“A proposal for a new Stata licensing scheme based on blockchain, cloud computing, and grid computing”

Alexander Zlotnik, PhD
Technical University of Madrid (Universidad Politécnica de Madrid)

David Arroyo Manzano, MsSc
Why?
Everyone will be using…

“Big data”
+
Complex algorithms
=
Lots of computational resources
Examples

• (very) big data & simple operations (such as sort)

• _____ big data & regression analysis

• _____ big data & multiple imputation

• (just) _____ data & bayesian analysis
Current Stata solutions

• Custom programming in C++

• Stata / MP

• Stata distributed processing (several computers)
  …example: Stata PARALELL
Custom programming in C++

• Example:

2015 UK Stata Users Group meeting

Big Data in Stata
Andrew Maurer
Quantitative Risk Management
Current Stata solutions

• Custom programming in C++

• **Stata / MP**

• Stata distributed processing (several computers)
  …example: Stata PARALELL
Stata / MP

Source: https://www.stata.com/statamp/
Current Stata solutions

• Custom programming in C++

• Stata / MP

• Stata distributed processing (several computers)
  …example: Stata PARALELL
Stata PARELLEL

Authors

George G. Vega [aut,cre] g.vegayon %at% gmail
Brian Quistorff [aut] Brian.Quistorff %at% microsoft

PARALLEL: Stata module for parallel computing

- stata
- parallelization
- bootstrap
- simulation
- hpc
- parallel

325 commits  3 branches  4 releases  3 contributors

Branch: master  New pull request

bquistorff update HTML from sthlp update

Latest commit 72ee46f on 29

- .github
- Update ISSUE_TEMPLATE.md
- ado
- update HTML from sthlp update
- man
- Normalize all the line endings
Figure 1: How parallel works

Starting (current) Stata instance loaded with data plus user defined globals, programs, Mata objects and Mata programs.

Splitting the data set

A new Stata instance (batch-mode) for every data-clusters. Programs, globals and Mata objects/programs are passed to them.

Task (Stata batch-mode)

The same algorithm (task) is simultaneously applied over the data-clusters.

Appending the data set

After every instance stops, the data-clusters are appended into one.

Ending (resulting) Stata instance loaded with the new data.

User defined globals, programs, Mata objects and Mata programs remain unchanged.

Source: http://fmwww.bc.edu/repec/bocode/p/parallel.pdf
Distributed processing

• Important concepts
  – Algorithmic complexity
  – Ahmdal’s law

• Decision criteria
Algorithmic complexity

![Graph showing algorithmic complexity](image)
Ahmdal’s law
Distributed processing

• Important concepts
  – Algorithmic complexity
  – Ahmdal’s law

• Decision criteria
Decision criteria

• High $n$?

• $O(n) = \text{algorithmic complexity}$?

• Parallelizable code (Ahmdal’s law)?
Example: Multiple Imputation

• High n ?
  – Many experiments => High n => Yes

• $\mathcal{O}(n) = \text{algorithmic complexity}$ ?
  – $\mathcal{O}(n) \approx n$ (regressions)

• Parallelizable code (Ahmdal’s law) ?
  – Many independent experiments => Yes
Ideas for future Stata versions
Ideas for future Stata versions

• Stata private cloud

• Stata public cloud (grid computing) 
  … with blockchain licensing
Chessbase private cloud

28...Qf7! 29.Re6! Vishy manages to find the best way to develop his counterplay.

29...Ng4? This natural move is not the best from an objective point of view, though White’s task is becoming extremely tough.

Komodo 9.02 64-bit Depth=27 29...Kh8 (1/39) 5049 K/N/s
Stata private cloud proposal

• Do some processing locally.

• Remove all identifying information (variable names, variable encoding, non-numerical values, et cetera).

• Send complex optimization problems to Stata cloud.

• Get results in local instance of Stata.
Ideas for future Stata versions

• Stata private cloud

• Stata public cloud (grid computing) … with blockchain licensing
Stata public cloud (grid)

• Many computers …
  … in different geographical locations
  … working on the same problem

• Example: SETI@home
Stata public cloud (grid)

• The same approach could be used with Stata.

• But… how could Stata users be incentivized to provide their instances of Stata for distributed processing?
  – With blockchain licensing!
What is a blockchain?
Blockchain applications

• blockchain = distributed database (distributed ledger) with transactional integrity guarantees not controlled by a single entity based on many processing nodes (anonymous or publicly known).

• It is very hard (almost impossible, given certain conditions) to falsify an entry in the blockchain.
Blockchain applications

• Civil registries.

• Land ownership registries.

• Notary registries.
Blockchain applications

• Cryptocurrencies (Bitcoin, “ether”, etc) which are not controlled by a central bank (or any kind of central entity).

• International financial transactions (alternatives to the SWIFT system).
Blockchain applications

• Smart electricity grids (intelligent electricity production, distribution and billing).

• Distributed organizations (such as cooperatives with no managers).

• e-Administration / Open Government (Malta, Russia, Ukraine, Estonia, …)
Stata public cloud + Blockchain licensing

• “Free” Stata license which is paid for by computational time for Stata Corp.

• Computational time given to Stata Corp is logged in a blockchain thus guaranteeing transparency and irrevocability.
Thank you !