OPPA European Social Fund
Prague & EU: We invest in your future.

DCGI

KATEDRA POČÍTAČOVÉ GRAFIKY A INTERAKCE

MODERN ALGORITHMS (not only in computational geometry)

PETR FELKEL

FEL CTU PRAGUE

felkel@fel.cvut.cz

http://service.felk.cvut.cz/courses/X36VGE

Based on [Kolingerova], [Brönnimann], and [Muthukrishnan]

Version from 22.12.2011

Modern algorithms

1. Computational geometry today

- Popular: beauty as discipline, wide applicability
- Started in 2D with linear objects (points, lines,…), now 3D and nD, hyperplanes, curved objects,…
- Shift from purely mathematical approach and asymptotical optimality ignoring singular cases
- to practical algorithms, simpler data structures and robustness => algorithms and data structures provable efficient in realistic situations (application dependent)

Space efficient algorithms - practical advantages

- Allow for processing larger data sets
  - Algorithms with separate input and output need space for 2n points to store – O(n) extra space
  - Space efficient algs – n points + O(1) or O(log n) space
- Greater locality of reference
  - Practical for modern HW with memory hierarchies (e.g., main RAM – ram on chip – registers, caches, disk latency, network latency)
- Less prone to failure
  - no allocation of large amounts of memory, which can fail at run time
  - good for mission critical applications
- Less memory => faster program

2. In-place / in situ algorithms

Space efficient algorithms

- output is in the same location as the input and
- need only a small amount of additionaly memory
  - in-place – O(1) extra storage
  - *in situ* – O(*log n*) extra storage

In-place sorting

- In array – continuous block in memory
  - Select sort, insert sort … yes, $O(1)$ memory, $O(n^2)$ time
  - Heapsort – yes, $O(1)$ additional memory
  - Quicksort – yes, $O(\log n)$ additional memory for recursion
  - Mergesort – not in-place
- In list – linked lists in dynamical memory
  - n[th] element in $O(n)$ time
  - Mergesort – $O(\log n)$ time, $O(\log n)$ additional memory

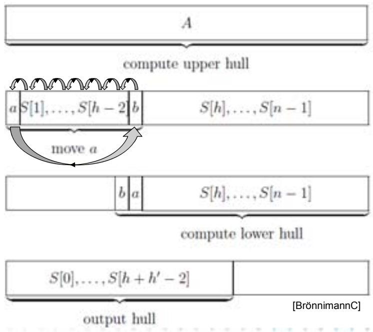
Graham in-place algorithm

Graham-InPlaceHull(S, n)
Input: S – an array of length n with points in plane
Output: Convex Hull in clockwise order

```
h ← Graham-InPlace-Scan(S, n, 1)   // CW upper hull
for i ← 0 . . . h − 2 do
    swap S[i] ↔ S[i + 1]            // bubble a to the right
h' ← Graham-InPlace-Scan(S + h − 2, n − h + 2, −1)  // lower hull
return h + h' − 2
```

Stack at the beginning of the array

Graham in-place algorithm



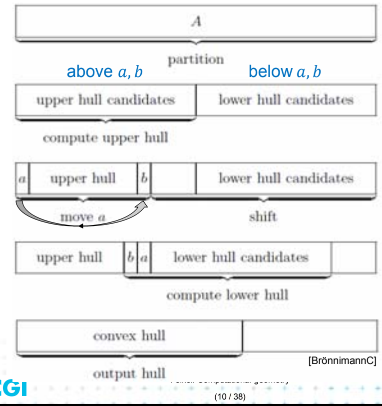
[BrönnimannC]

Graham in-place algorithm

Graham-InPlaceScan(S, n, d)
Input: $S[0, n-1]$ – array of length n with points in plane, $d = \pm 1$ direction
Output: Convex Hull in clockwise order

```
InPlace-Sort(S, n, d)   // d = 1 CW for upper hull, d = −1 CCW for LH
h ← 1
for i ← 1 ... n − 1 do
    while h ≥ 2 and not right turn( S[h − 2], S[h − 1], S[i] ) do
        h ← h − 1      // pop top element from the stack
    swap S[i] ↔ S[h]
    h ← h + 1
return h
```


Optimized Graham in-place algorithm



[BrönnimannC]

3. Data stream algorithms [Indyk]

- Data stream = a massive sequence of data
  - Too large to store (on disk, memory, cache,…)
- Examples
  - Network traffic
  - Database transactions
  - Sensor networks
  - Satelite data feeds
  - …
- Aproaches
  - Ignore It
  - Develop algorithms for dealing with such data

Motivation example [Muthukrishnan]

- Paul presents numbers $x = \{1..n\}$ in random order, one number missing
- Carole must determine the missing number but has only $O(\log n)$ bits of memory

Any idea?

- Compute the sum of the numbers and subtracts the incoming numbers one by one.

$$missing\ number = \frac{n(n+1)}{2} - \sum_{i<n} x[i]$$

- The missing number remains

Motivation example [Muthukrishnan]

- And two missing numbers?
- Store sum of numbers s and sum of squares s'

$$i + j = \frac{n(n+1)}{2} - s$$

$$i^2 + j^2 = \frac{n(n+1)(2n+1)}{6} - s'$$

Basic data stream model [Indyk]

- Single pass over the data: $a_1, a_2, \dots, a_n$
  - Typically n is known
- Bounded storage (typically n^α or $\log^c n$ or only c)
  - Units of storage: bits, words, or elements (such as points, nodes/edges, …)
  - Impossible to store the data
- Fast processing time per element
  - Randomness is OK (in fact, almost necessary)

Data stream models classification

- Input stream $a_1, a_2, \dots, an$
  - arrives sequentially, item by item
  - describes an underlying signal A, a 1D function A: $[1..N]$ $->$ R
- Models differ on how a_i's describe the signal A (in decreasing order of generality):
  a) Time series model - a_i equals $A[i]$, in increasing i
  b) Cash register model - a_i are increments to $A[j]$, $Ii > 0$
  c) Turnstyle model - a_i are updates to $A[j]$, $U_i \in R$

a) Time series model (*Časová řada*)

- Stream elements a_i are equal to $A[i]$ (samples of the signal)
- a_i's appear in increasing order of i
- Applications
  - Observation of the traffic on IP address each 5 minutes
  - NASDAQ volume of trades per minute

b) Cash register model (*registrační pokladna*)

- a_i are increments to $A[j]'s$
- Stream elements $a_i = (j, Ii)$, $I_i \geq 0$ to mean

$$A_i[j] = A_{i\,\_\,1}[j] + Ii$$

  where
  - A_i is the state of the signal after seeing i-th item
  - multiple a_i can increment given $A[j]$ over time
- A most popular data stream model
  - IP addresses accessing web server
  - Source IP addresses sending packets over a link
  - access many times, send many packets,…

c) Turnstile model (*turniket*)

- a_i are updates to $A[j]'s$
- Stream elements $a_i = (j, U_i)$, $I_i \in R$ to mean

$$A_i[j] = Ai_{\,\_\,1}[j] + U_i$$

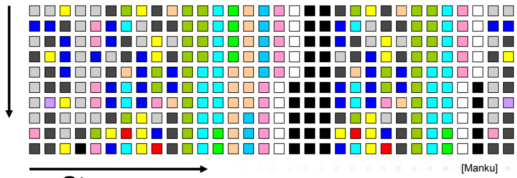
  where
  - A_i is the state of the signal after seeing i-th item
  - U_i may be positive or negative
  - multiple a_i can updategiven $A[j]$ over time
- A most general data stream model
  - Passengers in NY subway arriving and departing
  - Hard to get reasonable solution in this model

c) Turnstile model variants (for completness)

- strict turnstile model – $A_i[j] \geq 0$ for all i
  - People can only exit via the turnstile they entered in
  - Databases – delete only a record you inserted
  - Storage – you can take items only if they are there
- non-strict turnstile model – $A_i[j] < 0$ for some i
  - Difference between two cash register streams

Examples: Iceberg queries

- Identify all elements whose current frequency exceeds support threshold s = 0.1%.



Stream

[Manku]

Ex: Iceberg queries – a) ordinary solution

The ordinary solution in two passes

1. Pass – identify frequencies
   - a set of counters is maintained. Each incoming item is hashed onto a counter, which is incremented.
   - These counters are then compressed into a bitmap, with a 1 denoting a large counter value.
2. Pass – count exact values
   - exact frequencies for only those elements are maintained which hash to a value whose corresponding bitmap value is 1

- Hard to modify for datastream – unknown frequencies after only 1st pass

Ex: Iceberg queries – problem definition

- Input: threshold $s \in (0,1)$, error $\varepsilon \in (0,1)$, length N
- Output: list of items and frequencies $\quad \epsilon \ll s$
- Guarantees:
  - No item omitted (reported all items with frequency > sN)
  - No item added (no item with frequency < $(s - \epsilon)N$)
  - Estimated frequencies not less than ϵN of the true frequencies
- Ex: $s = 0.1\%, \epsilon = 0.01\%$ $-\ \epsilon$ about $\frac{1}{10}$ to $\frac{1}{20}$ of s
  - All element with freq. > 0.1% will output
  - None of element with freq. < 0.09% will output
  - Some elements between 0.09% and 0.1% will output

Ex: Iceberg queries – b) sticky sampling

- Probabilistic algorithm, given threshold s, error ϵ and probability of failure δ
  - Data structure S of entries (e, f),
    e element, f estimated frequency,
    r sampling rate, sampling probability $\frac{1}{r}$
- $S \leftarrow \emptyset, r \leftarrow 1$
- If $e \in s$ then $(e, f{+}{+})$
  else insert (e, f) into S with probability $\frac{1}{r}$
- S sweeps along the stream as a magnet, attracting all elements which already have an entry in S

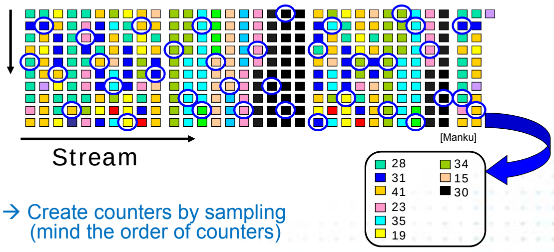
Ex: Iceberg queries – b) sticky sampling

- r changes over the stream, $t = \frac{1}{\epsilon}\log\left(\frac{1}{s\delta}\right), |S| < 2t$
  - $2t$ elements $r = 1$
  - next $2t$ elements $r = 2$
  - next $4t$ elements $r = 4$ ...
- whenever r changes, we update S
  - For each entry (e, f) in S
    - toss a coin until successful (head)
    - if not successful (tail), decrement f
    - if f becomes 0, remove entry (e, f) from S
- Output: list of items with threshold s
  i.e. all entries in S where $f \geq (s - \epsilon)N$

Ex: Iceberg queries – b) sticky sampling

- Space complexity is independent on N
- For
  - support threshold $s = 0.1\%$,
  - error $\epsilon = 0.01\%$,
  - and probability of failure $\delta = 1\%$
- Sticky sampling computes results
  - with $(1-\delta) = 99\%$ probability
  - using at most 2t = 80 000 entries
  - $t = \frac{1}{\epsilon}\log\left(\frac{1}{s\delta}\right) = 40\,000, |S| < 2t$

Ex: Iceberg queries – b) sticky sampling

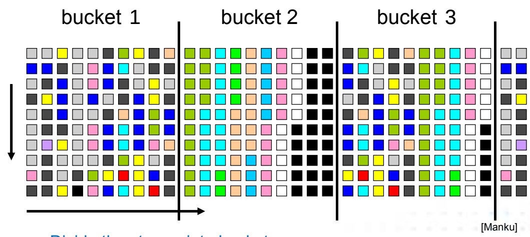


→ Create counters by sampling (mind the order of counters)

Ex: Iceberg queries – c) lossy counting

- Deterministic algorithm
- Stream conceptually divided into buckets
  - With $w = \lceil 1/\varepsilon \rceil$ items each
  - Numbered from 1, current bucket id is $b_{current}$
- Data structure D of entries (e, f, Δ),
  - e element,
  - f estimated frequency,
  - Δ maximum possible error of f (max number of occurences in previous buckets)
- At most $\frac{1}{\epsilon}\log\left(\frac{1}{\varepsilon N}\right)$ entries

Ex: Iceberg queries – c) lossy counting



- Divide the stream into buckets
- Keep exact counters for items in the buckets
- Prune entries at bucket boundaries (remove entries for which $f + \Delta \leq b_{current}$)

Ex: Iceberg queries – c) lossy counting alg.

- $D \leftarrow \emptyset$
- New element e
  - If $e \in D$ then increment its f
  - If $e \notin D$ then
    - Create a new entry $(e, 1, bcurrent - 1)$
    - If on the bucket border, i.e., $N \bmod w = 0$ then delete entries with $f + \Delta \leq b_{current}$
    - i.e., with zero or one occurence in previous buckets
  - New $\Delta = b_{current} - 1$ is maximum number of times e could have occured in the first $b_{current} - 1$ buckets
- Output: list of items with threshold s i.e. all entries in S where $f \geq (s - \epsilon)N$

Comparison of sticky and lossy sampling

- Sticky sampling performs worse
  - Tendency to remember every unique element
  - The worst case is for sequence without duplicates
- Lossy counting
  - Is good in pruning low frequency elements quickly
  - Worst case for pathological sequence which never occurs in reality

Number of mutually different entries

- Input: stream $a_1, a_2, \dots, an$, with repeated entries
- Output: Estimate of number of different entries
- Appl: # of different transactions in one day
- Precise deterministic algorithm:
  - Array $b[1..U]$, U = max number of different entries
  - Init by $b[i] = 0$ for all i, counter $c = 0$
  - For each a_i
    - if $b[a_i] = 0$ then $\mathrm{inc}(c)$, $b[i] = 1$
  - Return c as number of different entries in $b[]$
  - $O(1)$ update and query times, $O(U)$ memory

Number of mutually different entries

- Approximate algorithm
  - Array $b[1..\log U]$, U = max number of different entries
  - Init by $b[i] = 0$ for all i, counter $c = 0$
  - Hash function $h: \{1..U\} \to \{0..\log U\}$
  - For each a_i
    Set $b[h(ai)]$ = 1
  - Extract probable number of different entries from $b[]$

Sublinear time example

- Given mutually different numbers $a_1, a_2, \dots, an$
- Determine number in upper half of values
- Alg: select k numbers equally randomly
  - Compute their maximum
  - Return it as solution
- Probability of wrong answer = probability of all selected numbers are from the lower half = $\left(\frac{1}{2}\right)^k$
- For error δ take $\log\frac{1}{\delta}$ samples
- Not useful for MIN, MAX selection

4. Randomized algorithms

Motivation

- Array of elements, half of char ”a”, half of char ”b”
- Find ”a”
- Deterministic alg: $n/2$ steps of sequential search (when all ”b” are first)
- Randomized:
  - Try random indices
  - Probability of finding ”a” soon is high regardless of the order of characters in the array
    (Las Vegas algorithm)

Randomized algorithms

- May be simpler even if the same worst time
- We do not know a deterministic version (prime numbers)
- Deterministic algorithm does not exist
- Randomization can improve the average running time (with the same worst case time), while the worst time depends on our luck – not on the data distribution

Randomized algorithms

a) Incremental algorithms
   - Linear programming – see seminars
   - Convex hulls
   - Intersections, space subdivisions

b) Divide and conquer
   - Random sampling
   - Nearest neighbors, trapezoidal subdivisions

Random sampling

- Hierarchical data structures
- Sublinear algorithms
- Randomized quicksort
- Approximate solutions on random samples

Another classification

- Monte Carlo
  - We always get an answer, often not correct
  - Fast solution with risk of an error
  - It is not possible to determine, if the answer is correct
    → run multiple times and compare the results
  - Output can be understand as a random variable
  - Example: prime number test
    - Task: Find $a \in \left\langle 2, \frac{n}{2} \right\rangle$ such as n is divisible by a
    - Algorithm: Sample 10 numbers from the given interval, answer
- Las Vegas

Las Vegas algorithms

Las Vegas
- We always get a correct answer
- The run time is random
- Sometimes fails –> perform restart
- Example: Randomized quicksort
  - No median necessary
  - Simpler algorithm
  - Independent on data distribution
  - Return a correct result
  - The result will be ready to uphold with a high probability
  - $\theta(n \log n)$ čas na lib. Výstup s velkou pravděpodobností
  - Bad luck – we select the smallest element -> Selection sort

Randomized quicksort

RQS = Randomized Quicksort
Input: sequence of data elements $a_1, a_2, \dots, a_n \in S$
Output: sorted set S

1. Step 1: choose $i \in \langle 1, n \rangle$ in random
2. Step 2: Let A is a multiset $\{a_1, a_2, \dots, a_n\}$
   - if $n = 1$ then output(S)
   - else – create three subsets of $S_<, S_=, S_>$
     $S_< = \{b\ z\ A: b < a_i\}$
     $S_= = \{b\ z\ A: b = a_i\}$
     $S_> = \{b\ z\ A: b > a_i\}$
3. Step 3: Sort $S_<$ *and* $S_>$
4. *Výstup: RQS(S<), S=, RQS(S>)*

Conclusion

- Randomized algs. are often experimental
- We would not get perfect results, but nicely good
- We use randomized algorithm if we do not know how proceed

References

[Kolingerová] Nové směry v algoritmizaci a výpočetní geometrii (1 a 2), přednáška z předmětu Aplikovaná výpočetní geometrie, MFF UK, 2008

[Brönnimann] Hervé Brönnimann. Towards Space-Efficient Geometric Algorithms, Polytechnic university, Brooklyn, NY,USA, ICCSA04, Italy, 2004

[BrönnimannC] Hervé Brönnimann, et al. 2002. In-Place Planar Convex Hull Algorithms. In *Proceedings of the 5th Latin American Symposium on Theoretical Informatics* (LATIN '02), Sergio Rajsbaum (Ed.). Springer-Verlag, London, UK, UK, 494-507. http://dl.acm.org/citation.cfm?id=690520

[Indyk] Piotr Indyk. 6.895: Sketching, Streaming and Sub-linear Space Algorithms, MIT course

[Muthukrishnan] Data streams: Algorithms and applications, ("adorisms" in Google)

[Mulmuley] Ketan Mulmuley. Computational Geometry. An Introduction Through Randomized Algorithms. Prentice Hall, NJ,1994

[Manku] G.S. Manku, R. Motwani. Approximate Frequency Counts over Data Streams, Proceedings of the 28th VLDB Conference, Hong Kong, China, 2002. http://www.vldb.org/conf/2002/S10P03.pdf

OPPA European Social Fund
Prague & EU: We invest in your future.
