# Complexity of Finding a Duplicate in a Stream: Simple Open Problems

Jun Tarui

Univ of Electro-Comm, Tokyo

Shonan, Jan 2012

# Duplicate Finding Problem

Given: Stream $a_1, \ldots, a_m$ $a_i \in \{1, \ldots, n\}$

Assuming $m > n$, find a duplicate $d = a_i = a_l$ $(i \neq l)$

Finding just one, any duplicate suffices.

Exists by the pigeonhole principle.

# Duplicate Finding Problem

For $m = n+1$, a deterministic algorithm with $O(\log n)$ space and $O(1)$ passes exist? [Muthu, talk@Kyoto 05]

$\rightarrow$ No [T. 07] (-->Muthu's survey05)

Main-Point-of-Talk: Open Question 1:

the same question for the case $m = 2n$

( or $m = n^2$ or $m=\text{poly}(n)$ )

# Finding a missing item

Assuming $m < n$, find a missing item:

$x \in \{1,\ldots,n\}$ but $\notin \{a_1, \ldots, a_m\}$

a dual problem

but no known black-box reductions

Our lower bounds for space--#passes trade-off apply for both problems.

# Simple algorithms and our lower bounds

0. In RAM model, O(log n)-space O(n)-time by 2 pointers.

1. In the 1st pass, count # of $a_i$'s in [1 n/2] and in (n/2 n] ...

O(log n) space, O(log n) passes

$\rightarrow$ With O(log n) space, needs Ω(log n/loglog n) passes

2. With two passes: In the 1st pass, count # of $a_i$'s in
[1, √n], (√n, 2√n], …, (n-√n, n]. Space O(n^1/2 log n)
# of blocks n^1/2 $\rightarrow$ (n/log n)^1/2: Space O((n log n)^1/2 )

With k passes, space O(n^(1/k) (log n)^(1-1/k) )
$\rightarrow$ With k passes, needs space Ω( n^(1/ 2k-1) )

3. $m=2n$: Randomly choose $i \in \{1,\ldots, m\}$.

Check if $d=a_i$ occurs in $a_{i+1}, \ldots, a_m$.

If so, report d as a duplicate; otherwise report "failure"

one pass, O(log n) space, success prob ≥ ½, Las-Vegas

For $m=n+1$: one-pass Las-Vegas needs space $\Omega(n)$

For $m=n+1$:

One-pass Monte-Carlo (error < ¼) with O(log^3 n) space

[Gopalan-Radhakrishnan SODA09]

improved to O(log^2 n) [Jowhari-Sagiam-Tarods PODS11]

Open Problem 2: Reduce space to O(log n)

# Result for multiple-pass algorithms

Result 1:

Assume that m=n+1. A streaming algorithm with O(log n) space requires Ω(log n/loglog n) passes.

A k-pass algorithm requires Ω(n^(1/ 2k-1) ) space.

The same bounds apply for finding a missing-item with m=n-1.

# Results for one-pass algorithms

2. For any m > n (including $m=\infty$), if P is a deterministic read-once branching program that finds a duplicate, then the number of non-sink nodes in P is at least 2^n.

3. Assume that m = n+1. Let P be a Las-Vegas randomized oblivious read-once branching program that finds a duplicate with prob ≥ ½. Then, the number of nodes in P is at least 2^(n/4 – o(1)).

a result similar (but different) to 3 in

[Razborov-Wigderson-Yao02: Read-Once Branching Programs, Rectangular Proofs of the Pigeonhole Principle and the Transversal Calculus]

# Proof Sketch of Result 1

1. Relate to the Karchmer-Wigderson communication game for Majority

2. Apply well-known size lower bounds for constant-depth circuits computing Majority

Remark: First reduce to a comm complexity problem; but finish off using circuit bounds

Assume that $m=n+1$ is even.

Consider inputs:

$A=\{a_1, \ldots ,a_{m/2}\}$ all distinct $\rightarrow$ Alice

$B=\{a_{m/2+1}, \ldots, a_m\}$ all distinct $\rightarrow$ Bob

Alice and Bob must find some $j \in A \cap B$.

In one round, Alice $\rightarrow$ Bob or Bob $\rightarrow$ Alice

s-bit r-pass streaming algorithm

$\rightarrow$ s-communication-bit (2r-1)-round protocol

Karchmer-Wigderson communication game for a (monotone) Boolean function $f : \{0,1\}^n \to \{0,1\}$

Alice: $x \in \{0,1\}^n$ : $f(x)=1$ (minterm)

Bob: $y \in \{0,1\}^n$ : $f(y)=0$ (maxterm)

Find j such that $x_j \neq y_j$ ($x_j=1$ and $y_j=0$)

communication complexity

= min depth of AND/OR circuits for f

# of rounds $\leftrightarrow$ # of AND/OR alternations

Majority($x_1$, …, $x_n$) = 1 if $\Sigma x_i \geq n/2$, and 0 otherwise.

Assume n is odd.

minterms = maxterms = (n+1)/2-subsets of {1,…,n}

A, B: (n+1)/2-subsets of {1, …, n}

Alice gets A, Bob gets B; they must find $j \in A \cap B$

$\leftrightarrow$ monotone circuits computing Majority

Apply size lower bounds for monotone constant-depth circuits [Boppana86]. (the same bound for general circuits later given by [Hastad87] )

size $\rightarrow$ fan-in of each gate end-of-proof-sketch

# The proof breaks down for bigger m

Consider $f(x) = 1$ if $\Sigma x_i \geq n/2 + \varepsilon(n)$;

$0$ if $\Sigma x_i \leq n/2 - \varepsilon(n)$.

For $\varepsilon(n) = n/\text{polylog}(n)$, computable by poly-size O(1)-depth circuits [Ajtai-BenOr84]

$\rightarrow$The same argument applied to space O(log n) algorithms fails to yield an $\omega(1)$ bound for # of passes if $m \geq (1 + 1/\text{polylog}(n))n$

# Deterministic one-pass algorithms

Task 1: Find a duplicate d.

Task 2: Find d together with $i \neq l$ such that $d = a_i = a_l$.

an n-way read-once branching program

[RWY02] For Task 2, # of nodes ≥ $2^{\Omega(n \log n)}$.

Result 2: For Task 1, # of non-sink nodes ≥ $2^n$.

Both results hold for any m > n, including m=∞

# Proof sketch of Result 2

For node v, define $K[v] = \{ j \in \{1,\ldots,n\} :$ Every path to v includes "$a_i=j$" $\}$

Claim: $\{K[v] : v \text{ node}\}$ = the power set of $\{1, \ldots, n\}$

Assume otherwise and consider an inclusion-minimal $A \subseteq \{1,\ldots,n\}$ that does not appear as K[v].

E.g., $A = \{1,2,4\}$. For "$a_i=?$" at node v with $K[v]=\{1,2\}$, the adversary responds: $a_i=4$. end-of-proof-sketch

# Open Problems Restated

1. Show that O(log n)-space O(1)-pass is impossible for m=2n deterministic duplicate finding
(or no matter how big m is)

2. For the case m=n+1, give a Monte-Carlo randomized algorithm that finds a duplicate with 1 pass and O(log n) space.

connection to
the proof complexity of the pigeonhole principle?

Thanks!

## How should I get to Kyoto from here?

Go to “Shin-Yokohama” JR station, and take

a Nozomi Shinkansen (bullet train); takes 2 hours to get to Kyoto; runs every 10 minutes; reservation not needed

How should I get to Shin-Yokohama?

Get to Zushi station by bus or taxi;

go to Yokohama station by JR trains;

go to Shin-Yokohama by JR trains