Setting

- Association Rule Mining.
- Issue: what kind of synthetic database should be used in performance analysis.

CLAIM: databases generated by QUEST may not be entirely suitable for this purpose.

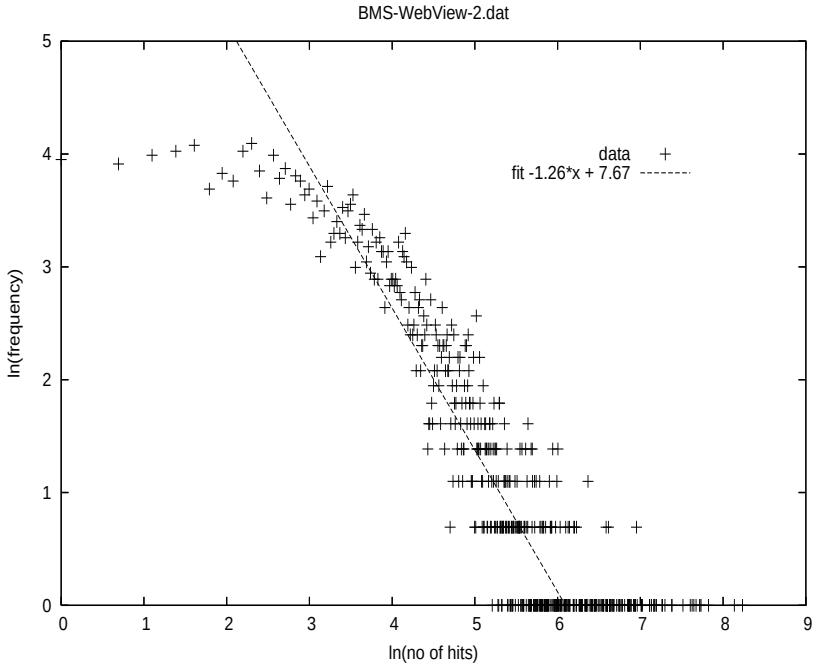
In the past, people have found differences (Zhang *et al.* (2001)) or argued (Brin *et al.* (1997)) that QUEST databases are relative "simple".

We focus on deeper differences that, we contend, depend on structural properties of the probabilistic model underlying Agrawal and Srikant's generator.

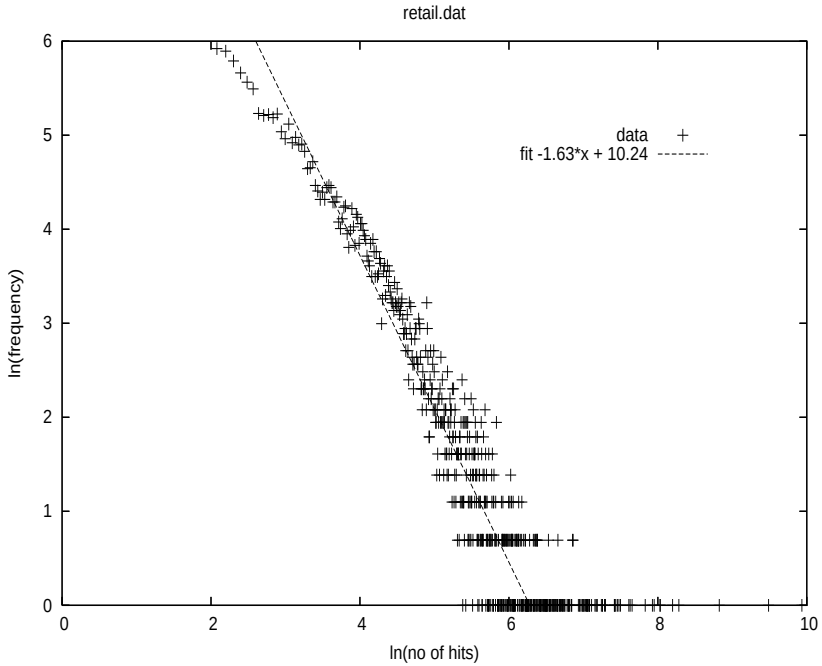
Item occurrence distribution: real data

Plotted on log-log scale.

Datasets from `http://fimi.cs.helsinki.fi/data/`.



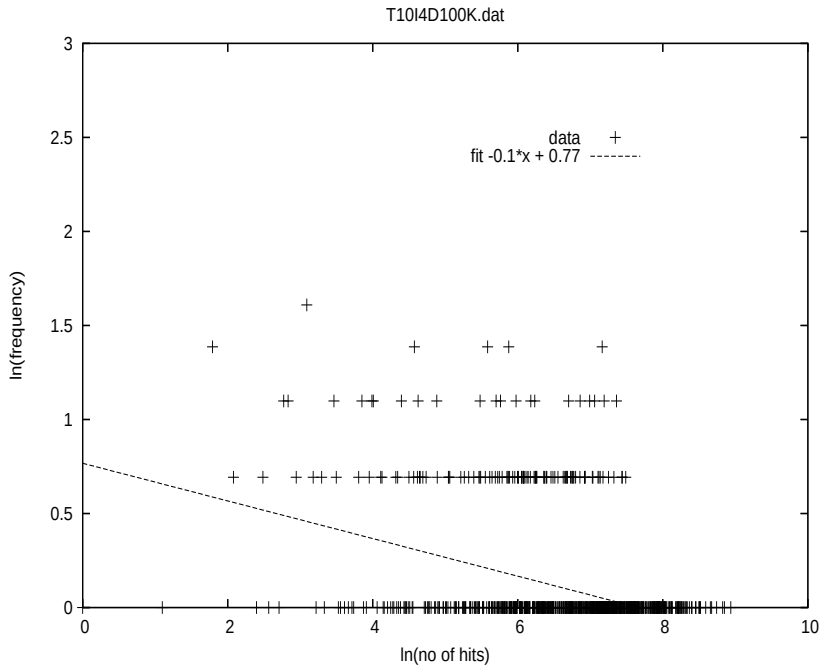
$h = 77,512, n = 3,340$



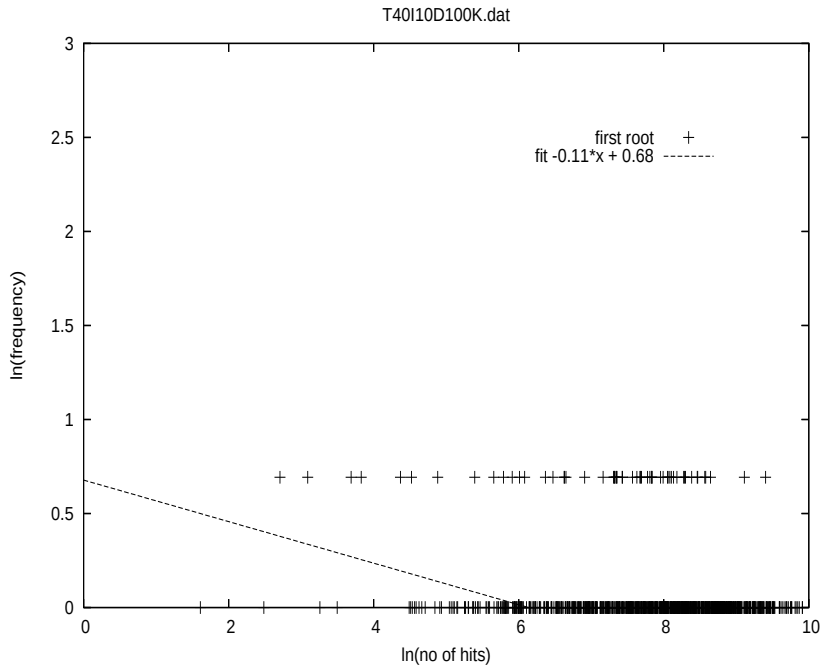
$h = 88,162, n = 16,470$

CLAIM: The item occurrence distribution shows a *power-law*. It decays like r^{-c}.

Item occurrence distribution: synthetic data



T10I4D100K.dat



T40I10D100K.dat

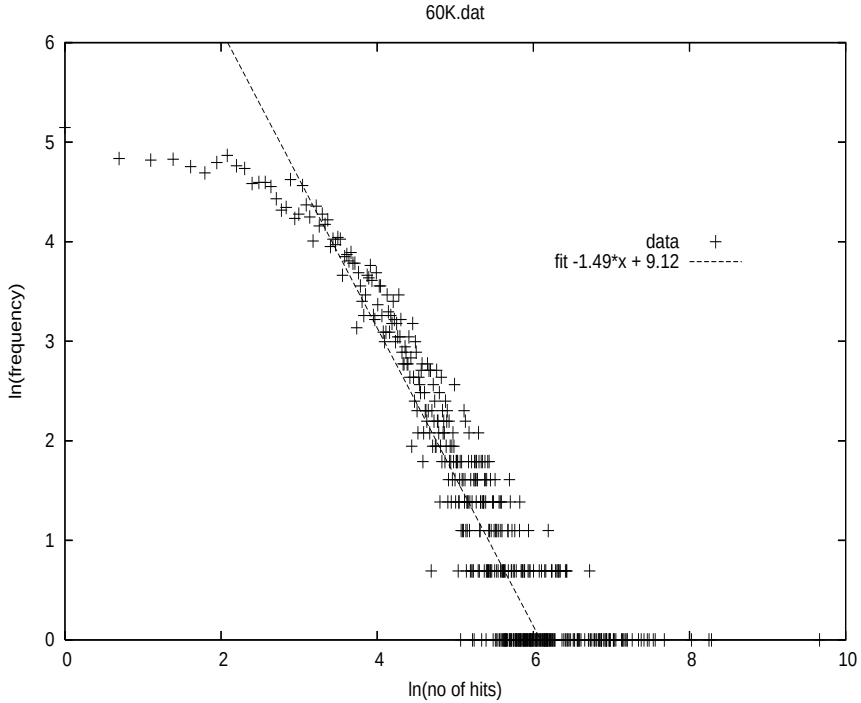
CLAIM: The item occurrence distribution in QUEST databases decays exponentially (i.e. like c^{-r}).

The Co-Zi generator[a]

http://www.csc.liv.ac.uk/~michele/soft.html

Input: μ, σ, α, P, h
Output: A database $\mathcal{D}$ with h transactions

Initially the database contains n_0 items and
    e_0 transactions chosen arbitrarily
for $t = 1$ to $h - e_0$
  select the size x of the next transaction
    as the absolute value of a normally distributed
    number with mean μ and deviation σ
  if $(x > 0)$
    with probability α add a NEW transaction to $\mathcal{D}$
    otherwise add an OLD transaction to $\mathcal{D}$.



[a] For lack of a better name!