Data analysis for benchmarking combinatorial solvers with R.

Stephan Gocht

April 29, 2022

Tools and Libraries

- R — programming language for data science
- RStudio — IDE for R (`https://www.rstudio.com/`)
- RNotebook — Feature of RStudio allowing to interleave text, R and resulting plots / tables
- tidyverse (`https://www.tidyverse.org/`)
  - ggplot2 — library for plotting
  - dplyr — library for data manipulation

What to Measure and Store

What to Measure and Store

- running time (wall clock system-time and user-time)
- memory usage
- how often did X happen
- proof size (number of steps file size)
- hash of input file
- computer name
- current date and time
- return code
- solver name
- instance name
- path of instance file
- instance category
- instance scaling parameter
- solver parameters
- used timeout
- ...

Storing and Importing Data — Best Practice

- use standard machine readable formats (CSV, SQLlite)
- following standards saves manual work for import
  (CSV = **comma** separated file, no space after commas, etc.)
- include table headers (avoid ' ', '-', and special characters)
- compress files individually ('.csv.gz' instead of '.zip')
- store information about used units (ms / s, byte / KB / MB, etc.)
- if CSV import takes long, use R's own format for local use
- less files are easier to handle
  (usually first normal form sufficient, for our use case)

Data Transformation

my favourite operations:

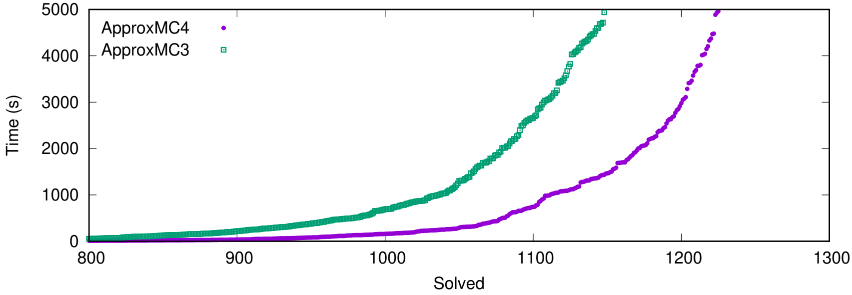
- `filter` — filters data
- `mutate` — change or create columns
- `arrange` — sort data
- `group_by` — group data for use with summarise
- `summarise` — collect statistics of grouped data
- `join` — combine columns of two tables
- `bind_rows` — append rows of to tables
- `pivot_wider` — create columns by removing rows
- `pivot_longer` — create rows by removing columns
- `distinct` — remove duplicate rows

cheat sheets: `https://www.rstudio.com/resources/cheatsheets/`

Tables

- easy way to show small number of data points
- usually used for accumulated data
  (number of time outs, speed-up, min, max, median, mean, quantiles)
- create table in R, export to latex, don't touch afterwards
  `print(xtable(dataframe, type = "latex"), file = "tabel.tex")`

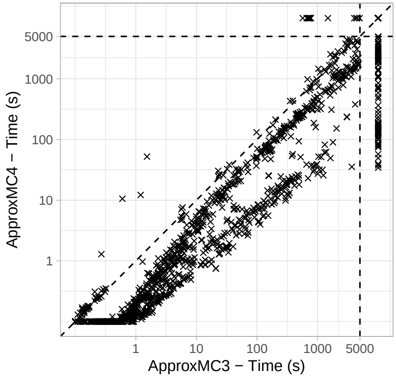
Cactus Plot / Cumulative Plot



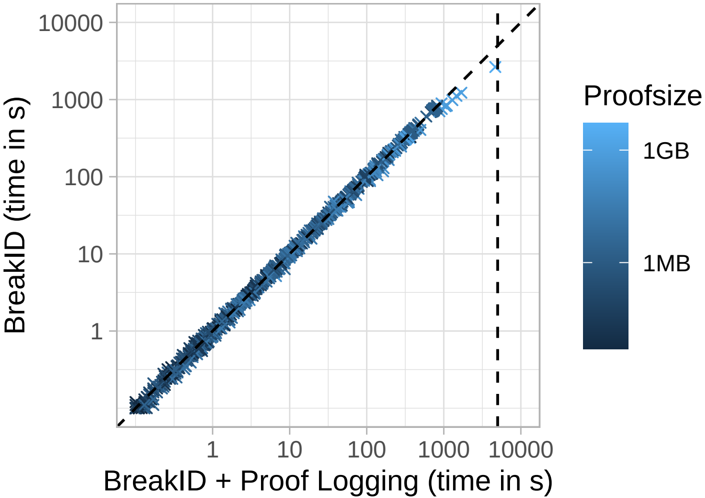
- used to compare many different algorithms / solvers
- shows number of solved instances at different timeouts
- pitfall: suggests that one solver is always better than other (usually not true)

Scatter Plot

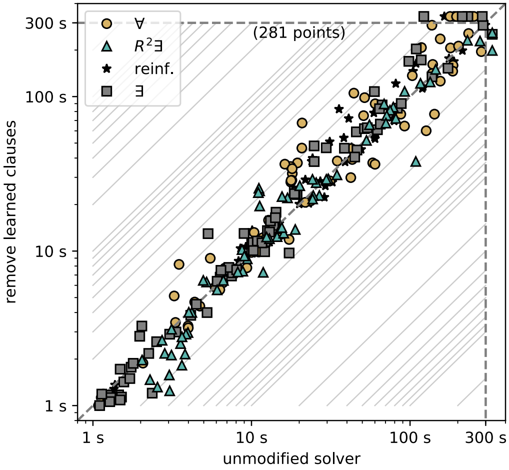
- used to compare two algorithms / solvers
- use log-log plot with 1:1 ratio
- can add points for timeout / error
- allows to see speed-up
- pitfall: where are points to be better?
- pitfall: difficult to see how many points are in point cloud



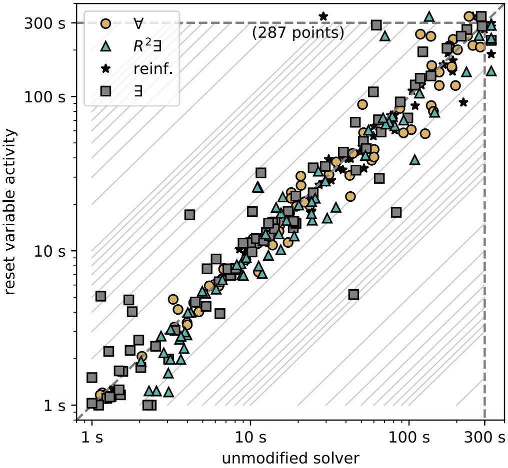
Scatter Plot — Variation in Running Time



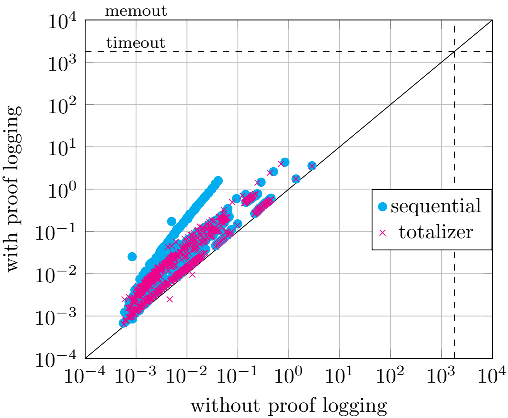
Scatter Plot — Solvers as Chaotic System (1)

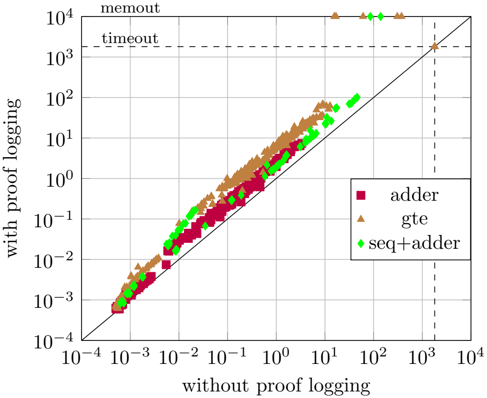


Scatter Plot — Solvers as Chaotic System (2)



Scatter Plot — Overhead





Empirical Asymptotic Scaling – Basics

from theory: this algorithm runs in time $O(n^2)$
$\Rightarrow$ can we empirically evaluate asymptotic behaviour?

problems:

- humans bad at judging curves
- polynomial with high enough degree will fit everything

assumptions:

- $f(n) = an^b + g(n)$
- $g(n) \ll an^b$

idea:

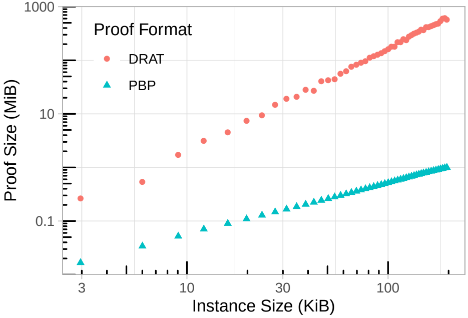
- plot $y = \log(f(n))$ against $x = \log(n)$
- $\log(f(n)) \approx \log(an^b) = \log(a) + b\log(n)$
- $\Rightarrow$ we should see a straight line ($y \approx \log(a) + bx$)

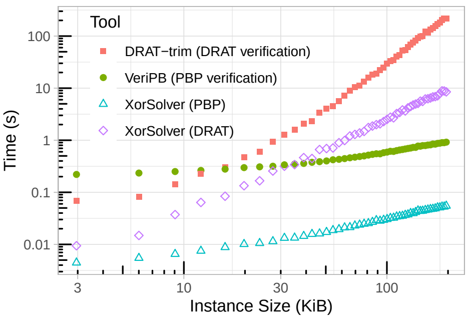
Empirical Asymptotic Scaling – Challenges

- need $g(n) \ll an^b$, is chosen n large enough?
  $f(n) = n + 0.1^{1'000'000} n^2$
- running time jumps when exceeding CPU cache limits
  $f(n) = n^2$ if $n > 512$ else $0.1n^2$
- data can be "noisy"
  $f(n) = \varepsilon(n) \cdot n^2$
- choice of scaling parameter
  e.g.: number of variables, number of clauses, file size

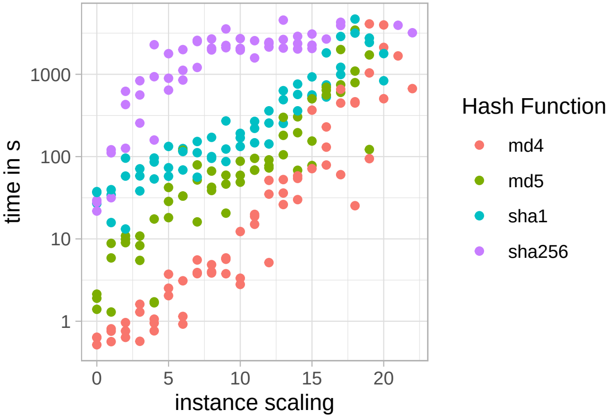
- can't **proof** asymptotic behaviour, only says what scaling is plausible
- can highlight problems in implementation (expected n^2 but got n^3)
- might show data **not** meeting assumptions (e.g. $f(n) = 2^n$)

Empirical Asymptotic Scaling – Polynomial Example





Empirical Asymptotic Scaling – Exponential Example



Box-plot

- used to visualize variation
- box ranges between quantiles, bold line is median, whiskers not standardized (1.5 IQR, all data points), points beyond whiskers are called outliers