

DISPERSING HASH FUNCTIONS

RASMUS PAGH

☰ BRICS

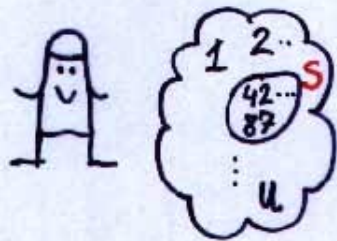
UNIVERSITY OF AARHUS

PLAN:

- "DISPERSING HASH FUNCTIONS" BY WAY OF A GAME.
- ELEMENT DISTINCTNESS: UNIVERSAL VS. DISPERSING.
- EXPLICIT CONSTRUCTIONS.
- "ALMOST PERFECT" HASH FUNCTIONS.

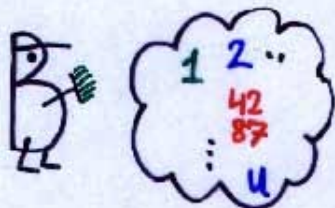
A TWO-PLAYER GAME

THE GAME "COLOURFUL" HAS ONE ROUND:



ALICE CHOOSES

$$S \subseteq U = \{1, \dots, u\}, |S| = n$$



INDEPENDENTLY, BOB

COLOURS U USING

COLOURS $\{1, \dots, r\}$.

BOB'S AIM: MAXIMIZE #COLOURS IN S
(WORST CASE OVER ALL S)

A RANDOMIZED STRATEGY

- PICK EACH COLOUR AT RANDOM.
($u \cdot \log r$ BITS).

↑ BOB CAN'T AFFORD THIS ☹

DISPERSING HASH FUNCTIONS

ASSUME $U \geq r^{1+\epsilon}$, $r \geq n$. FOR $S \subseteq U$, $|S|=n$,
A RANDOM FUNCTION $h: U \rightarrow \{1, \dots, r\}$ HAS

$$E[|h(S)|] = n - \lambda, \text{ WHERE } \lambda = \Theta(n^2/r)$$

DEFINITION

A FAMILY $\{h_i\}$, $h_i: U \rightarrow \{1, \dots, r\}$, IS **C-DISPERSING**
IF $E_i[|h_i(S)|] \geq n - c\lambda$, FOR ALL $S \subseteq U$ WITH $|S|=n$

PSEUDO-RANDOMNESS

IT IS USEFUL TO BE ABLE TO PICK AND STORE FUNCTIONS THAT "LOOK" LIKE RANDOM FUNCTIONS

MAIN CONCERNS:

- SIZE OF FAMILY
- EXPLICITNESS

PROPERTIES:

- LOW COLLISION PROBABILITY (UNIVERSALITY)
- k-WISE INDEPENDENCE
- NEAR-UNIFORMITY ON SUFFICIENTLY LARGE SETS (EXTRACTORS).

MOTIVATING EXAMPLE

ELEMENT DISTINCTNESS:

ARE $x_1, \dots, x_n \in \{1, \dots, 2^b\}$ DISTINCT?

ALGORITHM:

1. PICK $h: \{1, \dots, 2^b\} \rightarrow \{1, \dots, n^2\}$ AT RANDOM FROM A $(n/\log n)$ -DISPERSING FAMILY.
2. PUT x_1, \dots, x_n IN BUCKETS ACCORDING TO $h(x_1), \dots, h(x_n)$. (RADIX SORT)
3. FOR EACH BUCKET, PUT THE ELEMENTS IN A BINARY SEARCH TREE (STOP IF DUPLICATE)

ANALYSIS:

- $O(n/\log n)$ DISTINCT ELEMENTS IN NON-TRIVIAL BUCKETS, $\Rightarrow O(n)$ TIME
- LINEAR SPACE

DISPERSING VS. (ALMOST) UNIVERSAL FAMILIES

SIZE

UNIVERSAL: $O(r \cdot \log u)$ $\Omega(r \cdot \log_r u)$

C-DISPERSING: $O(r \cdot \log_c(u)/cn)$ $\Omega(r \cdot \log_{2r/n}(u)/cn)$
NON-CONSTRUCTIVE

EXAMPLES, CONTINUED

- ELEMENT DISTINCTNESS ALGORITHM NEEDS

$$O(\log(n^2 \cdot \log_{n/\log n}(2^b)/(n^2/\log n))) = O(\log b)$$

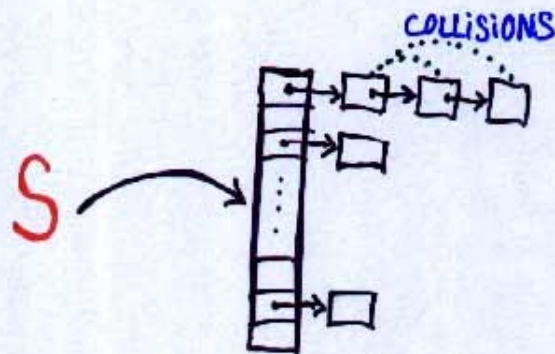
RANDOM BITS (RATHER THAN $O(\log n + \log b)$).

- BOB NEEDS, FOR $c=1+\epsilon$, $O(\log \log u)$
RANDOM BITS.

EXPLICIT CONSTRUCTION I

OBSERVATION (FREDMAN ET AL):

$$|h(S)| \geq n - \# \text{ COLLIDING PAIRS}$$

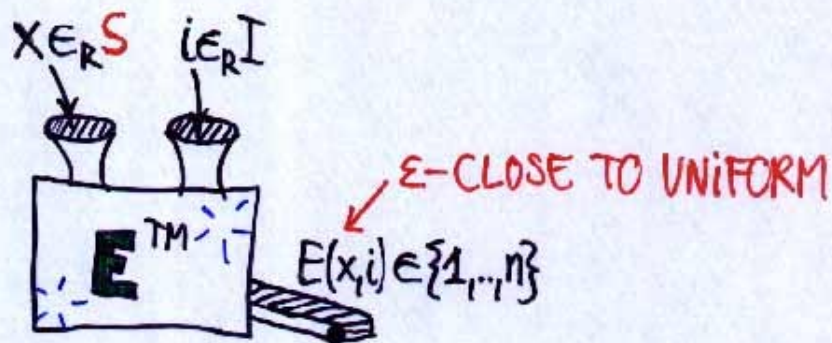


(ALMOST) UNIVERSAL FAMILIES
ARE $O(1)$ -DISPERSING

EXPLICIT CONSTRUCTION II_A

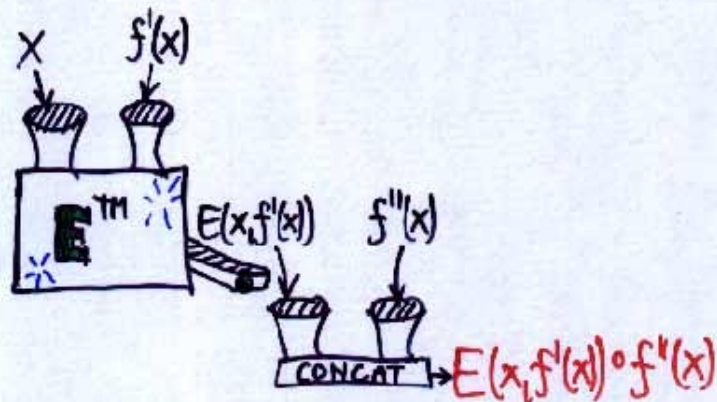
EXTRACTOR (SIMPLIFIED):

FOR ANY $S \subseteq U$ WITH $|S|=n$



EXPLICIT CONSTRUCTION II_B

- SELECT f' FROM A STRONGLY UNIVERSAL FAMILY, $f': U \rightarrow I$.
- SELECT $f'': U \rightarrow \{1, \dots, r/n\}$ FROM A UNIVERSAL FAMILY.



ANALYSIS:

LOOK AT COLLISIONS OUTSIDE A SET OF "BAD" POINTS, TO WHICH FEW ELEMENTS MAP.

A DIFFERENT VIEW OF EXTRACTORS

EXTRACTOR WITH OPTIMAL $\log |I|$
SEED LENGTH



NON-TRIVIAL

$O(1)$ -DISPERSING FAMILY WITH
OPTIMAL SAMPLE COMPLEXITY

WHEN DESIGNING EXPLICIT
EXTRACTORS, ONE MAY W.L.O.G.
CONSIDER COLOURFUL STRATEGIES

DETOUR: "ALMOST PERFECT" HASHING

- MINIMAL PERFECT HASH FUNCTION:

$$h: U \rightarrow \{1, \dots, n\}$$

$$|h(S)| = |S| = n$$

PROGRAM SIZE: $\Omega(n)$ BITS

- "ALMOST PERFECT" HASH FUNCTION:

$$h: U \rightarrow \{1, \dots, n\}$$

$$|h(S)| \geq \frac{2}{3}|S| = \frac{2}{3}n$$

PROGRAM SIZE: $\Omega(n)$ BITS

CONCLUSION

DISPERSING HASH FUNCTIONS

- + SEVERAL (MANY!) APPLICATIONS
- + POTENTIALLY LOW SAMPLE COMPLEXITY
- ⊖ EXPLICIT CONSTRUCTION PROBABLY DIFFICULT (AT LEAST FOR $c = O(1)$).
- ⊖ EVER AS FAST AS UNIVERSAL HASH FUNCTIONS?

• •
THAT'S ALL