e-Social Science Development in Taiwan: Scalable social simulation using commodity computing

Jing-Ya You

April 12, 2011, EGI User Forum 2011, Vilnius Lithuania

e-Social Science Projects

SimTaiwan – Migration in Taiwan

- Alex Voss[1], Jing-Ya You[2], Eric Yen[2], Simon Lin[2], Ji-Ping Lin[3], Andy Turner[4]

RAS (Receive-Accept-Sample model) – A theory-based model describing how an individual citizen forms its political preferences

- Frank C.S. Liu[5], Jing-Ya You[2], Yu-Ting Chen[2], Jing-Lung Sun[6]

[1]School of Computer Science, University of St Andrews, UK
[2]Academia Sinica Grid Computing, Academia Sinica, Taiwan
[3]Center for Geographic Information Science, Academia Sinica, Taiwan
[4]School of Geography, University of Leeds, UK
[5]Institute of Political Science, National Sun Yet-San University
[6]Department of Computer Science, National Tsing Hua University

Overview

- Social simulation and agent-based modeling
- [User Cases]:
  - SimTaiwan: a model of migration in Taiwan
  - RAS: a theory based model of citizens' politic preferences
- How commodity computing facilitates researches on social science
  - Social simulation at scale and tractability
  - Model validation and verification for social simulation
- Observation on computing infrastructures
- References

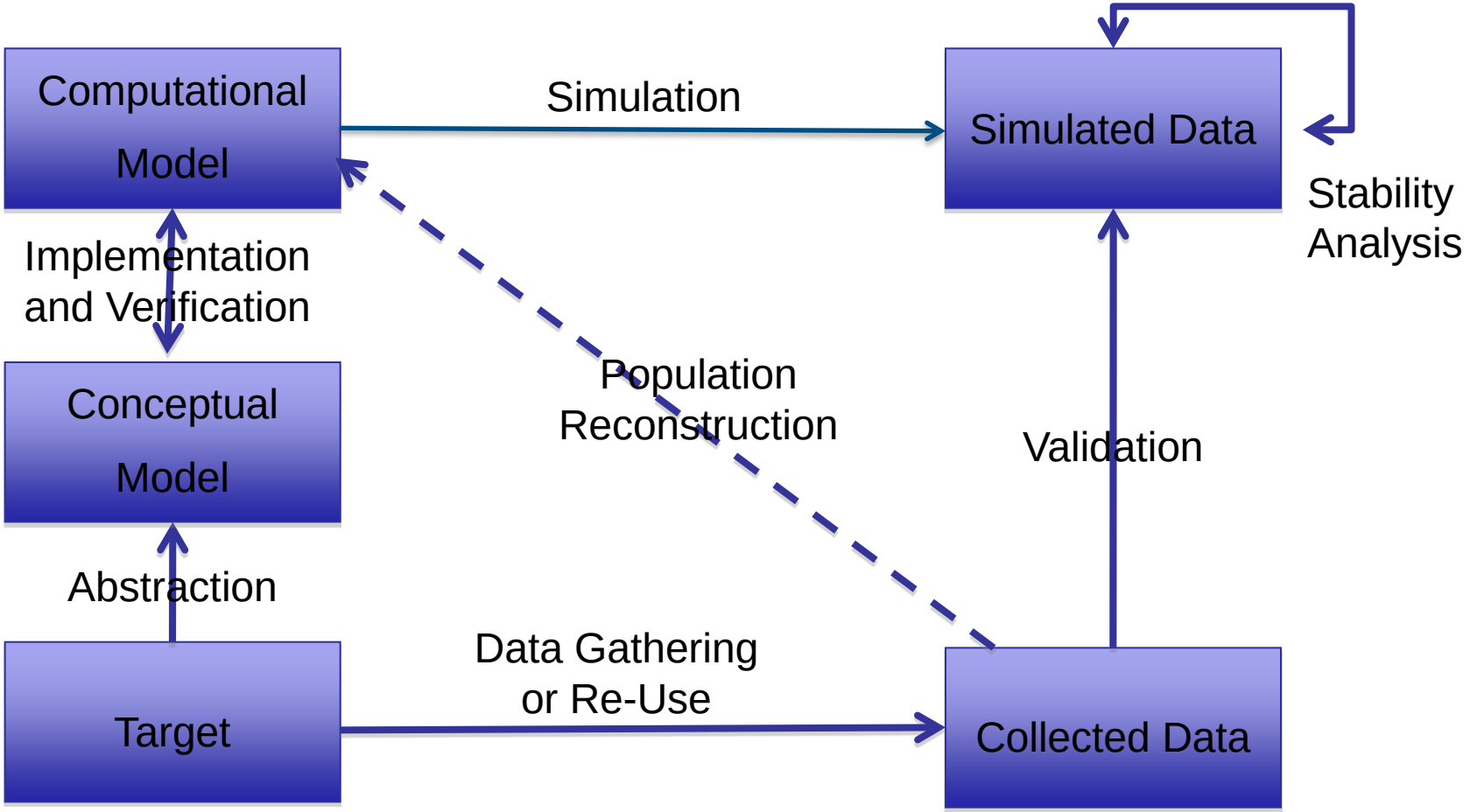
Agent-based Social Simulation

- A new approach to modeling social phenomena
- Based on empirical data
- Based on existing theories
- A new way to explore them, complementing other forms of modeling and prediction
- Used to understand and predict
- Interpretation between macro and micro levels

Agent-Based Social Simulation

- Simulating interactions between dynamic populations in changing environment
- Heterogeneous populations – each individual has specific attributes such as age, gender, socio-economic status, health, etc.
- Stochastic process – each run can differ from previous
- Model and analyse phenomena too complex for closed form, can be used in absence of knowledge about causality
- Supported by Agent-Based Modelling Frameworks: Repast Simphony, Mason, Swarm, NetLogo, FLAME, etc.

Social Simulation as a Research Method

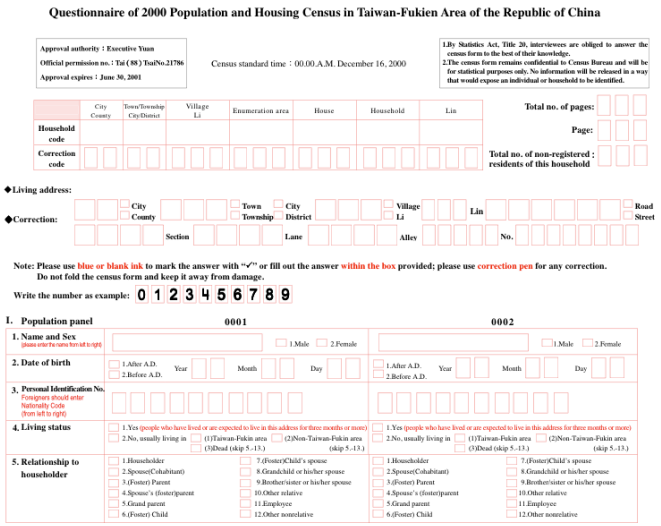


SimTaiwan: Migration in Taiwan

- Migration has been an important factor in Taiwanese social development and influenced by outside factors since the 1600s
- Aim is to test existing theories of migration constructively and to investigate recent developments such as increased outward migration to China.
- Timely as Taiwan is running another census in 2010.
- Based on existing statistical models of migration developed at Academia Sinica Center for Survey Research.

SimTaiwan: Migration in Taiwan

- Based on Taiwan 2000 Population and Housing Census
- Dataset is individual-level but with restricted variables, held at Academia Sinica
- Individual-level data needed to estimate further socio-economic variables (e.g., income)
- Additional datasets providing data about geographic regions: population size, employment and unemployment rate, average wages
- Need to scale up to ca. 22 million agents of reasonable complexity – ca. 20 attributes per agent



Questionnaire of 2000 Population and Housing Census in Taiwan-Fukien Area of the Republic of China

Approval authority : Executive Yuan
Official permission no. : Tai (88) TsaiNo.21786
Approval expires : June 30, 2001

Census standard time : 00.00.A.M. December 16, 2000

1.By Statistics Act, Title 20, interviewees are obliged to answer the census form to the best of their knowledge.
2.The census form remains confidential to Census Bureau and will be for statistical purposes only. No information will be released in a way that would expose an individual or household to be identified.

	City County	Town/Township City/District	Village Li	Enumeration area	House	Household	Lin
Household code							
Correction code							

Total no. of pages:
Page:
Total no. of non-registered residents of this household

◆Living address:
◆Correction: City/County, Town/Township, City/District, Village/Li, Lin, Road/Street, Section, Lane, Alley, No.

Note: Please use blue or blank ink to mark the answer with "✓" or fill out the answer within the box provided; please use correction pen for any correction. Do not fold the census form and keep it away from damage.

Write the number as example: 0 1 2 3 4 5 6 7 8 9

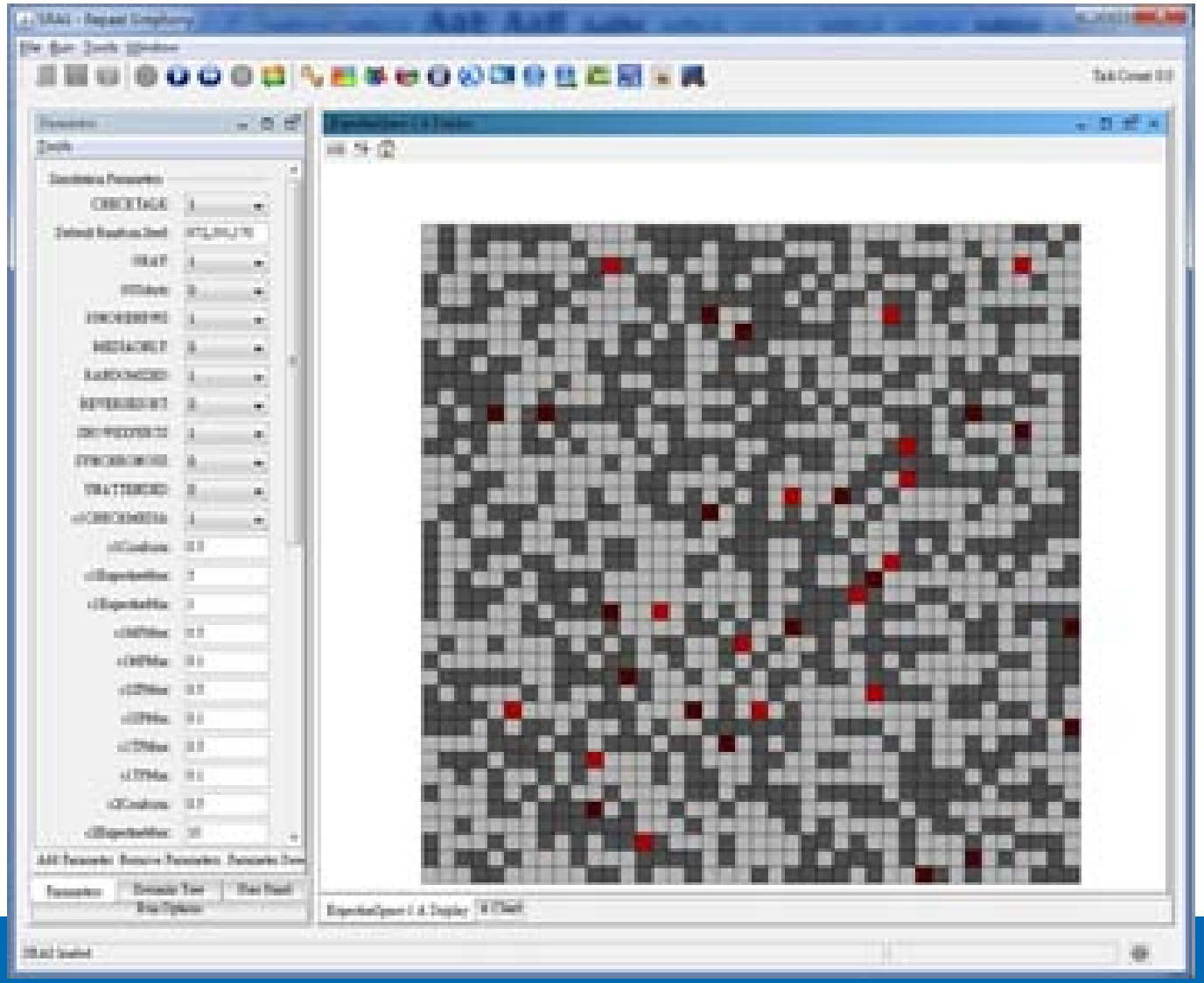
I. Population panel

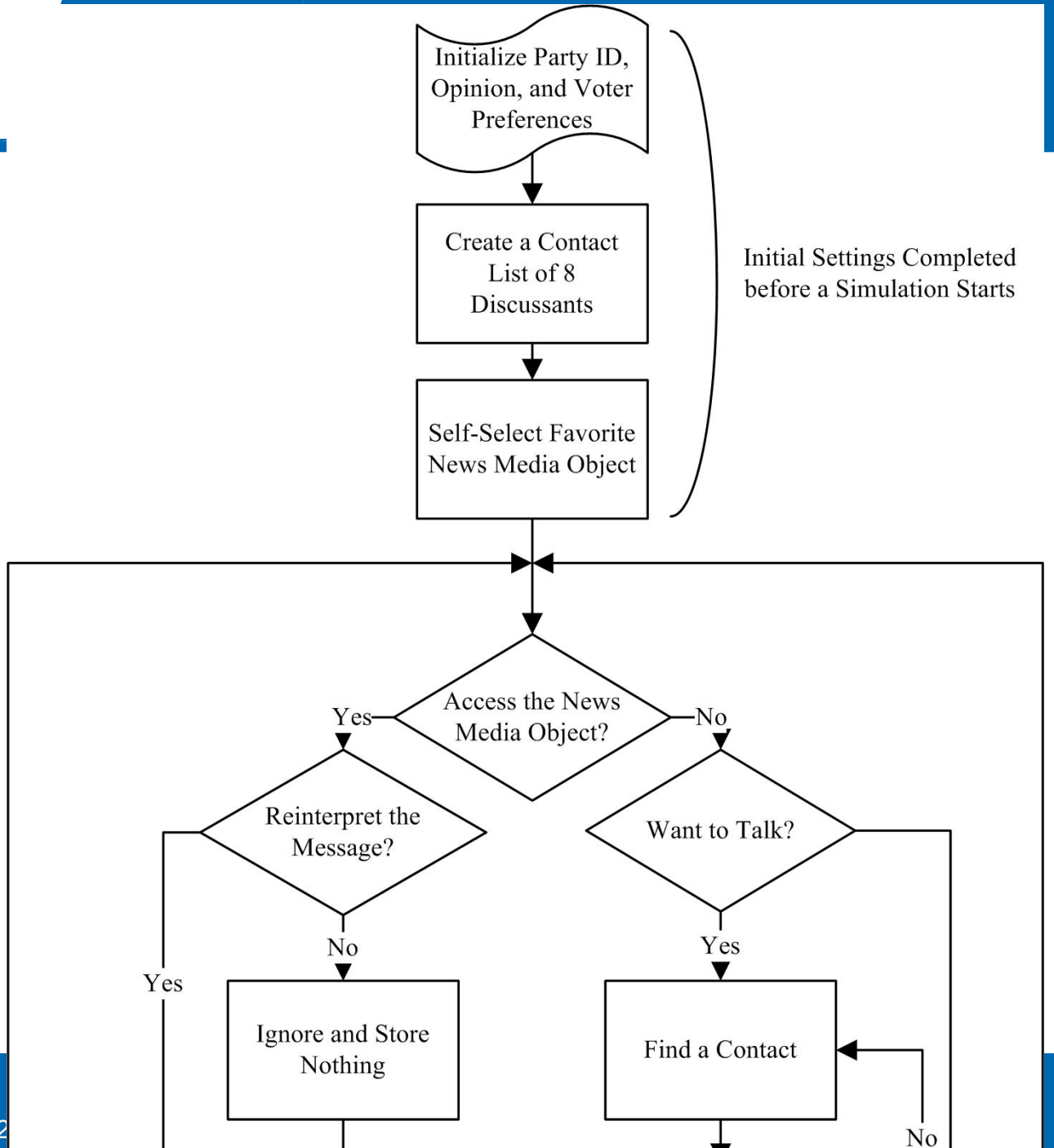
	0001	0002
1. Name and Sex (please enter the name from left to right)	1.Male 2.Female	1.Male 2.Female
2. Date of birth	1.After A.D. 2.Before A.D. Year Month Day	1.After A.D. 2.Before A.D. Year Month Day
3. Personal Identification No. Foreigners should enter Nationality Code (from left to right)		
4. Living status	1.Yes (people who have lived or are expected to live in this address for three months or more) 2.No, usually living in (1)Taiwan-Fukin area (2)Non-Taiwan-Fukin area (3)Dead (skip 5.-13.) (skip 5.-13.)	1.Yes (people who have lived or are expected to live in this address for three months or more) 2.No, usually living in (1)Taiwan-Fukin area (2)Non-Taiwan-Fukin area (3)Dead (skip 5.-13.) (skip 5.-13.)
5. Relationship to householder	1.Householder 2.Spouse(Cohabitant) 3.(Foster) Parent 4.Spouse's (foster)parent 5.Grand parent 6.(Foster) Child 7.(Foster)Child's spouse 8.Grandchild or his/her spouse 9.Brother/sister or his/her spouse 10.Other relative 11.Employee 12.Other nonrelative	1.Householder 2.Spouse(Cohabitant) 3.(Foster) Parent 4.Spouse's (foster)parent 5.Grand parent 6.(Foster) Child 7.(Foster)Child's spouse 8.Grandchild or his/her spouse 9.Brother/sister or his/her spouse 10.Other relative 11.Employee 12.Other nonrelative

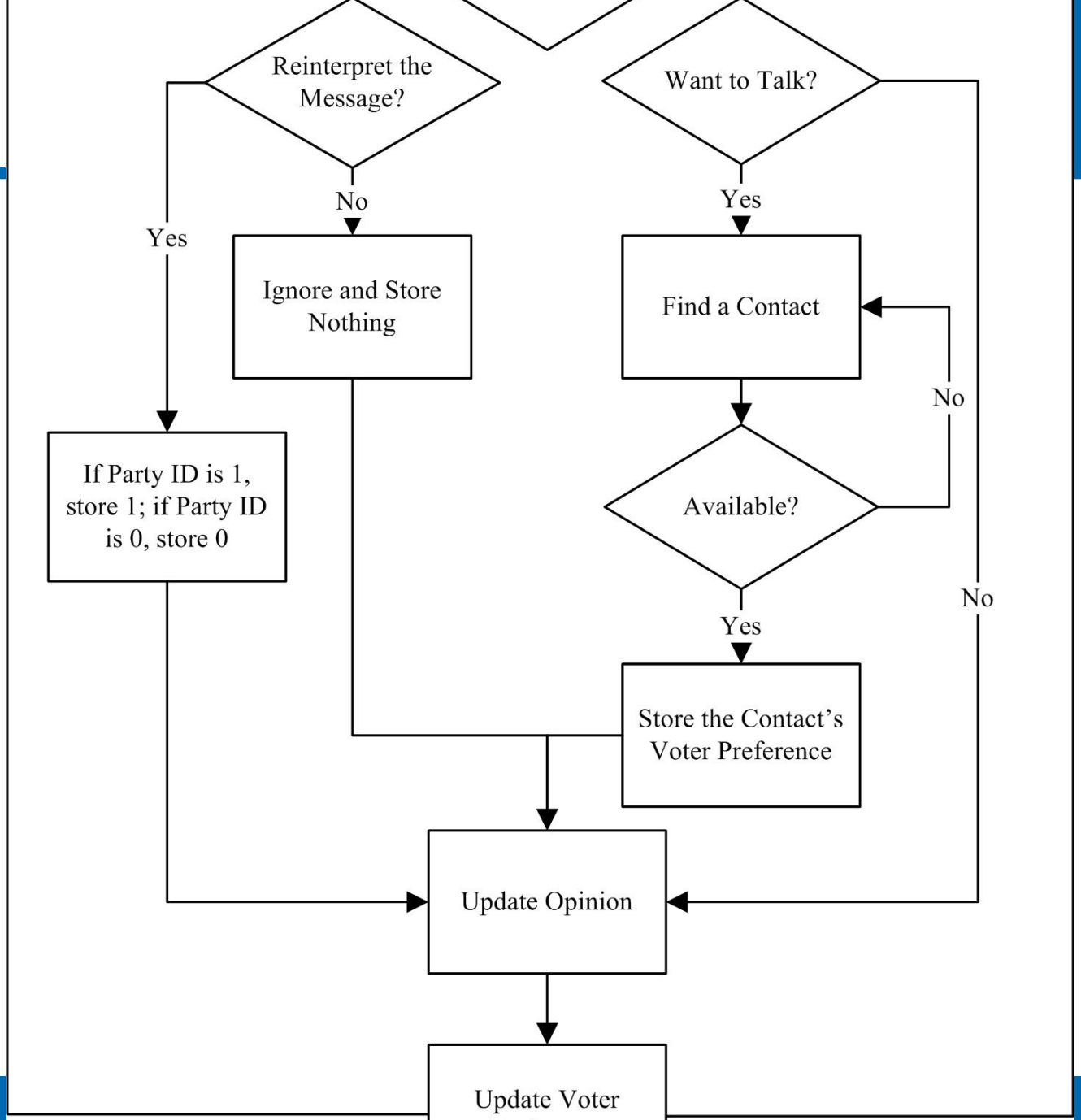
SimTaiwan: Migration in Taiwan

- Divided into two parts: departure and destination
- Probability of departure is a weighted sum of different factors such as: Sex, Age, Education, Marital status, Population size, and Employment at departure location
- Weights are provided by statistical model based on actual migration patterns measured by census.
- Note: we do not currently have linked socio-economic data such as income

RAS, the model




Initialize Party ID, Opinion, and Voter Preferences
Create a Contact List of 8 Discussants
Self-Select Favorite News Media Object
Initial Settings Completed before a Simulation Starts
Access the News Media Object?
Yes
No
Reinterpret the Message?
Want to Talk?
Yes
No
Yes
Ignore and Store Nothing
Find a Contact
No

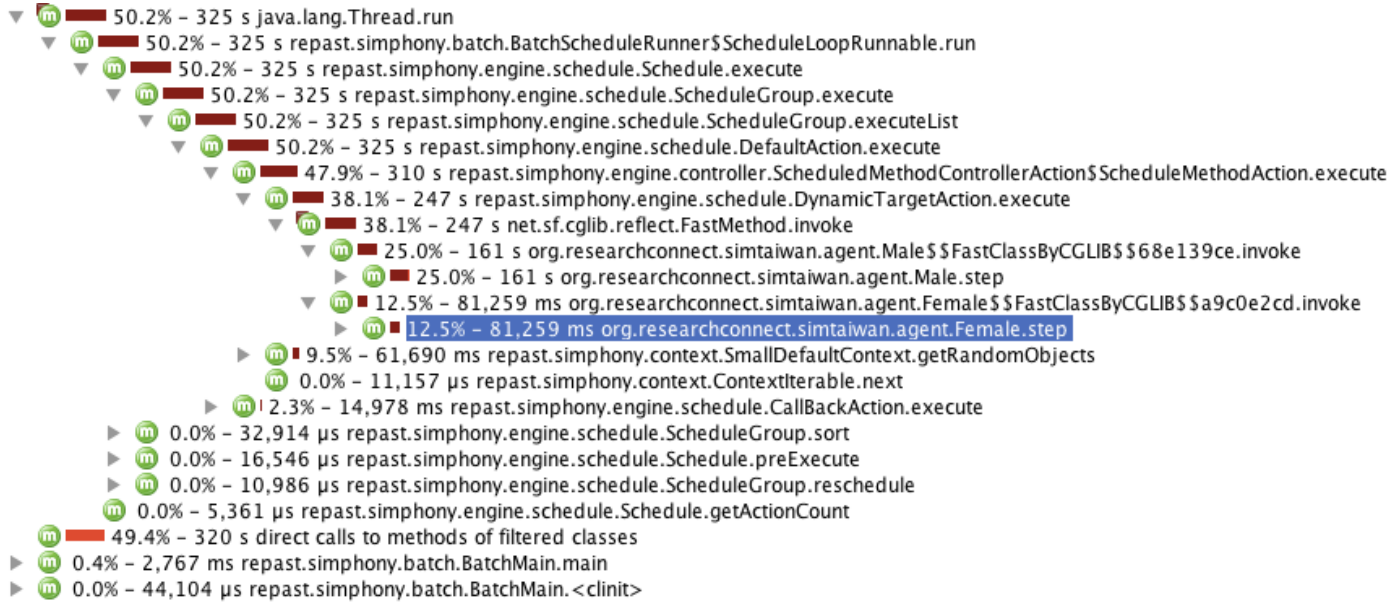

Reinterpret the Message?
Want to Talk?
No
Yes
Yes
Ignore and Store Nothing
Find a Contact
No
If Party ID is 1, store 1; if Party ID is 0, store 0
Available?
No
Yes
Store the Contact's Voter Preference
Update Opinion
Update Voter Preference

Test Code and Parameters

- Simplified model of SimTaiwan with only fertility and mortality, same for all measured models
- 250k male and 250k female random initial population, running for 365 ticks (=days)
- Measurements taken using JProfiler6.4
  - CPU sampling (5 sec intervals)
  - Memory allocations recording
- JVM parameters: -Xmx8192M –Xss128M
- Hardware: Dell PowerEdge R610 with 2xXeon E5504 @ 4x2GHz (8 cores total) and 16GB RAM

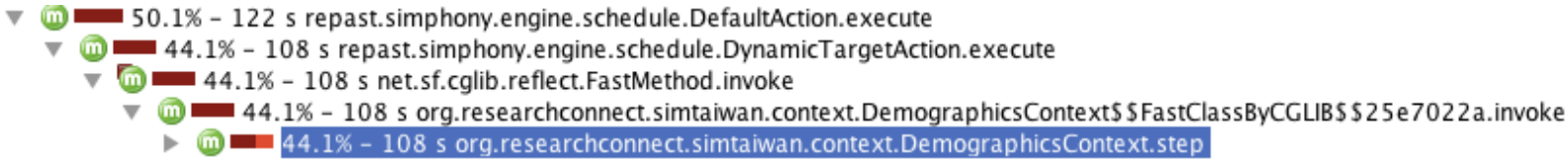
Naïve Serial Version

- More time spent in RePast scheduling code than in model code because events are scheduled for each individual agent every step.



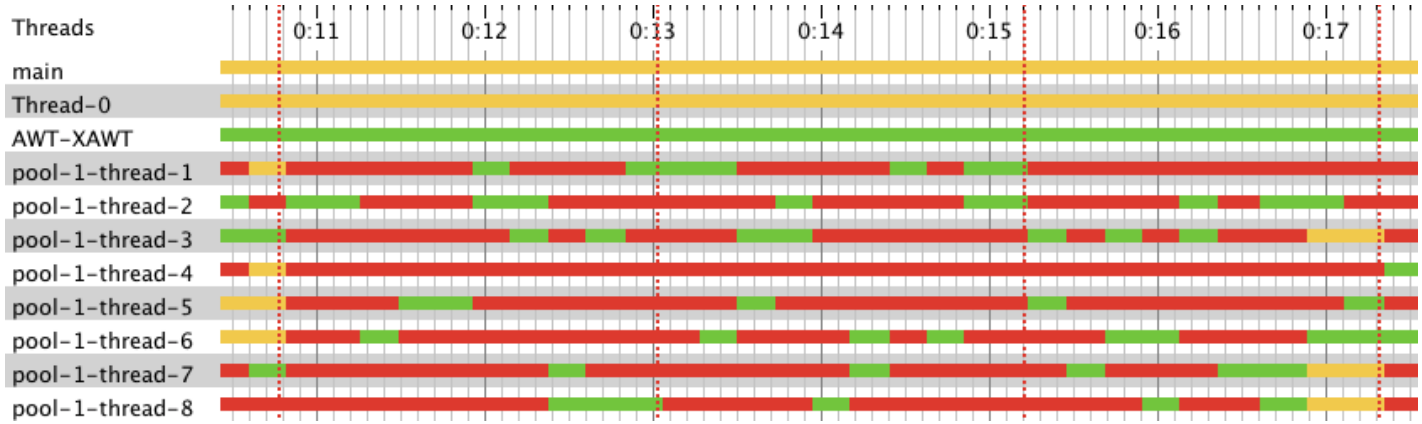
Improved Serial Version

- Event scheduled on DemographicsContext, code iterating through individual agents



- Wallclock time down from 5:32 to 2:23
- Opens up opportunities for parallelising code as well…

Initial Parallel Version (II)



- Worker threads blocking a lot on monitors placed around RePast constructs.
- Main issue seems to be that use of RandomHelper is not thread-safe
- Simulation schedule relatively minor issue
- Some contention around simulation objects

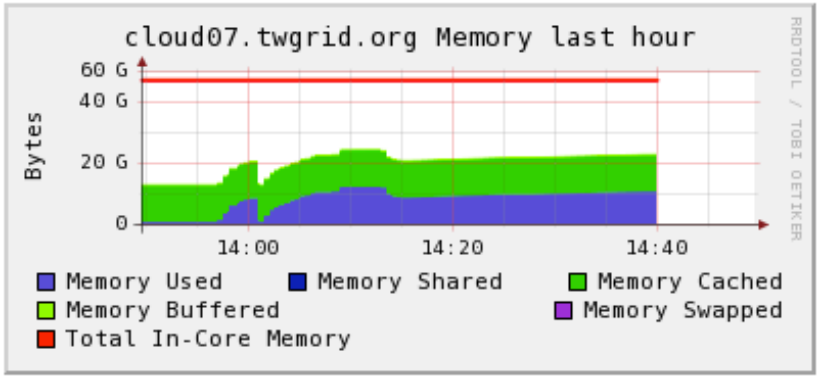
Improved Parallel Version

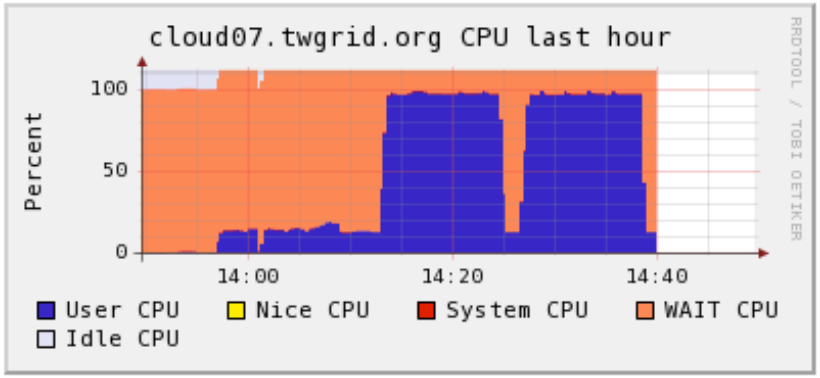
- Overloading some of RePast's code to make it thread safe.
- Reducing scope of monitor objects used and pulling code parts that are safe out of synchronized sections
- Introducing thread-local variable containing a per thread random number generator:

```
protected static ThreadLocal<Uniform> uniform = new ThreadLocal<Uniform>() {
  @Override
  protected Uniform initialValue() {
    RandomEngine generator = new MersenneTwister((int)System.currentTimeMillis());
    return new Uniform(generator);
  }
};
```

Improved Parallel Version (II)

- Monitor contention is eased signficantly
- Wallclock running time down to 1:03 and max. CPU utilisation up to ~ 600%
- Time spent in serial code for analysis and production of charts is now significant.
- Issue: Iteration through contexts needs to be randomised



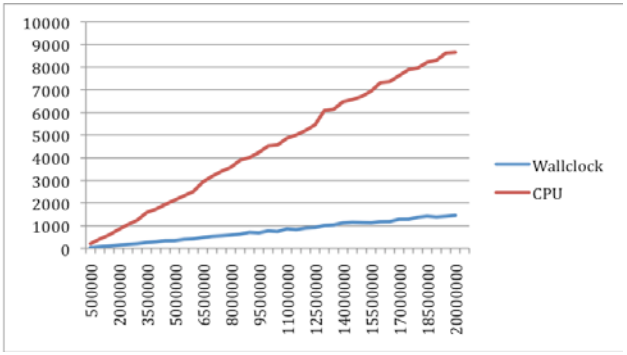


Debugging and Profiling on Grid

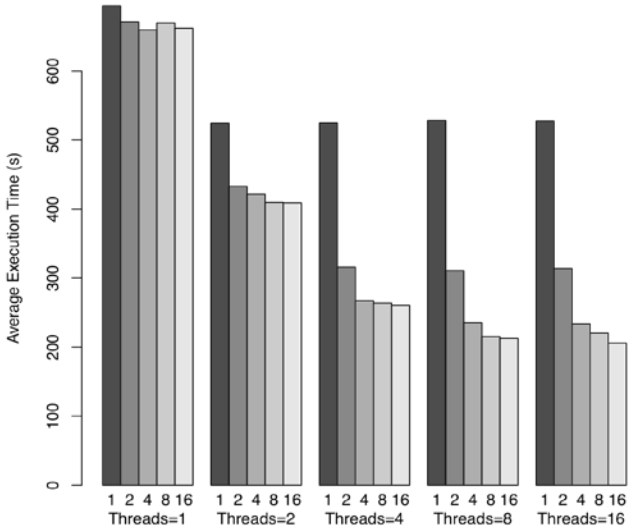
- Repast environment, JProfiler, and Ant installed on EUAsiaGrid resources at Academia Sinica
- Wallclock times not meaningful in a shared environment, need other measures – JProfiler
- Repeated runs provide:
  - Access to behaviour of model under different conditions, e.g. number of threads, number of partitions
  - Average measures of thread utilisation and runtime
  - Evidence of rare events (exceptions, deadlocks and non-termination)

Model Performance Study

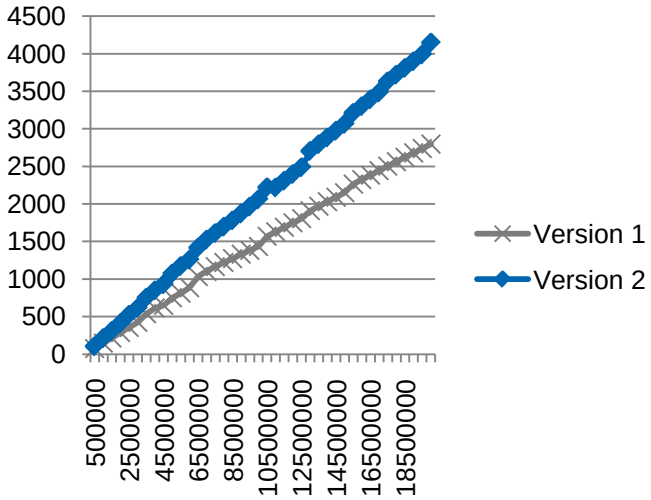
- Parallel Model Performance
  - Based on number of worker threads
- Memory and Running Time Estimation
  - Different size of population
  - Various complexity and variables of agents



Model Running Time



Parallel Model Performance Estimation



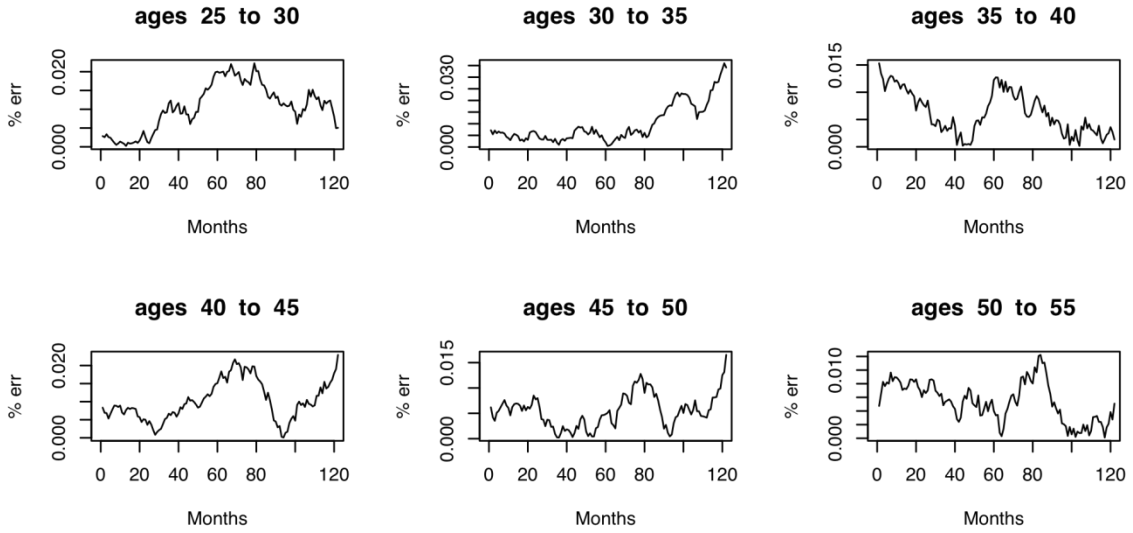
Model Heap Size Estimation

Stability and Sensitivity Test on Grid

- Validation and verification for models
  - Stability Test and Sensitivity Test
- Necessary to run multiple simulations that vary specific variables in order to study their impact
- Hide the complexity of Grid for social scientists
- Scale up duration and number of repeated runs for sensitivity test of simulation models
- Deploy statistics software on Grid/Cloud, such as R or SAS

Stability Test for SimTaiwan

- Stability analysis using count of female agents over time for a number of age groups.
  - Repeated runs: 80 runs of the model with 1 million initial agents over 10 year of simulated time (in 1-day steps)
  - Measures: Number of population of each month (30-days steps) for each age group
  - Assessment of variation: Percentage error less than 0.1%



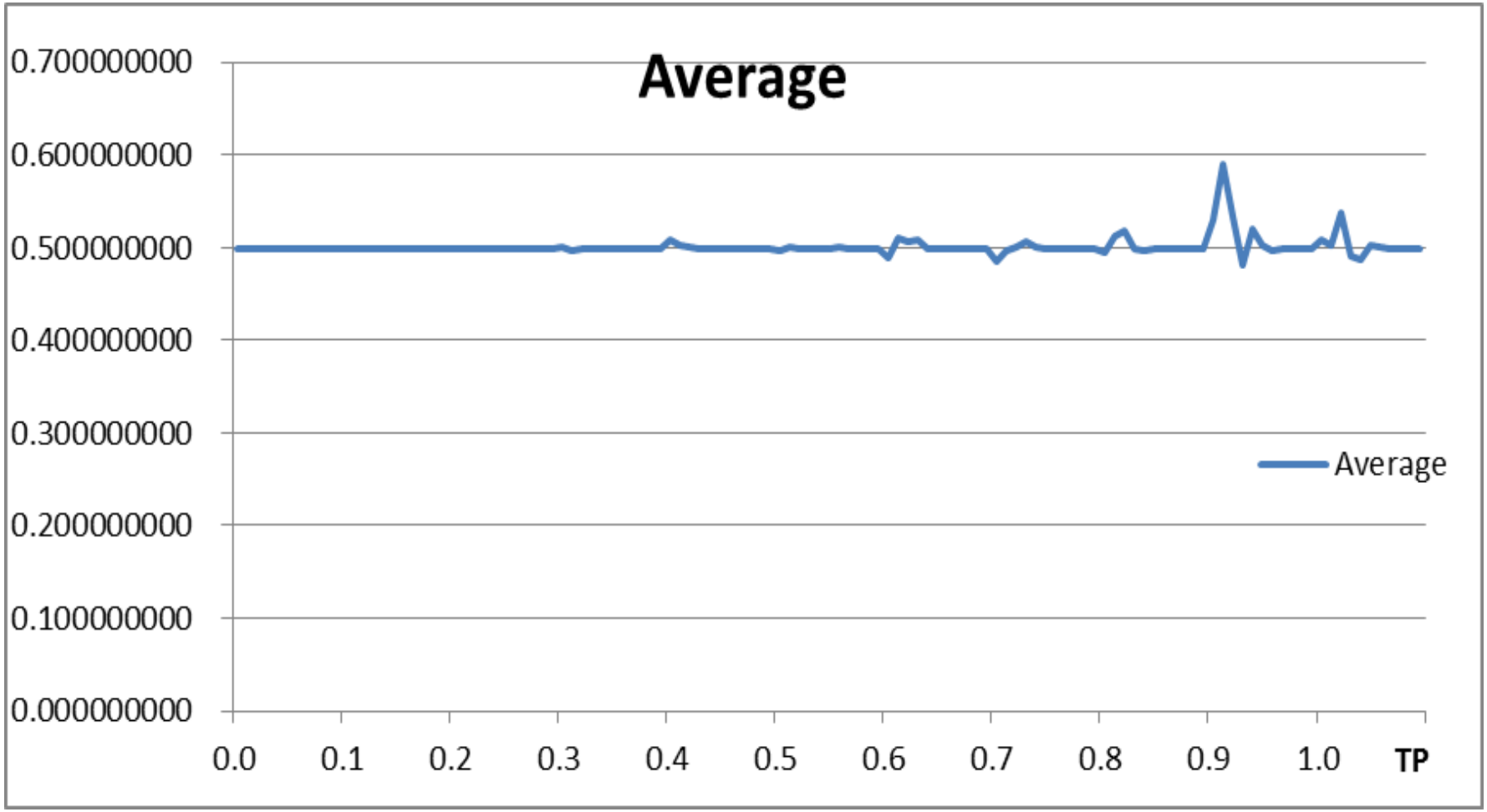
Sensitivity Test for RAS: Parameters

- Two Parameter sweeping:
  - **TP: propensity to talk politics**
  - **MP: propensity to access the news media**
- Vary the parameter values of TP from 0.0 to 1.0 with increments of 0.1, while fixing the value of all other parameters. And then vary MP from 0.0 to 1.0.
- This results in 11 * 11 =**121** sets of parameter combination (to be detailed below). We chose to monitor the proportion of agents and run each set of parameter combination for 70 times, each of which is given a unique seed and runs for **10,000** time steps. This sequence will lead to a total of 11 * 11 * 70 = **8,470** “tulples” or works.

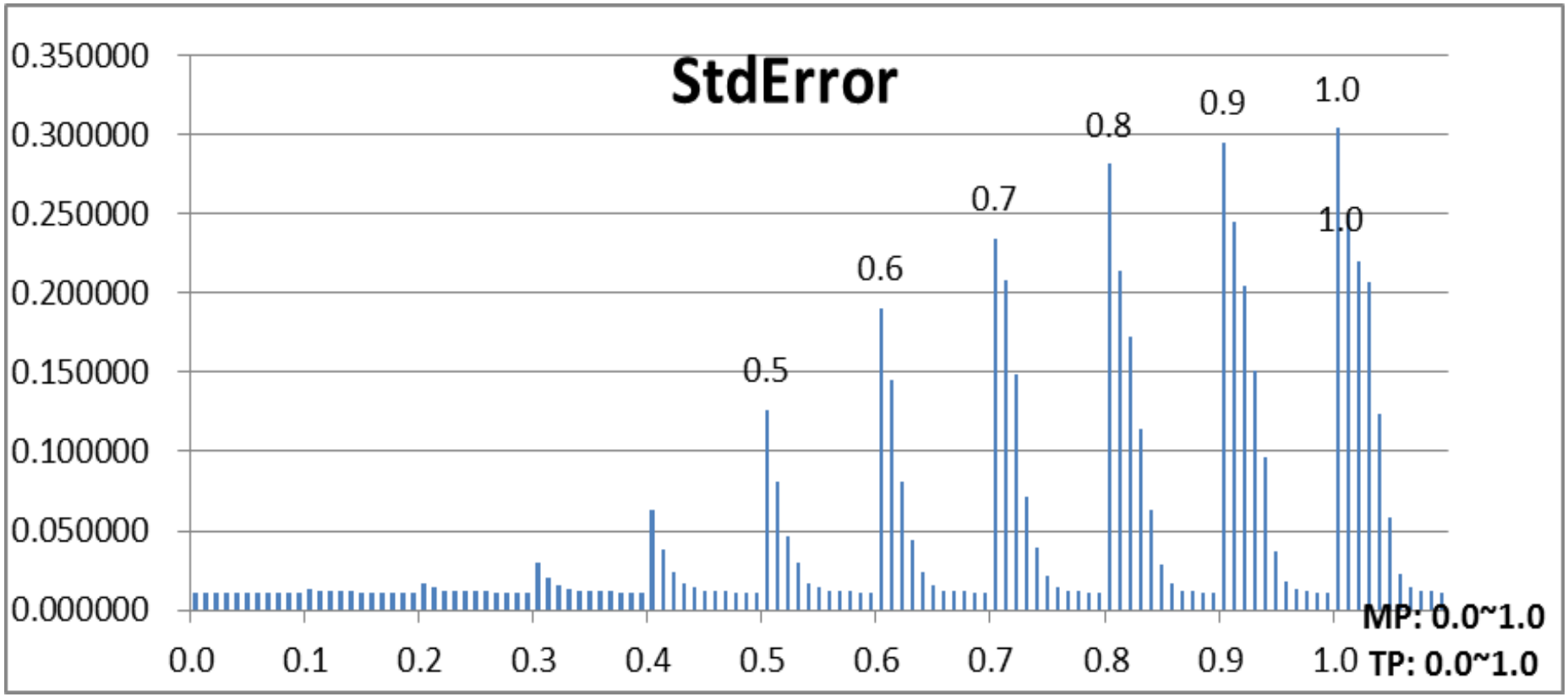
Sensitivity Test for RAS: Parameters

Parameters	Description	Set Value
duration	How many step to simulate.	10,000
worldXSize	The column size of the world.	40
worldYSize	The row size of the world.	40
GUIshots	Take a shot at the end	0
numMedia	How many media in simulation.	2
propExperts	The proportion of citizen2 in the model	0
propYES	The proportion of favoring "YES" for citizen1.	0.5
c1CHECKMEDIA	Citizen1 can access media or not.	1
c1Conform	The critical value of Opinion to change Preference	0.5
memLength	Citizen1's memory length or the capacity of storing past political information that influences vote preferences	10
c1TPMax	The maximum proportion of finding somebody to discuss politics of citizen1.	Max=Min [0.0, 1.0]
c1TPMin	The minimum proportion of finding somebody to discuss politics of citizen1.	
c1ExpertiseMax	The maximum level of expertise of citizen1.	5
c1ExpertiseMin	The minimum level of expertise of citizen1.	1
c1MPMax	The maximum proportion of accessing media of citizen1.	Max=Min [0.0, 1.0]
c1MPMin	The minimum proportion of accessing media	

Simulation Results: The Average of Agents Favoring “1”



Simulation Results: The Standard Errors of the 121 experiments



Compute Infrastructures

- Current production grids do not support multi-threaded code with high memory requirements well.
  - Suitable for runs with smaller population sizes for testing and stability analysis
- Amazon EC2 provides a suitable instance type m2.4xlarge that provides about 64GB of RAM
  - Suitable for occasional larger runs or production runs that exceed local capacity
- High-memory local servers integrated into local cloud environments.

Conclusion

- SimTaiwan shows that population-scale model can be built using commodity computing platforms
- Relatively simple programming models to utilise multi core CPUs.
- We argue that SimTaiwan is a member of a class of social science investigations that thanks to increasing compute capacity of commodity hardware are becoming tractable.
- Further work is now focusing on:
  - the development of the model logic,
  - analysis of results and validation against the statistical analyses
  - making the model accessible to our social science colleagues.

References

- Alex Voss, Andy Turner, Jing-Ya You, Eric Yen, Hsin-Yen Chen, Simon Lin, and Ji-Ping Lin, “Scalable Social Simulation: Investigating population-scale phenomena using commodity computing”, IEEE e-Science Conference, Australia Brisbane, December 7 – 10, 2010
- Frank C.S. Liu, Jing-Ya You, Yu-Ting Chen, and Jing-Lung Sun, “From Internal Validation to Sensitivity Test: How Grid Computing Facilitates the Construction of an Agent-Based Simulation in Social Sciences”, The International Symposium on Grids and Clouds and the Open Grid Forum, Academia Sinica, Taipei, Taiwan, March 19 - 25, 2011