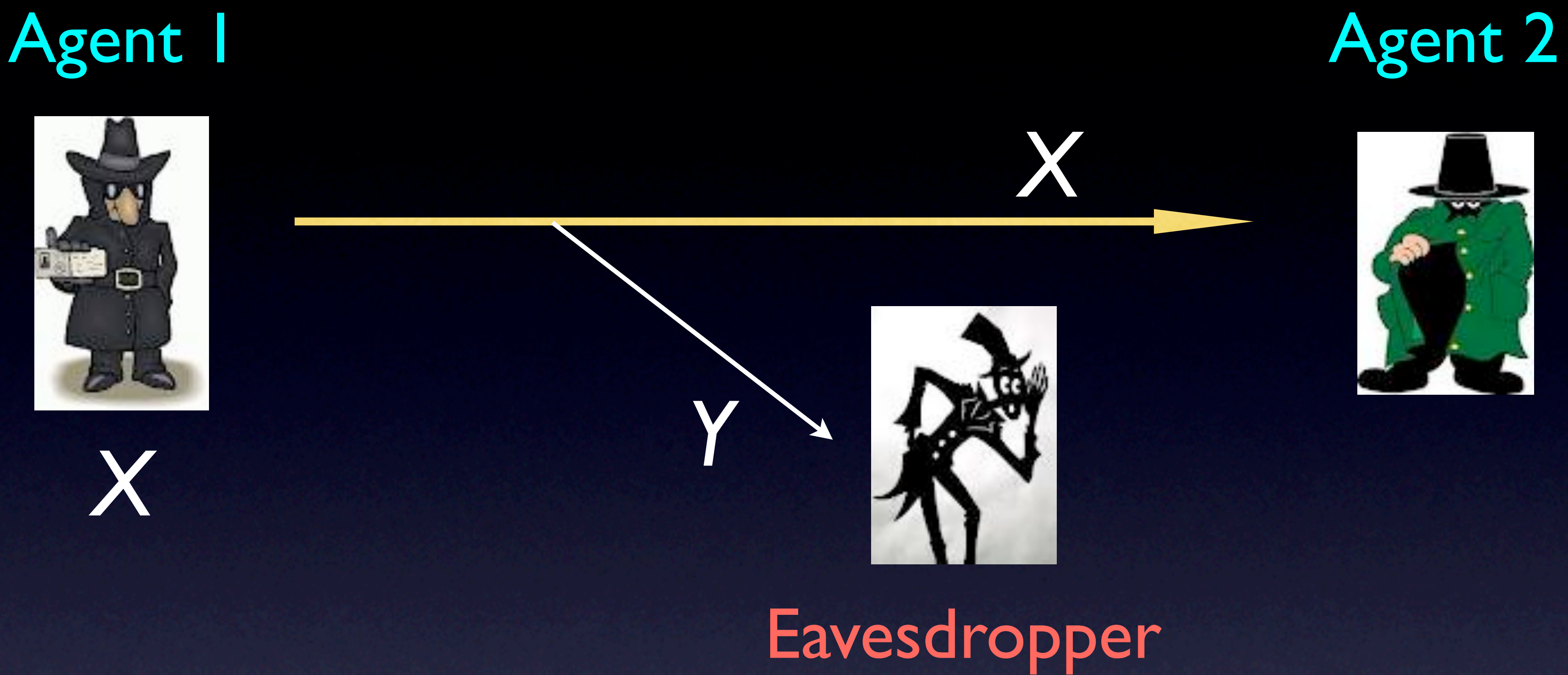
Communication complexity and the information cost approach

Ashwin Nayak
University of Waterloo

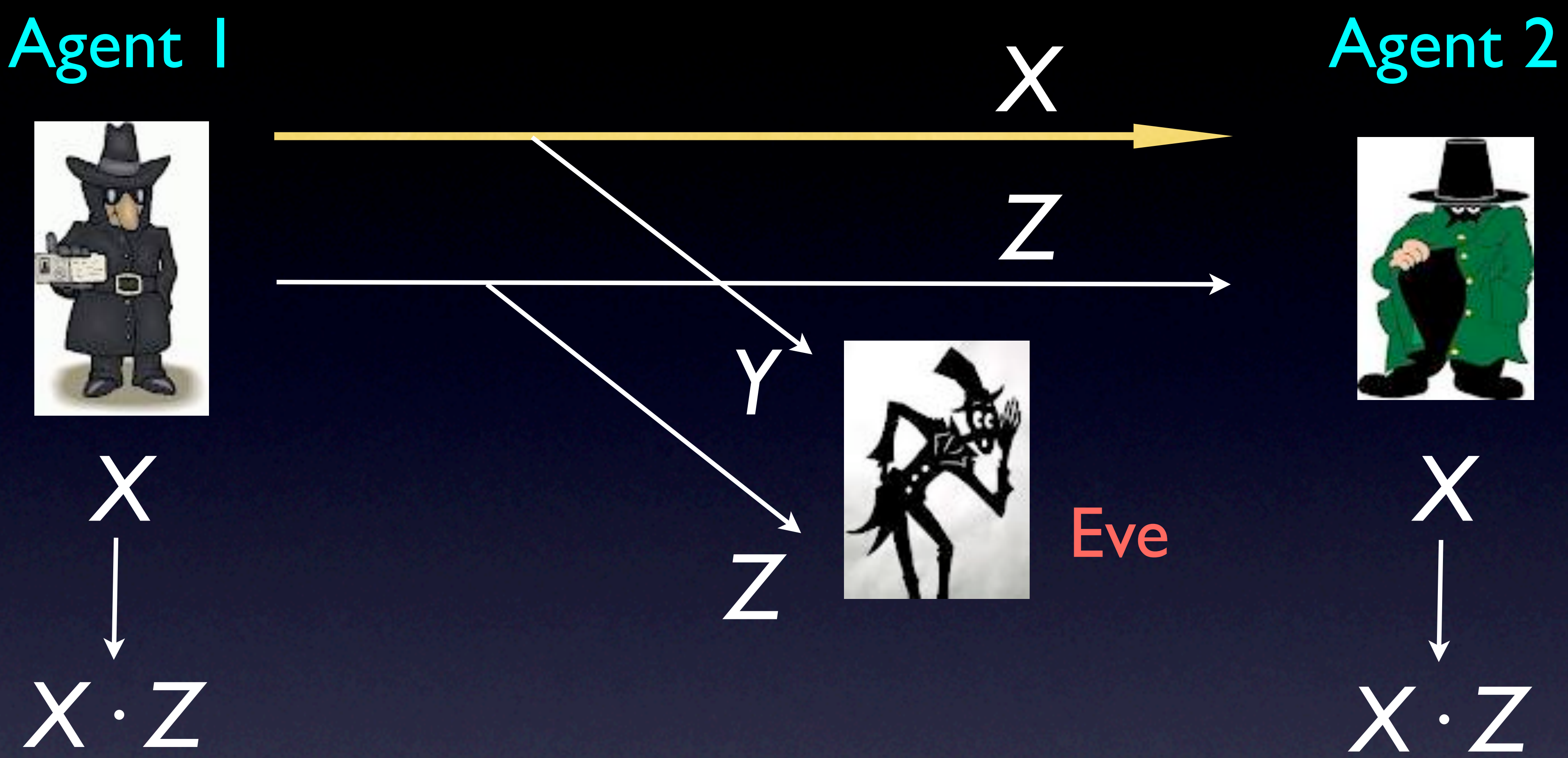
Application I

Privacy amplification



- A1 shared n uniformly random bits X with A2
- Leaked information Y with $m << n$ bits to Eve
- Can they distil more secure key ?

2-universal hashing



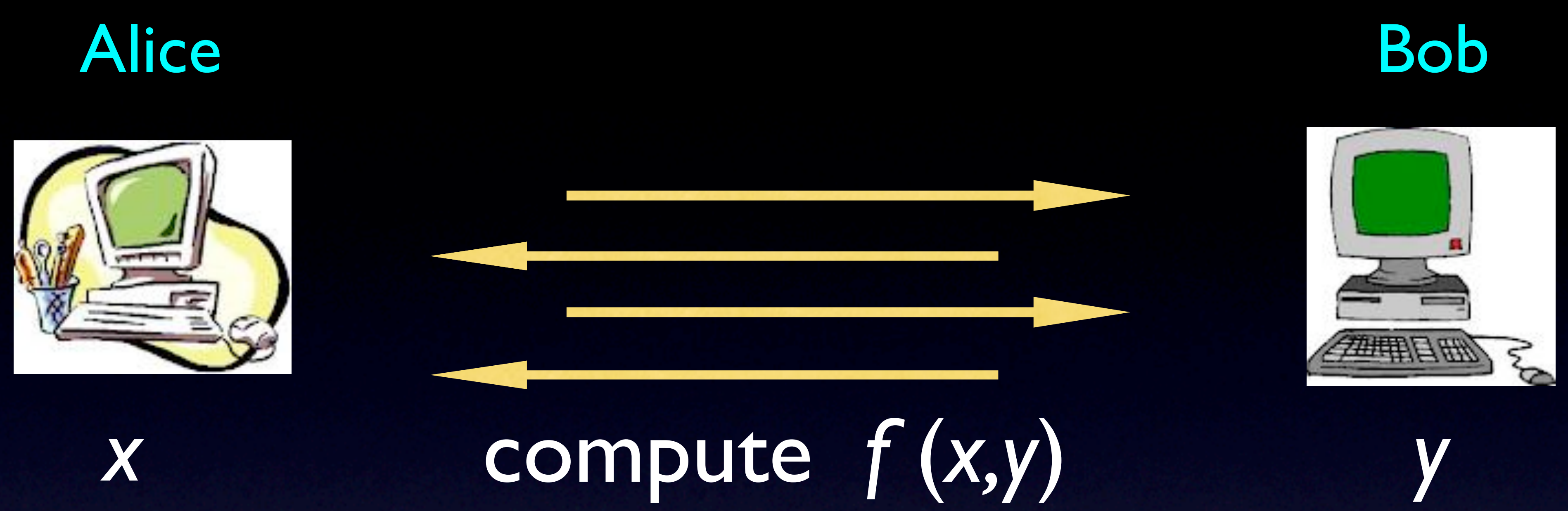
- A1 generates n uniformly random bits Z, sends to A2
- Both compute scalar product (mod 2) $B = X \cdot Z$
- Eve sees Z
- How well can she guess B from Y, Z ?

Communication complexity view [Ben-Or]

- A1 gets string X, Eve gets Z
- A1 sends m-bit message Y to Eve
- Eve estimates scalar product (mod 2) $B = X \cdot Z$
- What is the probability of correct estimate if Y is short ?
- Equivalently, for probability of correctness $1/2 + \varepsilon$, how long does Y have to be ?

The model of computation

Two-party communication



- Would like to compute function f on some input
- Input distributed among two computers as x, y respectively
- Alice and Bob send messages to each other, depending on input and previous messages, may use randomization
- Local computation of the messages is (cheap)
- Need to compute f with minimum communication, or messages, etc. (expensive)
- Can tolerate a small probability of error

Example: Equality

Alice Bob

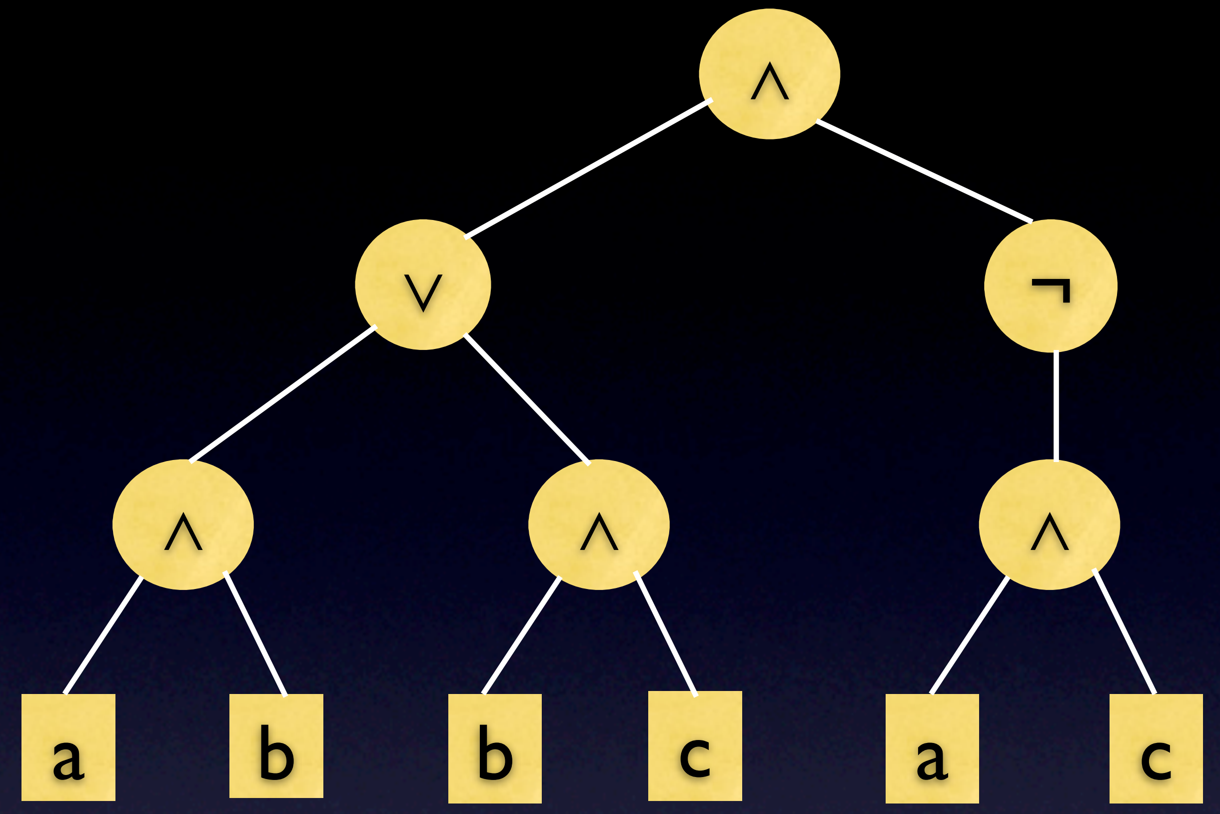
x Is $x = y$? y

- Alice may send x, Bob computes the answer
- Costs n bits, one message
- Cannot reduce cost with deterministic protocols, even with many message exchanges

- Randomization helps
  - Alice and Bob encode x, y using the same *good* error-correction code (length cn, distance δn) into $C(x)$, $C(y)$
  - Alice picks uniformly random i, sends i, $C(x)_i$
  - Bob outputs "equal" if $C(x)_i = C(y)_i$, "not equal"
- Correctness
  - If $x = y$, output is always "equal"
  - If not, the two bits are different with probability at least δ/c
  - By repeating for several indices, can increase probability of correctness
- Cost
  - $O(\log n)$, single message (constant rate, constant distance codes exist), is optimal

Application II

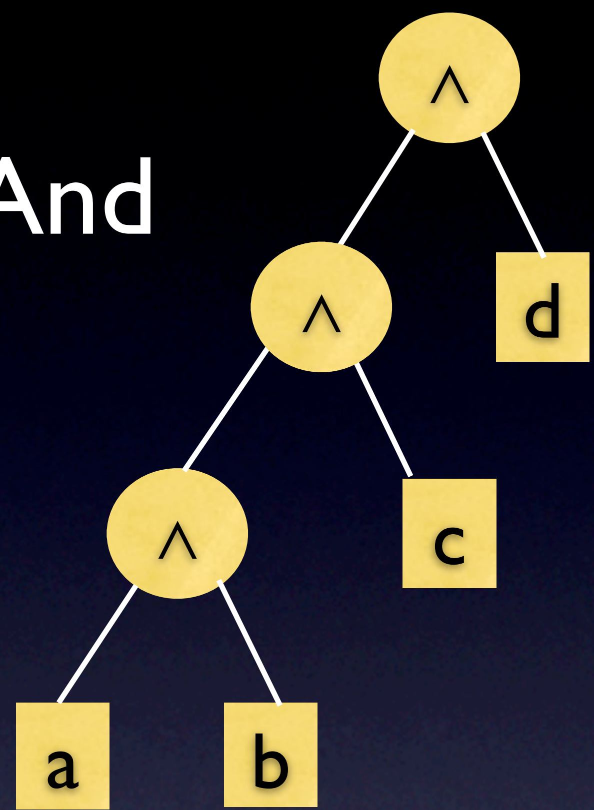
Formula size



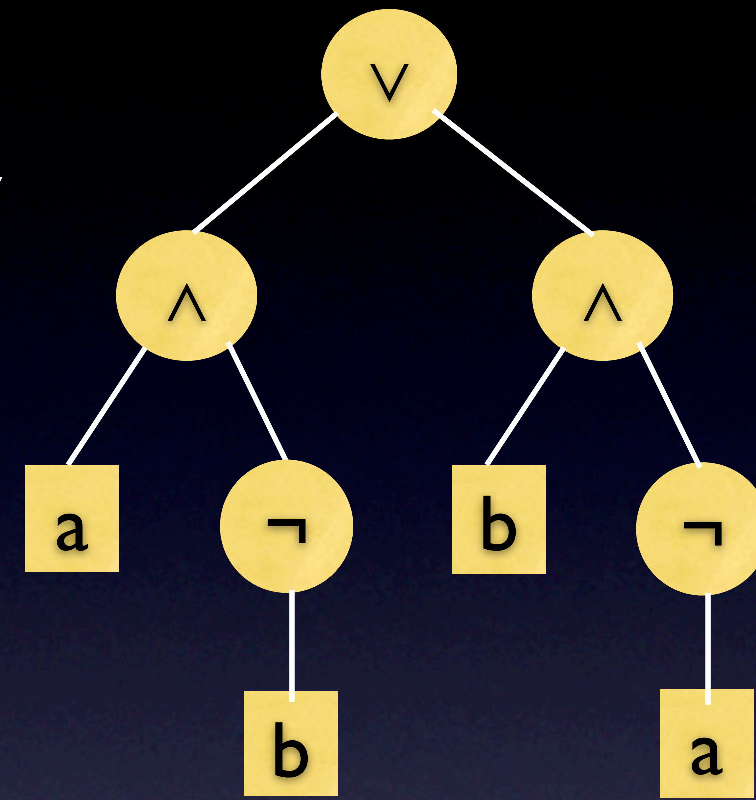
- Boolean formula over $\{ \neg, \wedge, \vee \}$, fan-in 2
- Computes function of variables in the leaves
- Size = number of gates ≈ number of leaves
- Given function f, what is the smallest formula that computes it ?

Examples

And

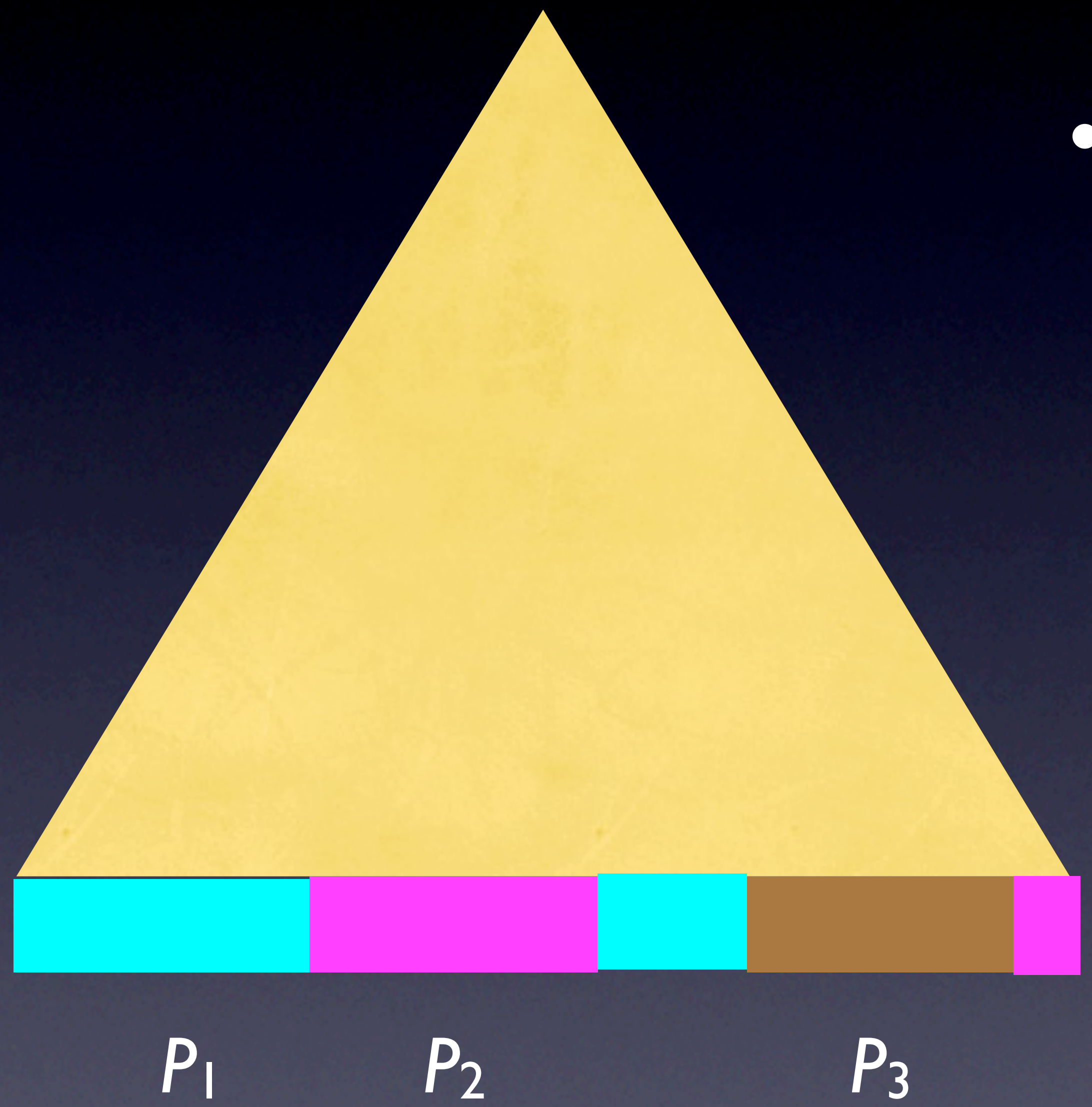


Parity

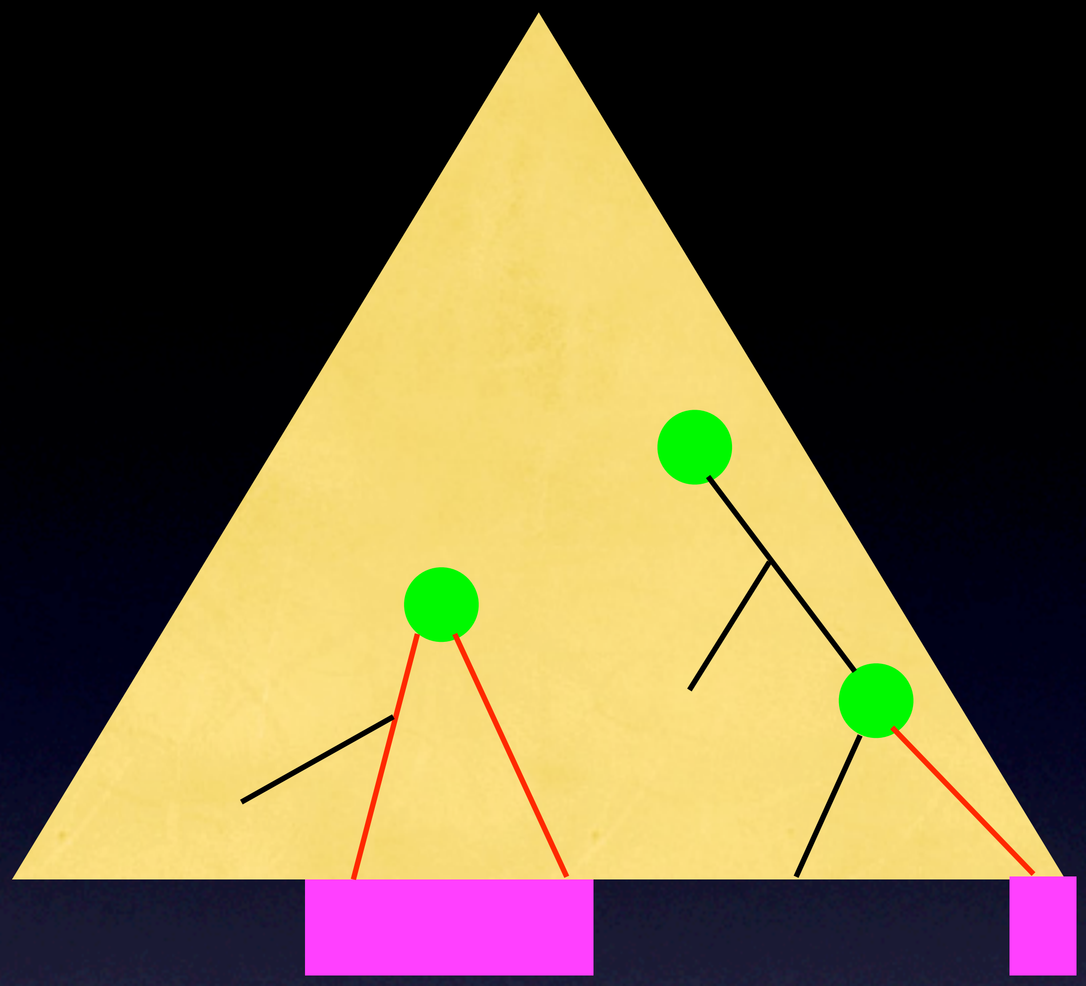


- And of n bits takes size n
- Parity of n bits takes size $O(n^2)$
- How do we show optimality ?

Bound à la Neciporuk [Klauck]



- Partition variables into sets $\{ P_i \}$
- For each i, consider a one-message communication problem:
  - Alice gets values of variables not in P_i
  - Bob gets values of variables in P_i
  - Let $D(f_i)$ be the communication complexity of computing the formula with one message from Alice
- Formula size $\geq$ $(1/4) \sum_i D(f_i)$



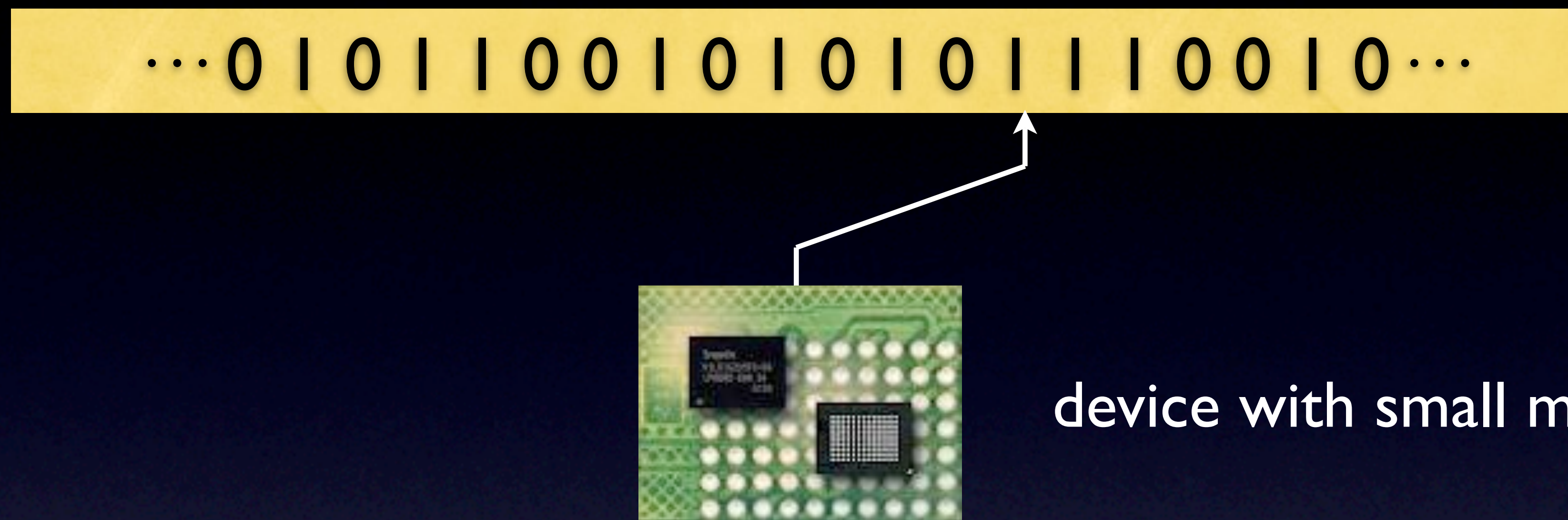
Formula size $\geq$ (1/4) $\sum_i$ D(f_i)

- Let L_i be the number of leaves with variables in P_i
- Formula size $= \sum_i L_i$
- Suffices to show D(f_i) $\leq$ 4 L_i
- Green nodes: those with at least one descendent in P_i

- Bob can evaluate formula if he knows the influence of Alice's inputs on the paths between these nodes (or root or leaf in P_i)
- For each such path, Alice can specify her influence by 2 bits
- # Paths $\leq$ 2 # green nodes + 2 $\leq$ 2 L_i
- D(f_i) $\leq$ 2 # Paths $\leq$ 4 L_i

Application III

Space complexity of streaming algorithms



Streaming model

- massive input, cannot be stored entirely in memory
- input arrives sequentially, read one symbol at a time
- device processes each symbol quickly, while maintaining small workspace

Important for network traffic analysis, genome decoding, web databases, ...

Streaming algorithms

Streaming algorithms with constant memory and time per symbol are precisely finite automata

Advantage for more complex natural problems ?

- Context-free languages: e.g., checking whether a sentence is grammatical
- For Dyck(2), checking if an expression in two types of parentheses is well-formed ?
    - ([ ] ( )) is well-formed
    - ([ )( ]) is not well-formed
- Canonical CFL, used in practice: checking well-formedness of large XML file

Streaming algorithms for Dyck(2)

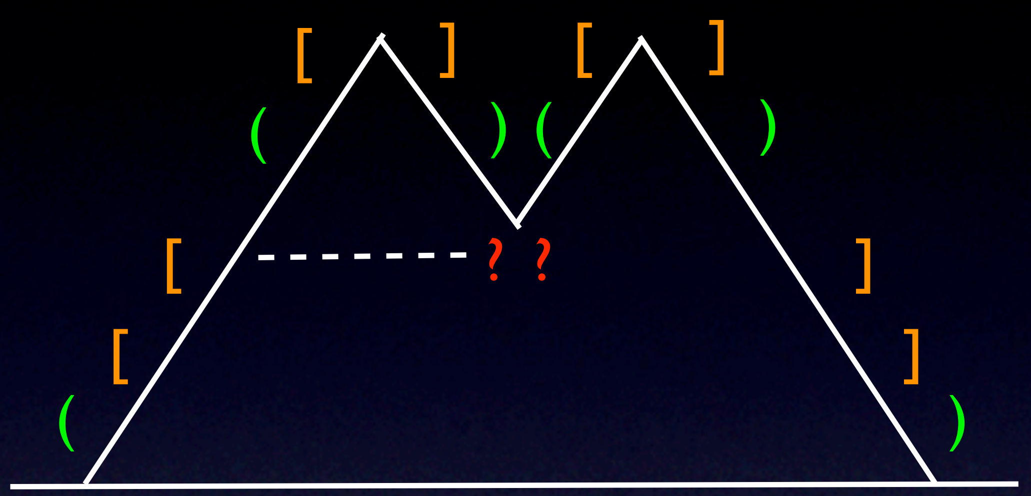
Magniez, Mathieu, N.'10:

- A single pass randomized algorithm that uses $O((n \log n)^{1/2})$ space, $O(\text{polylog } n)$ time/ symbol
- 2-pass algorithm, uses $O(\log^2 n)$ space, $O(\text{polylog } n)$ time/ symbol, second pass in reverse
- Space usage of 1 pass algorithm is optimal, via communication complexity

Jain, N.'10:

- Space usage of unidirectional T-pass algorithm is $n^{1/2} / T$

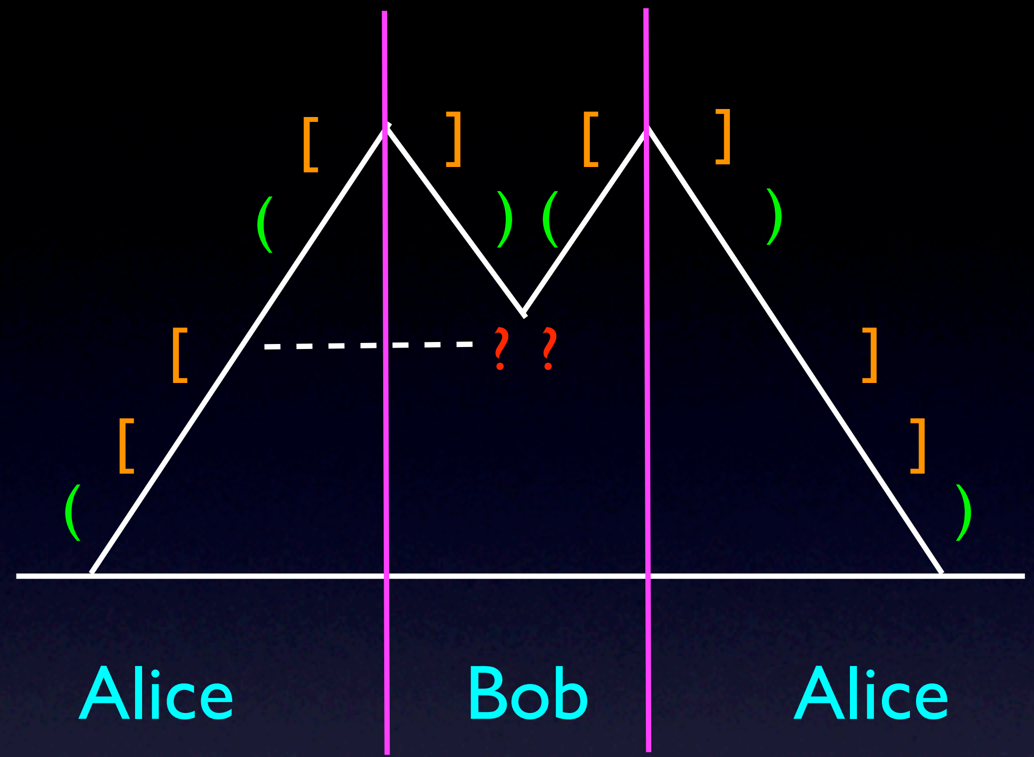
Connection to communication complexity [MMN'10]



Consider the following instance of length between $2n$ and $4n$

- n opening parenthesis, followed by k-1 closing, matched
- single closing parenthesis, of either kind
- followed by a mirror image of the same
- instance is well-matched iff kth closing parenthesis matches

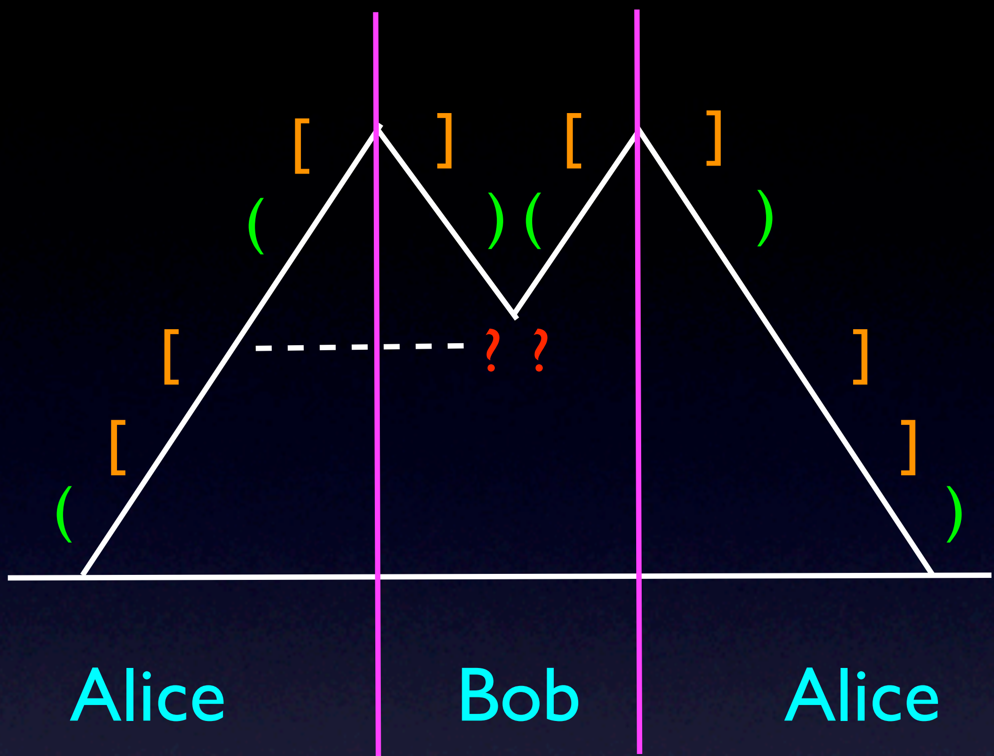
Communication problem



Distribute the instance between Alice and Bob as follows:

- first and last n symbols go to Alice
- middle $2k$ symbols go to Bob
- need only represent type of parenthesis by one bit, as opening/closing is evident

Equivalent problem: Augmented Index



Variant of Index function

- Alice has n-bit string x, Bob has the prefix $x[1, k\text{-}1]$, and a bit b.
- Need to check if $b = x_k$.

One-pass algorithm with space S implies a protocol with communication $2S$, with two messages:

$x = x_1\, x_2 \ldots x_n$

Is $b = x_k$?

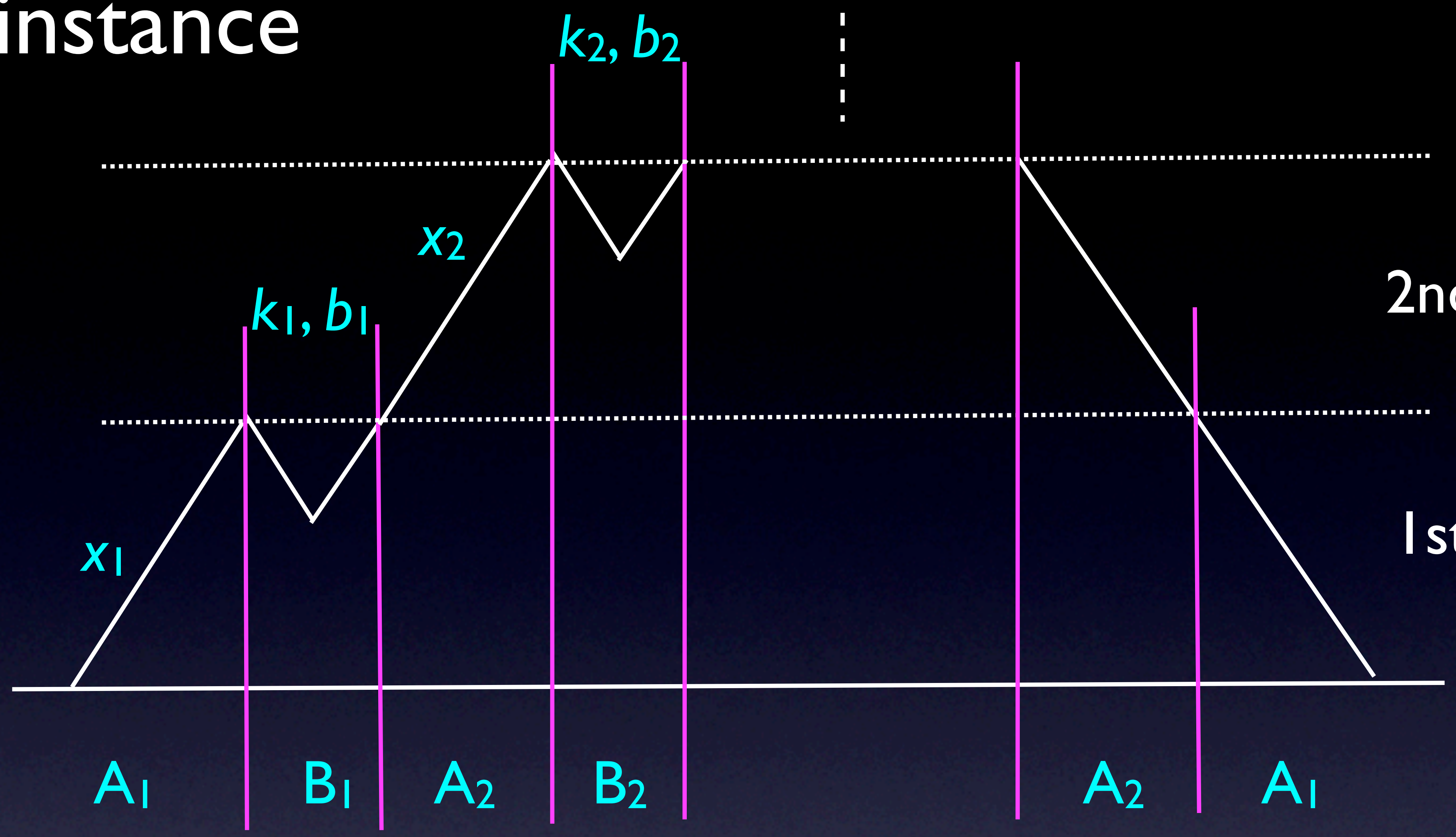
$k,\ x[1, k\text{-}1],\ b$

Proving space lower bound

$x = x_1 \, x_2 \ldots x_n$ Is $b = x_k$? $k,\ x[1, k\text{-}1],\ b$

- Need only show communication complexity is large
- However, it is $\log n + 1$: Bob sends k, b
- Need harder instance: interleave many such basic instances

Hard instance



- n independent basic units are nested at the second peak, and distributed among n pairs of players $\{A_i, B_i\}$
- Previous strategy fails because all k_i would have to be stored
- Intuition behind hardness: either A_i has to send all of x_i or the message from B_n would contain information about some k_i

Information cost trade-off

Theorem [MMN'10]

If a Alice-Bob-Alice communication protocol computes Augmented Index$_n$ with probability $1 - \varepsilon$ on the uniform distribution, either

Alice reveals $\Omega(n)$ information about x, or

Bob reveals $\Omega(\log n)$ information about k,

even when restricted to well-formed inputs.

Extension to multi-round protocols [JN'10, CCKM'10]

Implies space lower bound via a "direct sum reduction", the reason for restricting to well-formed inputs

Intuition behind proof

(2 messages [JN'10])

$x = x_1 x_2 \ldots x_n$

$\xrightarrow{M_A}$

$\xleftarrow{M_B}$

$k,\ x[1, k\text{-}1],\ b$

↓ output

Consider uniformly random X, K, let $B = X_K$ (well-formed / 0-input)

- Consider K in [n/2]. If M_A has $o(n)$ information about X, then it is nearly independent of X_L, $L > n/2$. Flipping Alice's L-th bit does not perturb M_A much.

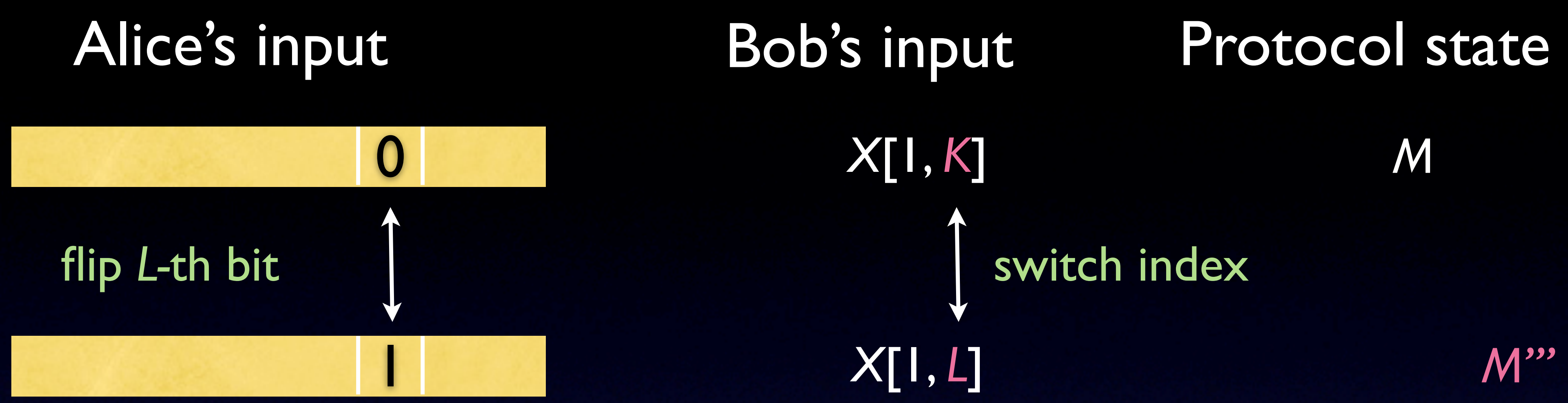
- If M_B has o(1) information about K, then M_B is nearly the same for most pairs $J \leq n/2$, $L > n/2$. Switching Bob's index from J to L does not perturb M_B much.

Consequences of Average Encoding Theorem [KNTZ'07, JRS'03]

Intuition continued...

Alice's input	Bob's input	Protocol transcript
0	$X[1, K]$	M 0-input
flip L-th bit	same index	
1	$X[1, K]$	$M' \approx M$
0	$X[1, K]$	M
same L-th bit	switch index	
0	$X[1, L]$	$M'' \approx M$
0	$X[1, K]$	M
flip L-th bit	switch index	
1	$X[1, L]$	M''' 1-input

Finally...



We have $M \approx M'$ and $M \approx M''$. Therefore, $M' \approx M''$.

Cut and paste lemma

In any (private coin) randomized protocol, the (Hellinger) distance between message transcripts on inputs (u,v) and (u',v') is the same as that between (u',v) and (u,v')

Therefore, $M \approx M'''$ and the protocol errs.

Final remarks

- Communication complexity captures a number of phenomena in information processing

  (also data structures, VLSI layout, time-space trade-offs, proof complexity, circuit depth, decision tree complexity, coding, game theory, ...)

- Recently, it has played a central role in streaming complexity of problems
- Much of this work uses the information cost approach
- Information theory may be the key to important questions regarding communication complexity