MADALGO
CENTER FOR MASSIVE DATA ALGORITHMICS


Jelani Nelson
MIT

Danmarks Grundforskningsfond
Danish National Research Foundation

An Optimal Algorithm for the Distinct Elements Problem

Problem and Results

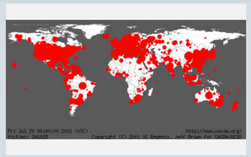
Sequence of integers

1 17 2 4 17 **9** 2 **5** 1 1 4 **6**

- One pass over a stream of integers each between 1 and n
- Query() – Output the number of distinct integers seen thus far
- Goals – Use little memory, and process each integer quickly

Applications

track spread of Code Red worm



network intrusion detection

database query optimization

	Memory	Update Time
Flajolet, Martin 1983	O(log n)	–
Alon, Matias, Szegedy 1996	*O*(log n)	*O*(log n)
Gibbons, Tirthapura 2001	$O((\log n)/\varepsilon^2)$	$O(1/\varepsilon^2)$
Bar-Yossef, Jayram, Kumar, Sivakumar, Trevisan 2002	$O((\log n)/\varepsilon^2)$	O(log(1/ε))
Bar-Yossef, Jayram, Kumar, Sivakumar, Trevisan 2002	$O((\log\log n + \log(1/\varepsilon))/\varepsilon^2 + \log n)$	$O(1/\varepsilon^2)$
Durand, Flajolet 2003	$O((\log\log n)/\varepsilon^2 + \log n)$	–
Kane, Nelson, Woodruff 2010	$O(1/\varepsilon^2 + \log n)$	O(1)

Algorithm Ideas

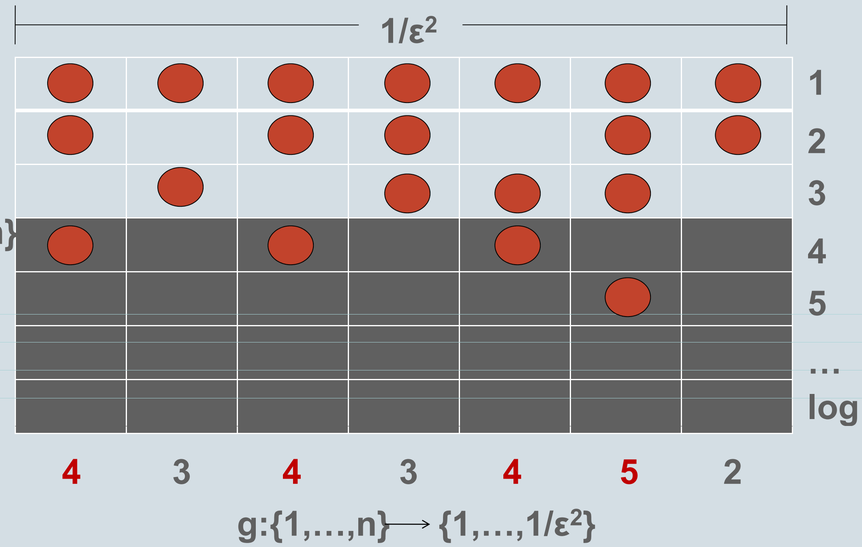
Balls-and-bins approach

Inspired by [Bar-Yossef et al. 2002]

- Subsample the stream at geometrically decreasing rates
- Perform balls and bins at each level

$h:\{1,\ldots,n\} \longrightarrow \{1,\ldots,\log n\}$

$\Pr[h(i) = j] = 1/2^j$



- When i appears in stream, put a ball in cell (g(i), h(i))
- For each column, store the largest row containing a ball
- Identify the largest row j which is at least half full, and count the number of columns with at least j written. Base estimate on this count.

Some implementation details

- Store column array in variable length array data structure of [Blandford, Belloch 2008] (bitpacked)
- In column array, store offset from topmost active row and not absolute index.
- Use deamortization of global rebuilding for worst-case time [Overmars 1983]
- Use high-performance hash functions of [Siegel 1989] and [Pagh, Pagh 2008]

References

[1] Alon, Matias, Szegedy. *On the Space Complexity of Approximating the Frequency Moments*. STOC 1996

[2] Bar-Yossef, Jayram, Kumar, Sivakumar, Trevisan. *Counting* distinct elements in a data stream. RANDOM 2002

[3] Blandford, Blelloch. *Compact dictionaries for variable-length keys and data with applications.* ACM Transactions on Algorithms 2008

[4] Flajolet, Martin. *Probabilistic Counting.* FOCS 1983

[5] Gibbons, Tirthapura. *Estimating simple functions on the union of data streams*. SPAA 2001

[6] Kane, Nelson, Woodruff. *An Optimal Algorithm for the Distinct Elements Problem*. PODS 2010

[7] Overmars. *The Design of Dynamic Data Structures.* Springer 1983

[8] Pagh, Pagh. *Uniform Hashing in Constant Time and Optimal Space.* SICOMP 2008

[9] Siegel. *On Universal Classes of Uniformly Random Constant-Time Hash Functions.* SICOMP 2004