Grid technologies for Social Science: the Seamless Access to Multiple Datasets (SAMD) project

Authors: Celia Russell, Keith Cole, M. A.S. Jones, S.M. Pickles, M. Riding, K. Roy, M. Sensier

IASSIST, University of Wisconsin, Madison
25th-28th May 2004
Structure of the talk

- What is e-Science?
- e-Social science
- About the SAMD project:
  - Method
  - Architecture
  - Results and outputs
- Extending the project to other social science applications
- Implications for social science
What are e-Science Grids?

- E-science grids are a new IT infrastructures that allow easier and faster access to distributed computing and data resources.
- An enquiry to a Grid search engine will not only find the data you need but also the data processing techniques and the computing power to carry them out before sending you the results.
- The scale of investment and the potential of the technology suggests Grid infrastructures will play a major role in future quantitative research in the social sciences.
Grid technologies run over existing internet infrastructures and offer a faster alternative to the current world wide web for the transfer and analysis of large datasets.

The Grid provides a way of managing very large databases (petabytes or even terabytes)

The Grid also uses a different security model to the web.

Currently Grid technologies are used in data intensive physical science applications. In this talk we look applications in the social sciences.
Benefits of the Grid

- Enable large-scale applications comprising thousands of computers
- Transparent access to "high-end" resources from your desktop
- Provide a uniform "look & feel" to a wide range of resources with no need to know unix etc
- Better handling of large and complex datasets
Some examples of existing Grid projects

- **High Energy Particle Physics**

In particle physics, the traditional approach of extracting data subsets across the Internet, storing them locally, and processing them with home-brewed tools has reached its limits.

Modern particle physics experiments might produce over a Petabyte ($10^{15}$ bytes, a billion megabytes) of data per year and the ability to analyze data and move it between international collaborators has not kept up with its increased flow.
Some examples of existing Grid projects 2

- **Climateprediction.net**

  Is creating the first probability-based ("Monte Carlo") 50-year forecast of human-induced climate change, using a full-scale 3-D climate simulation model.

  Grid technology makes it possible to utilise the idle processing capacity from millions of personal computers to obtain more computing power than is available by conventional sources.
Some examples of existing Grid projects

- **E-Star**

The e-STAR project aims to develop a network of robotic telescopes connected via appropriate middleware to enable distributed, dynamically scheduled, astronomical observations to be performed.

The principles developed in the project can be applied to other applications that rely on the availability of expensive and time-limited facilities, analysis of vast amounts of data and access to massive quantities of archived data.
The application of Grid (e-Science) technologies in a social science context is called e-Social Science.

The Economic and Social Research Council in the UK is funding a series of programmes to stimulate the uptake and use by social scientists, of new and emerging Grid-enabled computing infrastructures, both in quantitative and qualitative research.

The first successful demonstrator project to be funded by this program was the SAMD project.
Seamless Access to Multiple Datasets

- A project to demonstrate the benefits of applying e-Science grid technologies to an ordinary social science query
- We solve a genuine problem from the UK academic social science community - a multivariate analysis using a complex mathematical algorithm
- Based on a major social science databank, the UK Office for National Statistics Time Series Data, hosted at MIMAS
The problem


- The research query looks at the effect interest rate changes had on Gross Domestic Product in the UK over the period 1960 – 2000
Interest Rates in the UK

Interest rate, %, 1957 -2000

Supercomputing, Visualization & eScience
UK GDP – quarterly changes
The Model

\[ y_i - \phi y_i \mid \sum_{t=1}^{\delta} \phi_t y_{i-t} \cdot \sum_{t=1}^{\delta} \delta_{t}z_{i-t}. \]

\[ F(r) [\phi y_i + \sum_{t=1}^{\delta} \phi_t y_{i-t} - \sum_{t=1}^{\delta} \delta_{t}z_{i-t}] - \epsilon \]

\[ F(r) = \frac{1}{1 + \exp\left( \gamma(r - c) \sigma(r) \right)} \]

Where \( y \) is the quarterly change in GDP and \( z \) is the quarterly change in interest rates.
Before SAMD

Best part of a day
Grid Model Used
We built a mini demonstrator grid for SAMD by:

- Grid-enabling the NS Time Series Databank
- Parallelising the code to represent the HPC facilities
- Using Grid protocols for data transfer
- Creating a graphical user interface that included a single sign-on
- It all worked, and cut the data collection and analysis time down to around 8 minutes.
The SAMD solution

- Use Grid Security Infrastructure for "single sign-on" authentication everywhere
  - Modified standard Apache web server to accept proxy credentials
    - Permits re-use of existing CGI code
- Use third party file transfers (grid-ftp) to move data directly to where it's needed
- Use standard globus mechanisms to
  - Locate HPC facility for analysis
  - Stage analysis binary from local repository and run analysis job on HPC facility
  - Retrieve results
SAMD user interfaces
Data Request

Data moved to GridFTP server
- 1: send references to data
- 1,2,3: authentication & authorisation
- 4: ask datastore to move data (5)
- 6,7: datastore returns XML ticket
Finding an HPC Resource

GIIS MDS Server
- e.g. ginfo.grid-support.ac.uk

Search for:
- OS type e.g.: IRIX64
- Minimum No. Processors
- Jobmanager
- or manually enter your favourite

Data Analysis panel
Using the HPC Resource

- Select an executable on the local machine
- Stage job using Globus
- Check status using Globus
- Retrieve results using Globus
- Clean-up using Globus
- Even delete job using Globus

Data Analysis panel
The approach and methods of SAMD are applicable to more general social science applications involving data collection and analysis.

Some of the SAMD resources reused in other Grid applications. These are available on the SAMD website: [http://www.sve.man.ac.uk/Research/AtoZ/SAMD](http://www.sve.man.ac.uk/Research/AtoZ/SAMD)

SAMD shows that such an e-social science environment is technically possible. For e-Social science to develop, key datasets need to be Grid-enabled in a commonly understood, well-documented way.
What’s new with SAMD?

- More efficient handling of datasets – data is moved to where it's needed, not just to web browser
- The single sign-on for all databanks means users can cross search datasets and perform cross analyses of multiple datasets from different providers
- Grants access to high performance computing facilities without the user having to learn how to use them
- Can automate routine enquiries
- Cuts the time taken to run computing intensive problems by a factor of around 100
A Grid approach allows the social scientist to scale up their quantitative research by:

- Including many more data points in their analysis
- Developing more complex models incorporating more variables
- Dropping assumptions
- Visualising data
- Creating new communities and collaborations
- Exploring new types of analyses
# SAMD Acknowledgments

<table>
<thead>
<tr>
<th>Keith Cole</th>
<th>Mark Riding</th>
<th>Geoff Lane</th>
</tr>
</thead>
<tbody>
<tr>
<td>Celia Russell</td>
<td>Kevin Roy</td>
<td>Tim Hateley</td>
</tr>
<tr>
<td>Marianne Sensier</td>
<td>Stephen Pickles</td>
<td></td>
</tr>
</tbody>
</table>

*Funded by the [MIMAS](https://www.mimas.ac.uk), [CSAR](https://www.csar.ac.uk), and the [ESNW](https://www.esnw.org.uk).*