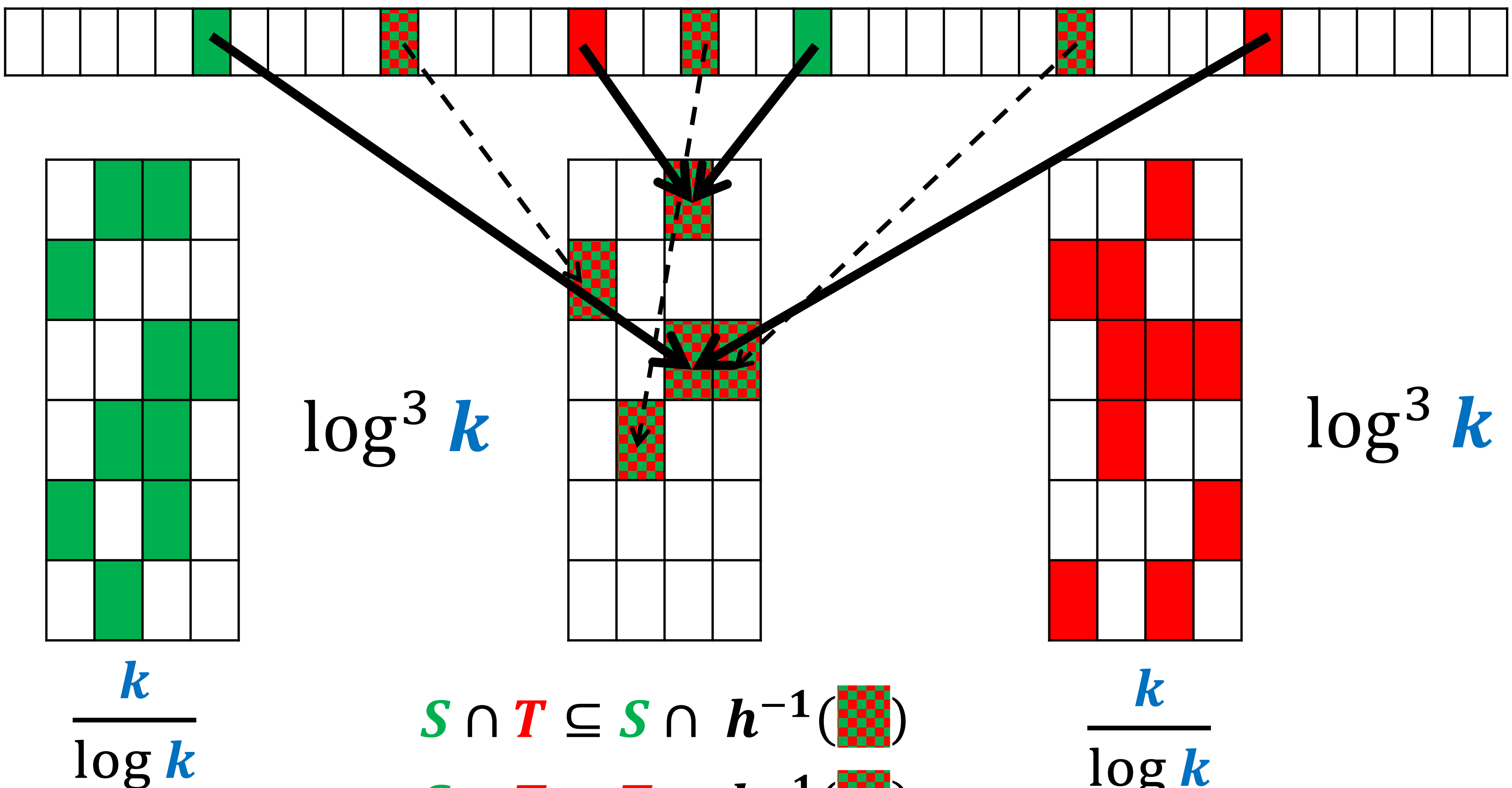
Total communication = $\frac{k}{\log k} O(\log k \log \log k)$ = $O(k \log \log k)$

Collisions



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

Collisions



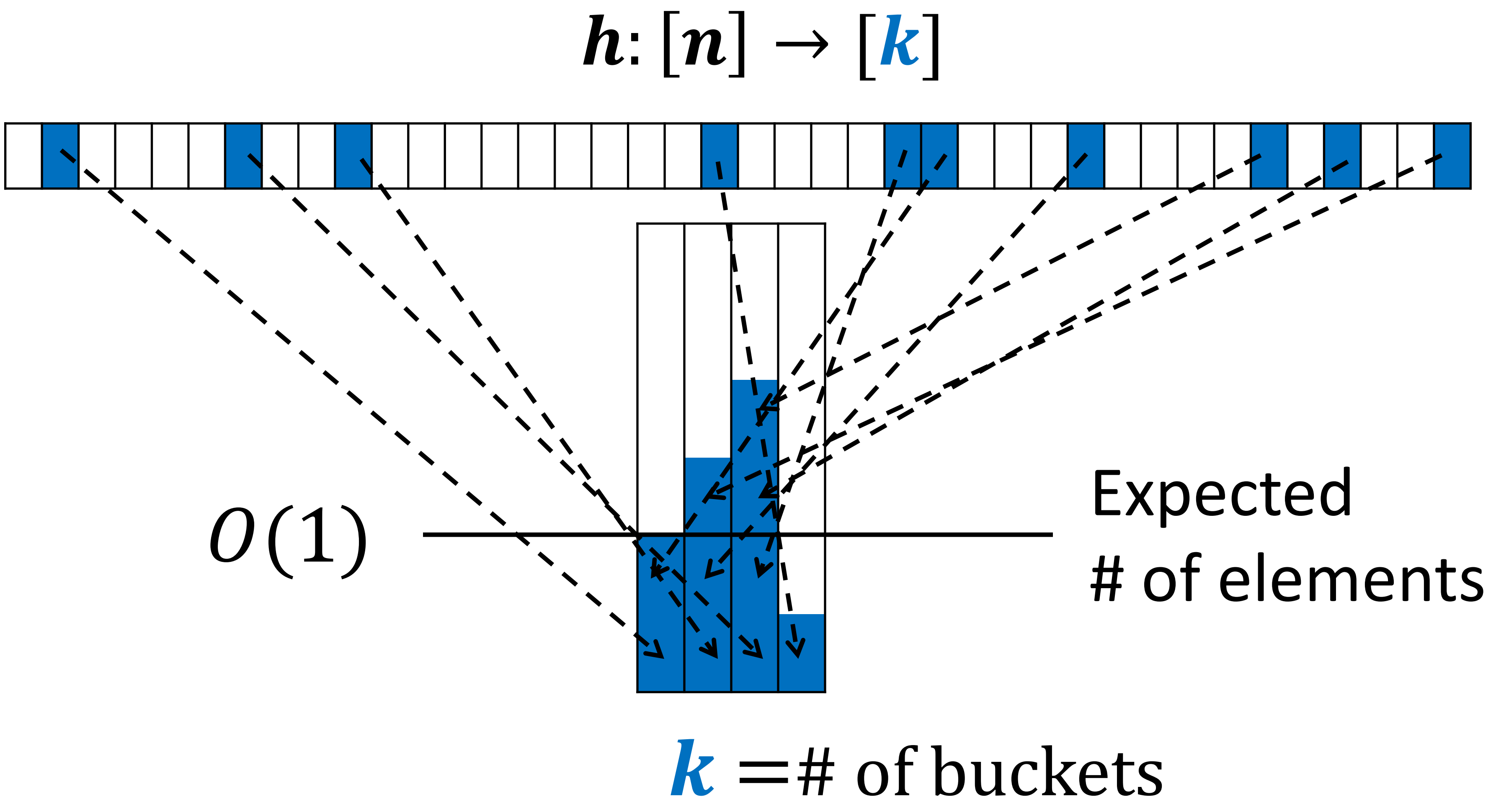
$$S \cap T \subseteq S \cap h^{-1}(\blacksquare)$$

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

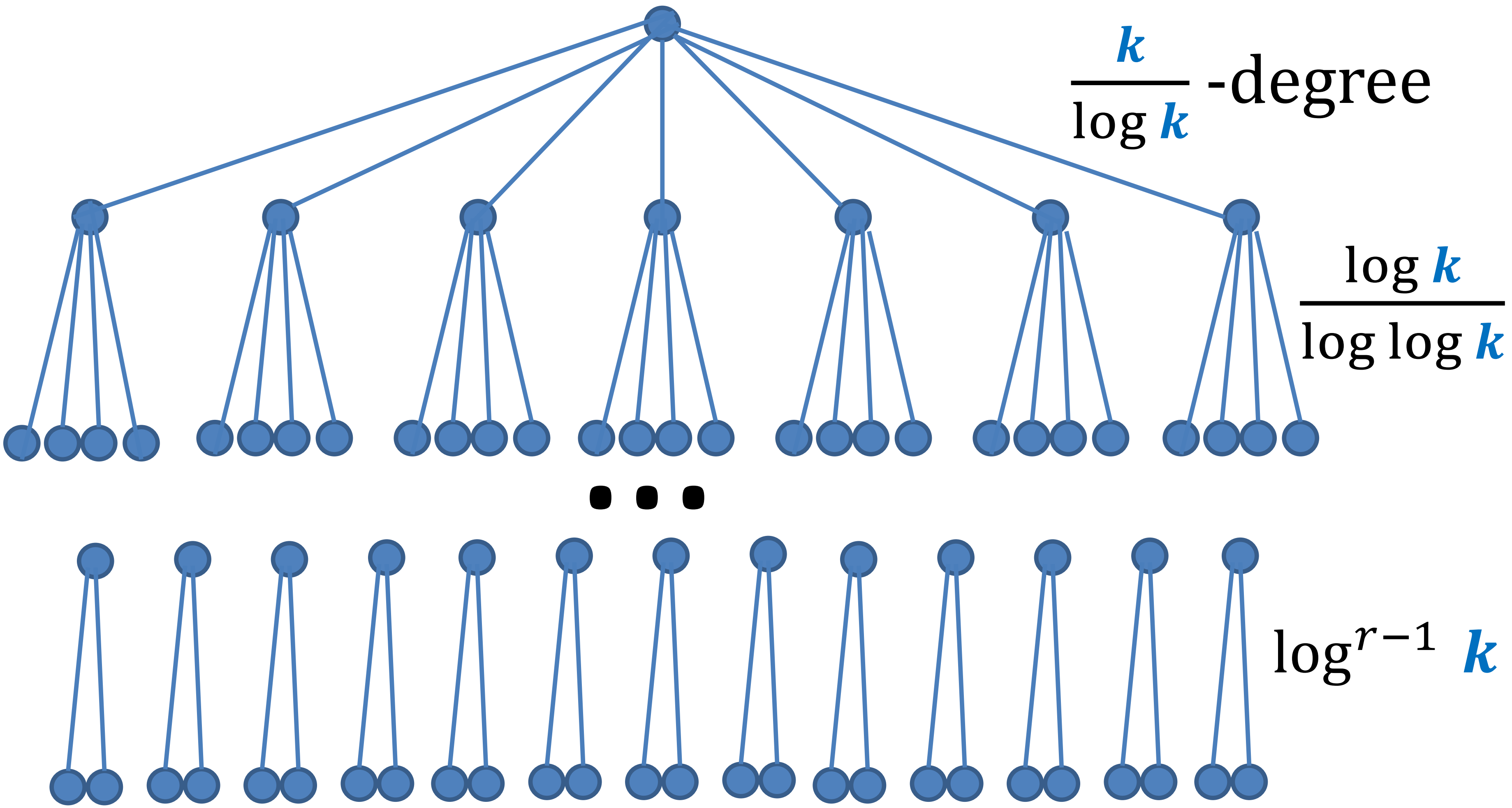
Collisions

- 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
  $$\boldsymbol{S} \cap \boldsymbol{h_i^{-1}}(\blacksquare) = \boldsymbol{T} \cap \boldsymbol{h_i^{-1}}(\blacksquare)$$
  - Expected # of 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

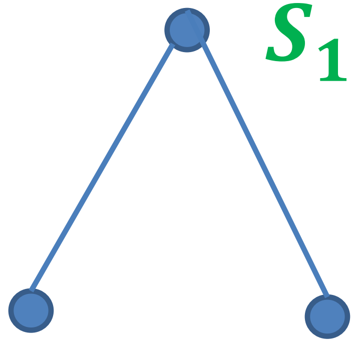


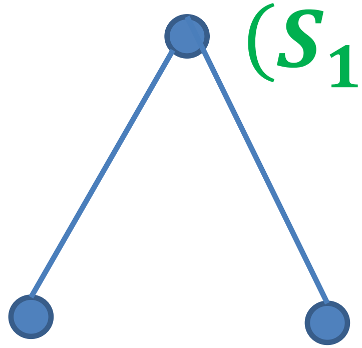
Verification tree



k buckets = leaves of the verification tree

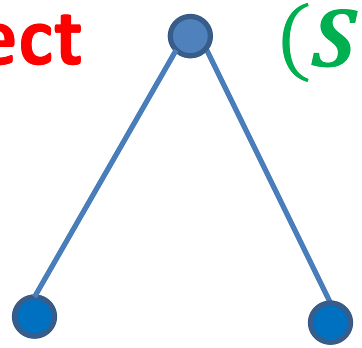
Verification bottom-up





Verification bottom-up

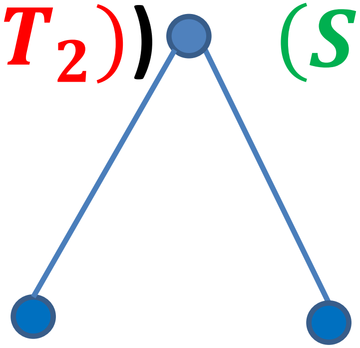
Incorrect $(S_1 \cup S_2) \cap (T_1 \cup T_2)$



Correct Incorrect

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

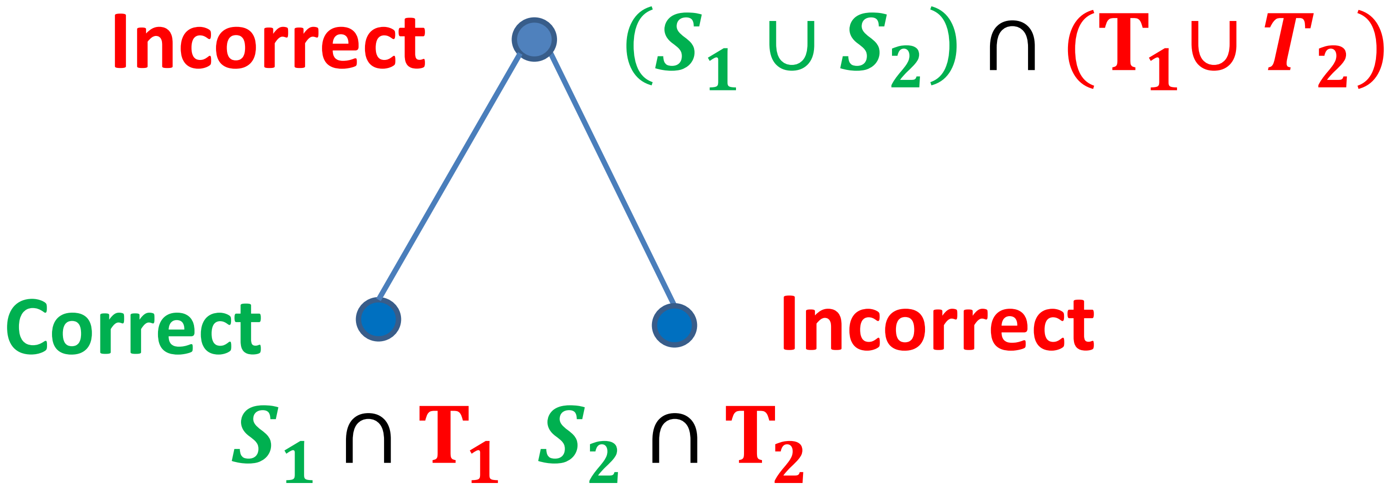
$\mathrm{EQ}((S_1 \cup S_2), (T_1 \cup T_2))$ $(S_1 \cup S_2) \cap (T_1 \cup T_2)$

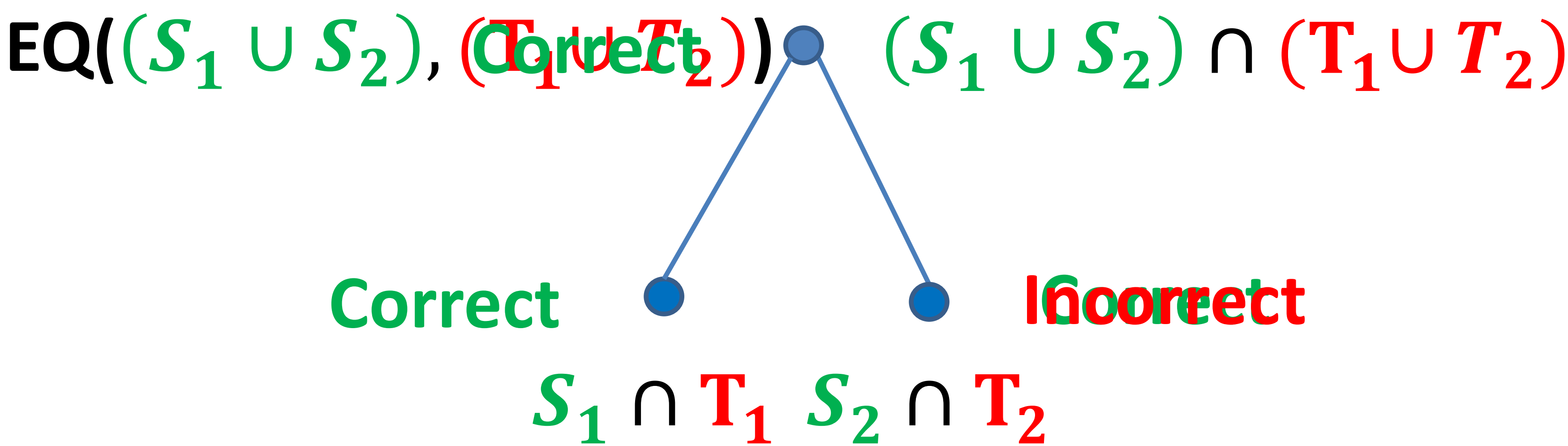


Correct Incorrect

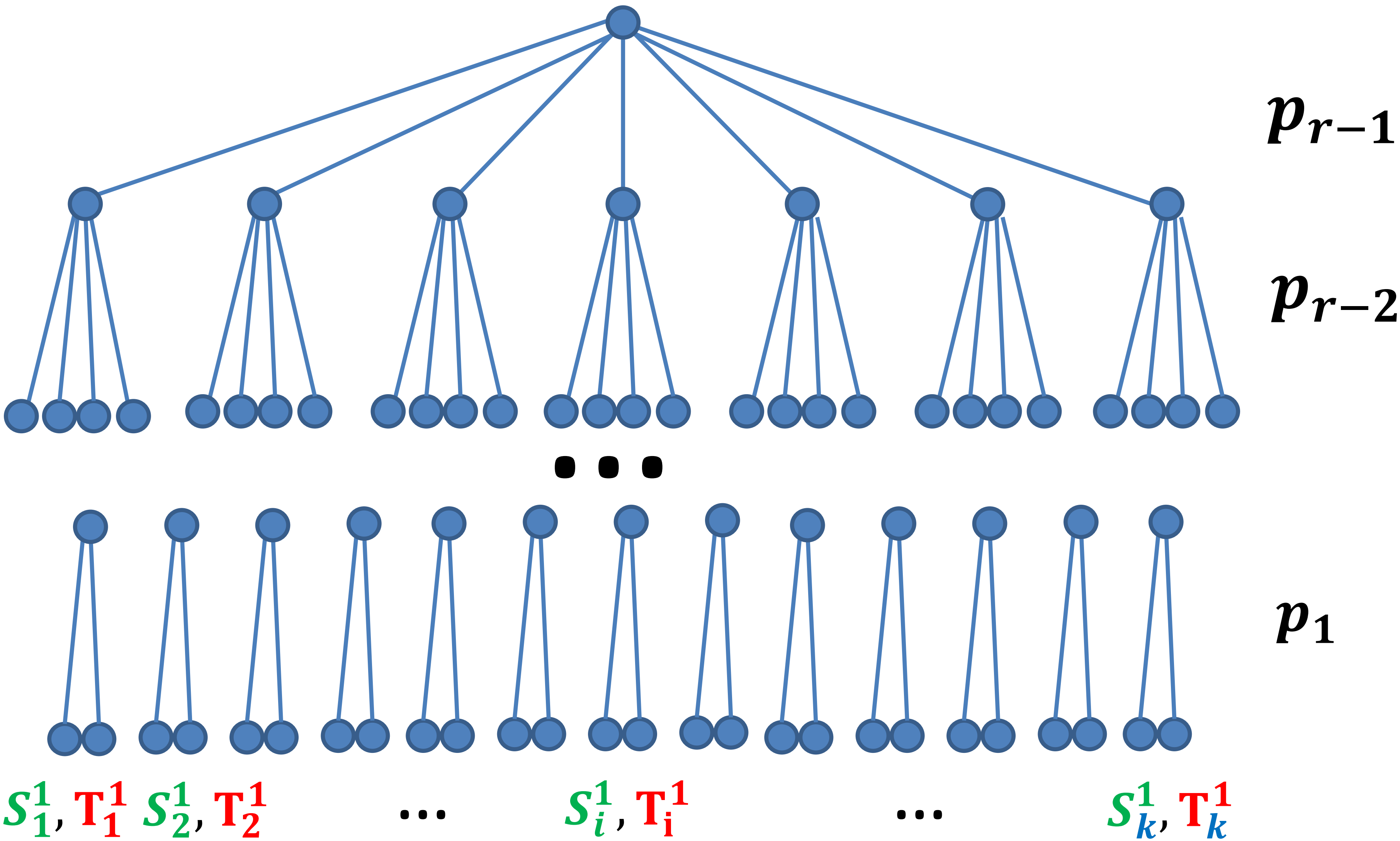
$S_1 \cap T_1 \quad S_2 \cap T_2$

Verification bottom-up





Verification bottom-up



Analysis of Stage i

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

Lower Bound

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

[Brody, Chakrabarti, Kondapally, Woodruff, Y.'13]

- $EQ_{\boldsymbol{m}}(\boldsymbol{x},\boldsymbol{y}) = 1$ iff $\boldsymbol{x} = \boldsymbol{y}$, where $\boldsymbol{x},\boldsymbol{y} \in \{0,1\}^{\boldsymbol{m}}$
- $EQ_{\boldsymbol{m}}^{\boldsymbol{k}}$= solving $\boldsymbol{k}$ independent instances of $EQ_{\boldsymbol{m}}$
- $EQ_{\boldsymbol{m}}^{\boldsymbol{k}}$ reduces to $\boldsymbol{k}$-Intersection:
  - Given $(x_1, \dots, x_{\boldsymbol{k}})$ and $(y_1, \dots, y_{\boldsymbol{k}})$
  - Construct sets with elements $(1, x_1), \dots, (\boldsymbol{k}, x_{\boldsymbol{k}})$ and $(1, y_1), \dots, (\boldsymbol{k}, y_{\boldsymbol{k}})$

Communication Direct Sums

"Solving **m** copies of a communication problem requires **m** times more communication":

$$\boldsymbol{R}^r(\boldsymbol{f}^{\boldsymbol{m}}) = \Omega(\boldsymbol{m})\boldsymbol{R}^r(\boldsymbol{f})$$

- For arbitrary $\boldsymbol{f}$ [... Braverman, Rao 10; Barak Braverman, Chen, Rao 11, ...]
- In general, can't go beyond

$$\boldsymbol{R}(EQ_{\boldsymbol{m}}) = O(1)$$
$$\boldsymbol{R}\left(EQ_{,\boldsymbol{m}}^{\boldsymbol{m}}\right) = O(\boldsymbol{m})$$

Specialized Communication Direct Sums

Information cost $\leq$ Communication complexity

- $\mathrm{R}(\mathrm{Disjointness}) = \Omega(n)$ [Bar Yossef, Jayram, Kumar,Sivakumar'01]

$$\mathrm{Disjointness}(x, y) = \wedge_i(\neg x_i \vee \neg y_i)$$

- Stronger direct sum for bounded-round complexity of Equality-type problems (a.k.a. "union bound is optimal") [Molinaro, Woodruff, Y.'13]

$$\boldsymbol{R}^1\left(EQ^{\boldsymbol{k}}\right) = \Omega(\boldsymbol{k} \log \boldsymbol{k})\boldsymbol{R}(EQ)$$
$$\boldsymbol{R}^r\left(EQ^{\boldsymbol{k}}\right) = \Omega(\boldsymbol{k} \; i\log^r \boldsymbol{k})\boldsymbol{R}(EQ)$$

Extensions

- Multi-party: $\boldsymbol{m}$ players, $S_1, \ldots, S_m$, where $|S_i| \leq \boldsymbol{k}$
  - $S = S_1 \cap \cdots \cap S_m = ?$
  - Boost error probability to $1 - 1/2^{\boldsymbol{k}}$
  - Average per player (using coordinator):

    $O(\boldsymbol{k}\ ilog^{\boldsymbol{r}}\boldsymbol{k})$ in $O\left(\boldsymbol{r} \max\left(1, \frac{\log \boldsymbol{m}}{\boldsymbol{k}}\right)\right)$ rounds
  - Worst-case per player (using a tournament)

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

Open Problems

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