The loglog algorithm of Durand and Flajolet

Keith Briggs

Keith.Briggs@bt.com

research.btexact.com/teralab/keithbriggs.html

2003 June 09

The preprint

Loglog counting of large cardinalities, by Marianne Durand and Philippe Flajolet. In the *Engineering and Applications Track* of the *11th Annual European Symposium on Algorithms* (ESA03). To be published by Springer, Lecture Notes in Computer Science. April 1, 2003.

`http://algo.inria.fr/flajolet/Publications/DuFl03.ps.gz`

The algorithm

$$\rho(b_1 b_2 b_3 \dots) \equiv \mathrm{argmin}_k \{k \text{ s.t. } b_k = 1\}$$

choose paramater k (typically 10-12)

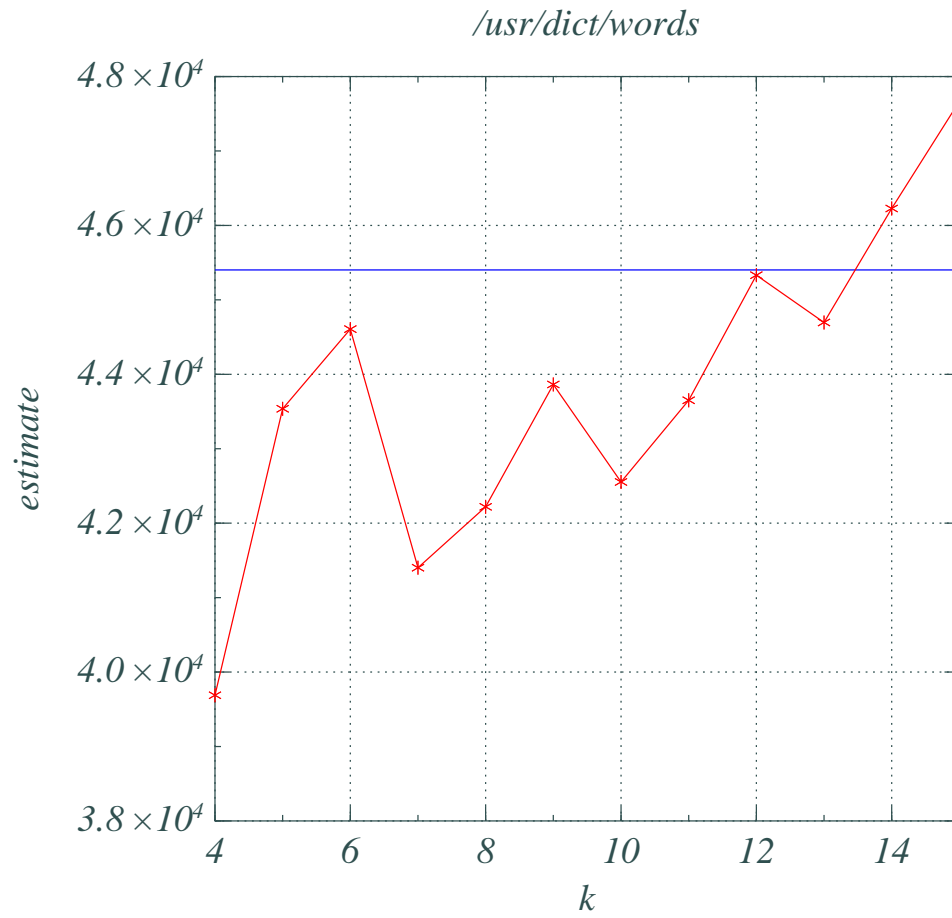$m = 2^k$, buckets $M_1, M_2, M_3, \dots, M_m$ initialized to 0

$h =$ a hash function (e.g. 32 bits)
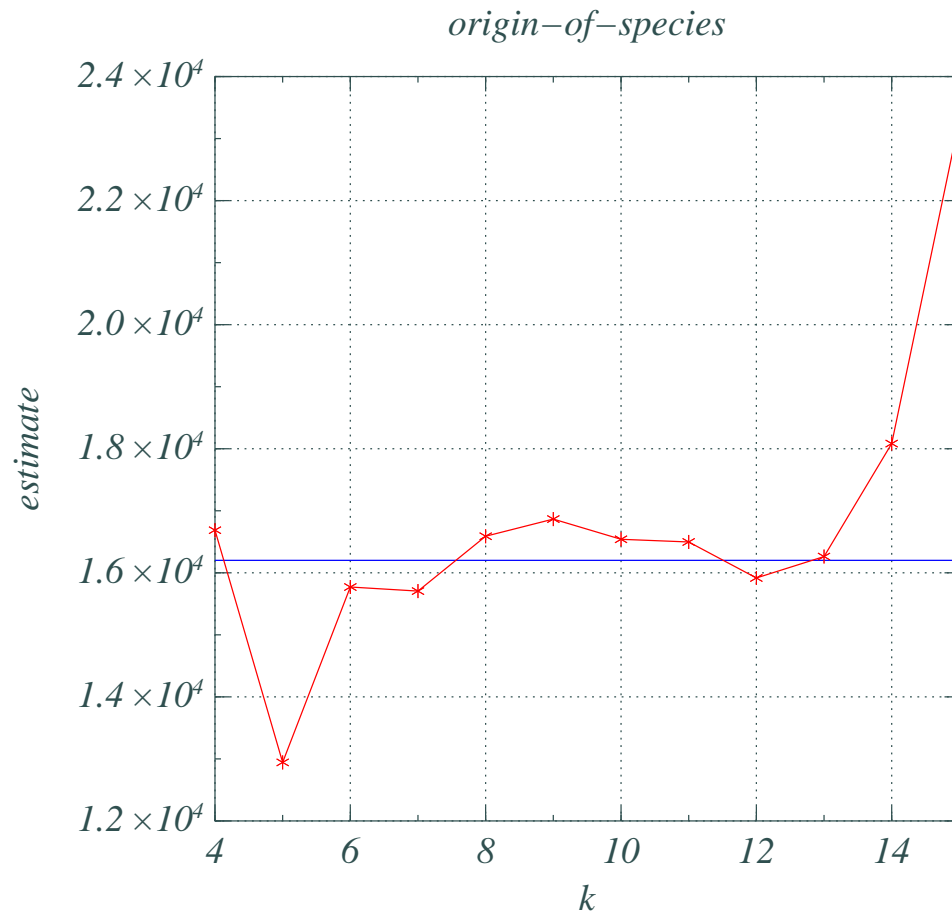
For each word x in the file:

- $y = h(x)$
- j=value of first k bits of y
- l=value of last (hash size $-k$) bits of y
- set M_j to the maximum of M_j and $\rho(l)$

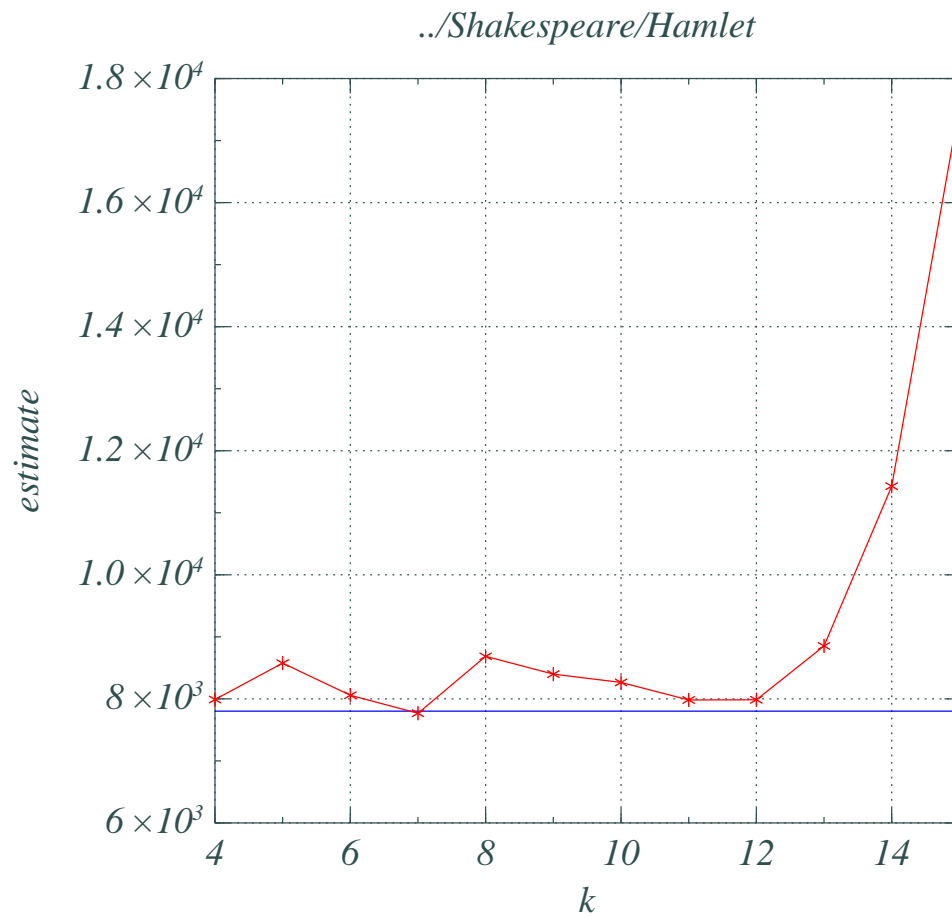size estimate is $m \left[\Gamma(-1/m) \frac{2^{-1/m} - 1}{\log 2} \right]^{-m} 2^{(\sum_j M_j)/m - 1}$
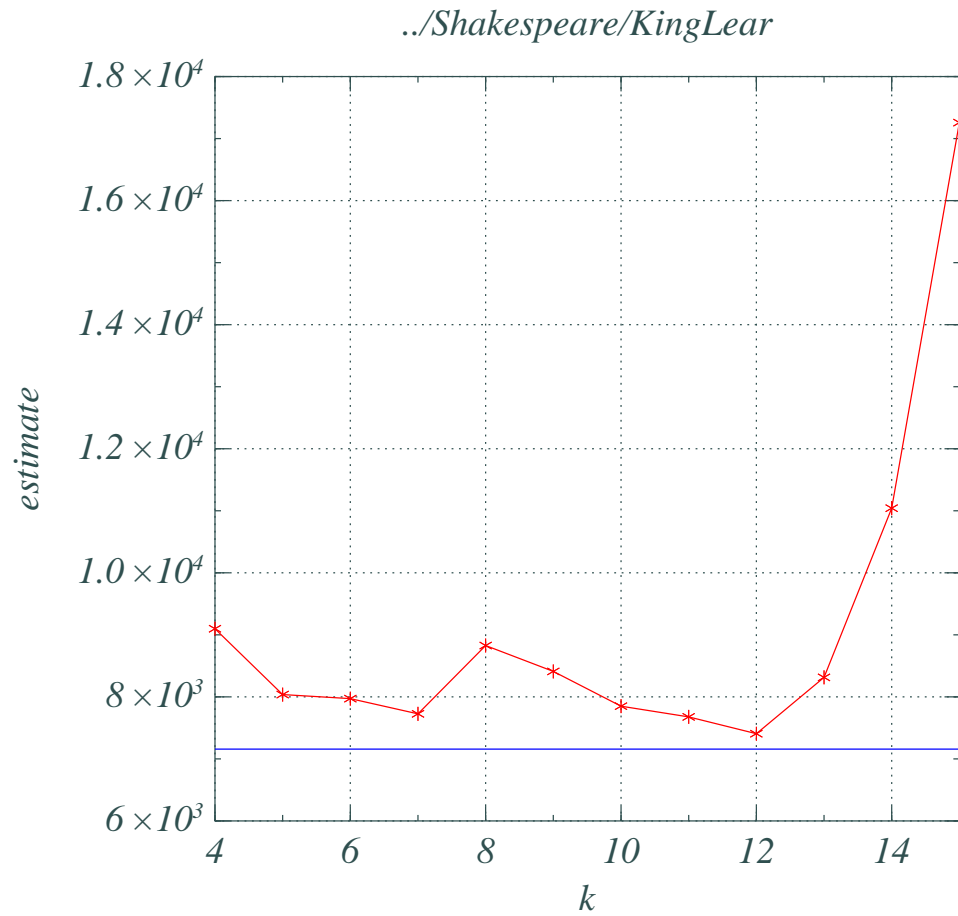
Results 1

Results 2

origin−of−species

Results 3

../Shakespeare/Hamlet

Results 4

../Shakespeare/KingLear

Results 5

../Shakespeare/JuliusCaesar