PROCEEDINGS OF SCIENCE

Micro Discrete Events and Macro Continuous Social Outcomes: Migration Flows Analysis and Scientific Computing Challenges for Social Scientists

Ji-Ping Lin[1]
Research Center for Humnanilities and Social Sciences, Academia Sinica
128, Sec. 2, Academia Rd., RCHSS, Academia, Nankang, Taipei 115, Taiwan
E-mail: jplin@sinica.edu.tw

This paper has two goals: (1) to describe the computing process and simulation design that are not fully addressed in my recent publication(see [4]) due to page limitation of publication; (2) to present the author's research experiences and propose some thoughts regarding computing challenges that might be common for researchers in social sciences. In [4], the author proposes a theory for the micro-macro link that functions to bridge individual micro-discrete events and social macro-continuous phenomena, with simulation of immigration impact as empirical study case. The empirical study on immigration impact simulation was at first implemented with ordinary digital hardware and software settings. Using a high-end workstation, the research finds that in-memory computing and moderate overcolocking of CPUs and I/O bus are ideal solution in accelarating social computing. However, computing starts becoming a serious issue as the micro data sets grow in size or/and spatial-temporal unit under research becomes smaller. Under these circumstances, the author gradually acknowleges the importance of Grids/Clouds/HPC computing, but finds two major barriers that need to overcome. The first barrier is how to make interrelated tasks of social computing implementable within distributed storage/computing framework, while the second one is the lack of friendly computing packages and technical supports. In the context of collaboration, simulation, modeling, and data analytics, the author proposes personal thoughts regarding challenges and possible mitigation methods from the perspective of social scientist, including (1) seeking a balance between using commercial and opensoure software as a key dimension of developing theory- and evidence-based modelling and data analytics; (2) development of Grids/Clouds/HPC middleware for social science computing; (3) consolidation of collaboration and seeking for mutual understanding between social scientists and computing experts as a gateway to overcome interdependence of social computing tasks that are not implementable within distributed computing/storage system.

The International Symposium on Grids and Clouds (ISGC) 2013
March 17-22, 2013
Academia Sinica, Taipei, Taiwan

[1] Speaker

1. Introduction

Constructing a theoretical framework to explain socioeconomic macro-level phenomena that are outcomes of countless individual choice behavior is a fundamental core aim for social scientists, e.g., see [1] [2] [3]. This paper has two goals: (1) to describe the computing process and simulation design that are not fully addressed in my recent publication(see [4]) due to page limitation of publication; (2) to present the author's research experiences and propose some thoughts regarding computing challenges that might be common for researchers in social sciences. The general aim of [4] is to address the micro-to-macro link research in general, with immigration impact analysis and simulation as an example of empirical study. It proposes a theory-based, evidence-based, and implementable integrated study by combining the dimensions regarding macro conditions, micro conditions, micro outcomes, and macro outcomes, and how macro outcomes in turn affect initial macro condition within a complex system.

The fundamental thought of the constructed micro-macro link theory in [4] resembles path integrals in quantum mechanics. Within the constructed micro-macro link theory, the micro theoretical framework is based on individual discrete choice theory, while the so-called transformation rules that integrate individual discrete outcomes to continuous-macro phenomena is constructed by using existing statistical theorems and asymptotic theory. During the research, however, computing starts becoming a serious issue as the micro data sets grow in size or/and spatial-temporal unit under research becomes smaller. Under these circumstances, the author considers possible solutions; and gradually acknowleges the importance of Grids/Clouds/HPC computing, but finds some barriers that need to overcome before taking advantage of Grids/Clouds/HPC computing.

All of the aforementioned issues in social computing are essentially originating from issues embedded in data integration (including volume and variety), development of modeling, complexity of micro-foundation, velocity of analytics, and capability of providing useful intelligence. Thus, they resemble issues commonly seen in other disciplines. Physics might be the most noteworthy one. In terms of micro-macro theory development and modeling, if we look backwards at the development history of modern physics, physics has achieved the abovementioned core goal that social scientists are striving to achieve. But it has been known that it was not without controversies to achieve this goal in physics at first. The most well-known should be the long Einstein-Bohr Debate that originated from Einstein's strong objections to “Bohr plays poker”, leading to a famous Einstein quote “God does not play dice with the universe”. What so bothered Einstein about quantum theory was that it was inherently probabilistic and that quantum mechanics might serve as the end of determinism deeply held by metaphysics [5].

Let's briefly look at the past development in social sciences. Originally stimulated by the success of metaphysics and availability of computers in the 1950s-60s, quantitative revolution of social sciences did happen, as can be seen in various disciplines of social sciences like geography, economics, sociology, etc.[1][2]. Nevertheless, unlike physics that has established theoretical link and testable experiments between metaphysics and quantum mechanics, the micro-macro link in social sciences have not yet been well established. It is thus not surprising that similar to early development of modern physics, approaches using macro and micro methodology in social sciences turn out to be distinctive and irrelevant to each other for each discipline, e.g., macro versus micro in economics; structural versus individualistic in sociology.

The term the micro-macro link used in this paper is well explained in a recent review work done by [3] who provide us with a comprehensive literature review and in-depth commentaries regarding the micro-macro link studies in social sciences, with special focus on micro-foundations in sociology. As illustrated by Figure 1, the author uses the known Coleman's Scheme to explain what the micro-macro link stands for (e.g. see [2] [6]).

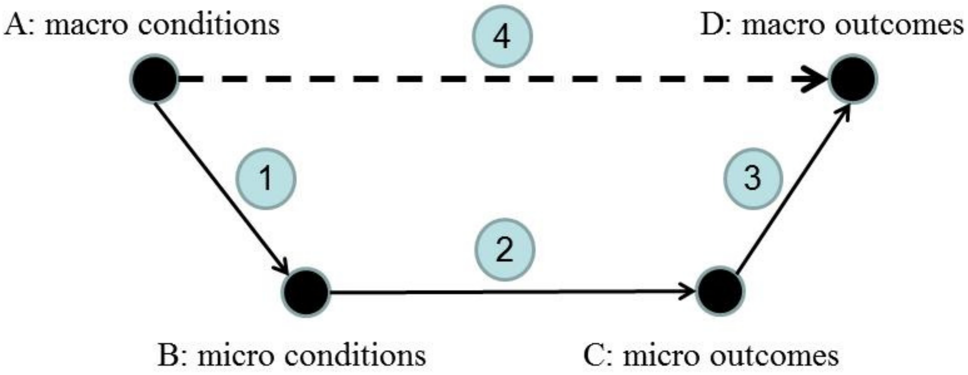


Fig. 1. The Coleman Scheme: (1) *the macro-micro link (data)*: on the basis of bridge assumptions, it aims to explore the effects of macro variables on micro conditions; (2) *the micro link (micro modeling)*: using a set of micro and macro variables, it aims to construct individual-level behavioral model; (3) *the micro-macro link (analytics)*: based on the macro-micro link and the micro link, it aims to construct transformation rules that bridge micro outcomes and macro phenomena; (4) *the macro-regularity link (outcome)*: based on observed macro regularities, it aims to associate macro conditions with macro phenomena.

In Figure 1, nodes A and D represent macro conditions and macro phenomena, with the arrow A-D referring to the propositions about observed regularities derived from associations between A and D. Thus, arrow A-D represents a macro theory. Likewise, node B represents micro conditions, serving as individual explanatory variables in accounting for micro outcomes C with individual-level model. Thus arrow B-C is called the micro link and represents a micro theory. Regarding the bilateral relationship between micro and macro, arrow A-B is academically called the link of bridge assumptions which represents a fact that micro conditions are subject to the effects of macro conditions. Arrow C-D represents propositions about the ways that actors' individual behavior generate macro outcomes, with the propositions regarding arrow C-D being called the micro-macro link and the propositions to construct the micro-macro link are called transformation rules.

Among the abovementioned four links, the micro-macro link is far less developed than the remaining three links. Because of this situation, the micro-macro link becomes the most ambiguous link but turns out the most important one in social sciences. Although some disciplines like neoclassical economics, institutional economics, evolutionary economics [3] have contributed to the construction of micro-macro link, it remains far less developed in theoretical foundation. For example, in production of collective goods [7][8], the macro-regularity, macro-micro, and micro links have been well developed, but it is lack of developing the micro-macro link. The school of standard rational choice micro model provides us with a well-established micro-foundation theoretical framework, but is criticized as being too parsimonious to establish micro-macro link, e.g., see [4].

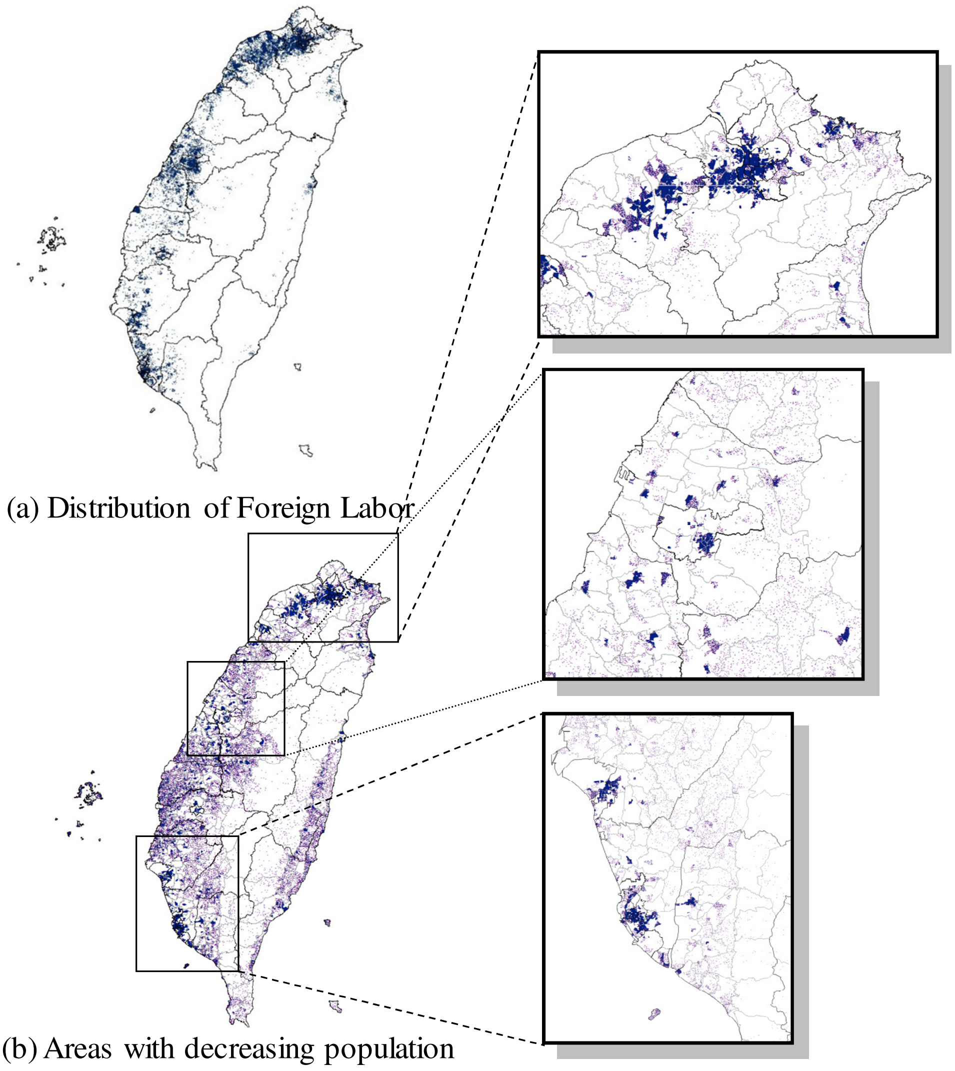
Nevertheless, it is worthy of noting that Coleman's Scheme is a highly stylized and simplified scheme. It provides us with a broad and clear picture in describing micro-macro models on one hand, but leaves issues of system complexity implicit. As a result, the micro-macro link could be viewed as a missing link in computational social sciences. Partly because of the less theoretical development in the micro-macro link and mainly because of increasing computing capacity and decreasing costs, social simulation in the past two decades has emerged as a good alternative for researches using deduction or induction. But a successful simulation must be based on a theoretical framework that not only has successfully established the micro-macro link, but also is grounded on the real-world evidence.

2.Context and Research Questions

2.1 Macro Regularities from Observed Patterns

This part briefly introduces the dynamics of a migration system by using Taiwan as an example[4][7]. The following introduction in this subsection aims to provide a systematic description related to the macro-regularity link shown in Figure 1. For a more comprehensive description, please see [9]. Because of the formation of dual-pole development system, the internal migration in Taiwan was characterized by northward-and-southward pattern in 1920s - early 1980s. Due to serious labor shortages, the government of Taiwan officially opened up the domestic labor market for low-skilled immigrants in 1992. The number of low-skilled foreign labor in Taiwan has risen to about 350 thousands by 2010, and the share of low-skilled foreign labor to native labor and to low-skilled native labor keeps at a rate of about 3.3% and 5.5% on average since the year of 2000, respectively. The period of 1996-2000 had the most noteworthy growth in terms of volume and rate that was only a few years after the aforementioned prominent transition in native destination choice preference.

Because (1) northern Taiwan, particularly the Taipei Metro, became the unique winner of internal migration in the late 1980s and early 1990s, (2) only a few years later, Taipei Metro became the most important immigrants' port of entry after the Taiwanese government opened up domestic labor market to low-skilled contract workers in 1992, and (3) Northern Taiwan (Taipei metro in particular) was observed to have a net out-migration of low-skilled blue collar native-born workers since the second half of the 1990s, such observed pattern with macro regularity inspire us to explore whether immigrants of international migration have affected Taiwan's internal migration and consequently domestic population redistribution. In short, the macro-regularity link is shown by Figure 2.



(a) Distribution of Foreign Labor

(b) Areas with decreasing population

Fig. 2. The macro-regularity link: (a) the observed concentration of foreign labor in major metropolitan areas is found to be highly correlated with (b) net out-migration of native-born population.

2.2 Research Questions

Research questions of the study are inspired by the observed patterns described in subsection 2.1 that represents the macro-regularity link. In short, the study wishes to answer the following questions:

1. Does immigration have impact on internal migration of native-born population?
2. If immigration impact exists, then to what extent does immigration affect internal migration? how the affected native-born labor respond to immigration impact?

3. If immigration volume can be regulated by policy, then how do the aggregate flows of domestic labor respond to regulated immigration level?

3. Methodology

This section describes the methodology developed to conduct immigration imapct analysis based on a simulation design proposed by the author. The simulation in the study is grounded on theory-, evidence-based, and data-driven methods that are not only pertinent to the theoretical framework of researdch in interest, i.e., migration, but are also empirically linked to the real world situation, i.e. Taiwan‘s labor market and population system. Detailed reviews regarding researh theme and eimprical findings as well as discussions can be found in [4]. The following part focuses on methodolotical issues that are not addressed within [4].

3.1 The Macro-Micro Link as Data Integration

The study chooses the period of 1996-2000 as the main research period, because immigration into Taiwan is seen to increase substantially in terms of volume and growth rate during this period. Micro data sets used to reflect micro conditions are the 1996-2000 micro data of Taiwan Manpower Utilization Surveys (MUSs). The MUS is a well-established large-scale survey conducted by Taiwan Census Bureau in May of each year since 1978, resembling the the US CPS (Current Population Survey) and the Canadian LFS (Labor Force Survey). Each MUS of 1996-2000, comprising around 60,000 individuals aged 15 and over, records abundant personal information on demographic characteristics, human capital, socioeconomic status, labor market participation and work experience, place of work and residence, labor mobility and job turnover, wage, etc.

Data sets utilized to reflect macro conditions are collected from a number of official aggregate statistics ranging from the year of 1995 through 1999, including population size, population density, employment growth, unemployment rate, household income level, geographic contiguity of labor market. They are used to control for the contexture effects of regional labor market. It is worth stressing that the measured time point of aggregate regional statistics is one year earlier than the micro data of MUSs, aiming at controlling for the causal effect of labor market condition preceding the incidence of individual migration and thus at avoiding confounding estimation results. Programming languages utilized to process micro data and to integrate micro and macro data for analyses in the subsequent part, i.e., the micro link, include Delphi and SAS.

3.2 The Micro Link as Micro Modeling

There are a number of micro models that can be chosen from the framework of rational choice theory. The author decides to choose micro models based on discrete choice theory [10] [11]. Reasons for choosing micro models based on discrete choice theory are twofold. First, individual choice in the real world is mostly discrete, i.e., either nominal or ordinal type. Second, the nested logit model derived from discrete choice theory comes with a so-called inclusive variable that has its unique theoretical merit and empirical implication, i.e., individual interaction with the remaining of the system. In brief, the nested logit model based on discrete choice theory is not only very suitable for us to establish interpersonal interactions, but also to construct individual-environment interactions with the system.

The specification of the migration micro model is a two-level nested logit model that has been widely used in the discipline of migration study, with $P_i(o)$ denoting the probability of worker *i* departing from origin *o*, and $P_i(d|o)$ representing the probability of *i* choosing *d* from the choice set *D* (the set of all possible destinations) given *i* decides to make migration from *o*. The top level is called departure model that formulates an individual migration decision between choosing "*stay*" or "*migrate*", while the bottom level being called destination choice model which formulates multinomial probabilities of choice from potential destinations once an individual decides to migrate.

Parameter estimators of the two interrelated migration micro models, departure and destination choice, are estimated on the basis of maximum likelihood estimation (MLE). Since explanatory variables in the macro-micro link are numerous and complex, parameter estimation in establishing the micro link using widely adopted numerical method like Newton-Raphson is very unstable and often leads to diverging estimations processes. Under practical considerations of (1) stabilizing parameter estimation processes, (2) ensuring convergence of parameter estimation, and (3) enhancing estimation efficiency and performance, the author adopts a line search algorithm to implement parameter estimation of the micro model.

Using line search algorithm to approach local/absolute maxima/minima has long been a crucial branch in the field of numerical optimization [12] [13]. In general, it involves four main considerations that need to be overcome: (1) to determine the best initial point of search that will enable subsequent steps of search to become implementable; (2) to compute search direction; (3) to compute the optimal step length of search along a given optimal search direction; and (4) to determine criteria of search convergence. The line search algorithm reads

```
Initialize an initial point $b^{(0)}$;
$i \leftarrow 0$;
repeat
  determine search direction $s^{(i)}$;
  determine step length $t^{(i)}$ along $s^{(i)}$;
  $b^{(i+1)} = b^{(i)} + t^{(i)} s^{(i)}$;
  $i \leftarrow i + 1$;
until $b^{(i)}$ converges to $b*$;
/* end of line search */
```

The study employs exact line search algorithm (e.g., see [12]) which ensures the total number of iterations of approaching MLE to be reduced to the minimum. Based on existing migration theories and empirical researches, the study then looks for the so-called optimal model. The optimal model in the study refers to the model in which the estimated coefficients of explanatory variables are not only statistically significant, but are also required to be substantively meaningful and to be consistent with existing theories. Consequently, it may need more than one hundred parameter estimations to get this optimal model, with each estimation requiring about 15~30 iterations to obtain the MLE of micro model's parameters.

3.3 The Micro-Macro Link as Foundation of Analytics Development

3.3.1 Transformation Rules

Migration of individuals is a joint outcome triggered by various socioeconomic and cultural forces. If numerous probabilities of individual migration decision are regarded as countless particles triggered by a gigantic collider, continuous macro migration outcomes that exhibit macro regularities, such as aggregate volume and rate of in-migration, out-migration, and net migration, can be obtained in a way through aggregating discrete individual migration probabilities. This thought is fundamentally similar to the method of path integrals in quantum mechanics that uses a sum over an infinity of possible trajectories to compute a quantum amplitude.

The transformation rules (TRs) constructed in the study is quite straightforward and simple by using existing known statistical theorems and asymptotic theory. The study specifies two Rules to formulate statistical distributions associated with macro migration outcomes. Note that the assumption of identical distribution is not required in TRs. They are specified as follows:

Transformation Rule 1

Aim: aggregation of individual discrete random choices to a continuous random flow;

Methods:

1. according the weak Central Limit Theorem (*weak CLT*) in statistics, the asymptotic distribution of any linear combination of *n* independent Bernoulli random variables with parameters $\{P_i \mid i\text{=1…n}\}$ in (0,1) is normally distributed, as *n* approaches infinity;
2. according to the *weak CLT*, the asymptotic distribution of any linear combination of n independent multinomial random variables with parameters $\{(P_{ki}) \mid k\text{=1…K}, i\text{=1…n}\}$, $\sum P_{ki} = 1$ over *k* for a given *i* is multinormally distributed;

Transformation Rule 2

Aim: aggregation of continuous random flows to a single continuous random flow;

Method: any linear combination of normal distributions is still normally distributed.

3.3.2 Aggregate Statistical Distributions:

The micro-macro link is developed based on the principle of TRs 1 and 2. For detailed theoretical derivation about the distribution function of macro flows in migration network, see [4]. Given TRs 1 and 2, we have

1. Regarding aggregate out-migration, the distribution function of a macro outflow originating from a given geographic area *i* is normally distributed, reading as

$$Y_i^{+} \overset{ind.}{\sim} N(\mu_i^{+}, \sigma_i^{+2}), \forall i \tag{1}$$

2. Regarding aggregate in-migration, the distribution function of a macro inflow into a given geographic area *j* is also normally distributed. It is jointly determined by departure and destination probabilities of choosing this given area for individuals who decide to migrate outside this given area, with a form

$$Y_{+j}^{+} \sim N(\mu_{+j}^{+}, \sigma_{+j}^{+2}), \forall j \tag{2}$$

3. Thus, net migration (= (2) – (1)) is still normally distributed.

The underlying importance of this part lies not only in providing us with aggregate statistical distributions which enable us to predict the likelihood of a macro regularity, but it also provides us with sufficient complexity within a system in the sense that the derived aggregate statistical distributions are inherently embedded with very complex micro behavior models. In other words, we are able to explain and predict macro regularities that interest us from very tiny microscopic events without the need to simplify conditions and assumptions associated with the micro model.

4. Simulation Design and Digital Infrastructure

4.1 Simulation design

The implementation of the proposed micro-macro link's theoretical framework was at first realized by a program coded with Gauss. Its kernel consists of two main procedures, Proc DepaAnalysis and Proc DepaDestAnalysis, which serve as the main analytics of the micro-macro link for departure and destination choice analyses. The author will provide the source code to interested researchers on request. To explore immigration impact, the study designs a simulation framework that is grounded on the aforementioned four stylized links, i.e., the macro-micro link (Link 1), the micro link (Link 2), and the micro-macro link (Link 3), with a core goal to examine the validity of the proposed research questions that are inspired by the macro-regularity link (Link 4). To reflect the dynamics of system evolution, an endogenization link is also taken into consideration in designing simulation framework, as illustrated in Figure 3.

The main reason for adding an endogenization link is due to the fact that traditional Coleman's Scheme does not reflect mutual causality. Adding endogenization link will provide us with a complete causality loop. In other words, macro outcomes should not be treated as the end of observed phenomena, rather they often serve a role in endogenizing the macro conditions of a micro model in turn. In other words, it serves a role in affecting initial conditions of an existing complex system. To account for the evolution and complexity within such system, the author proposes a revised Coleman Scheme in which Link 5 is added to represent the endogenization link from node D to node A. On the basis of the revised Coleman Scheme, the design of the simulation for immigration impact analysis is demonstrated by the algorithm below that reads as:

```
/* begin of migration impact on population system */
t ← 1;
Initialize MigrationLevel by the observed migration level at t;
Initialize _scalelevelcell; /* e.g. 0.50 */
Initialize _scaleincrementlevel; /* e.g. 0.05 */
ScaleLevel ← _scalelevel;
/* e.g., _scalelevel = -0.5 decrease by 50% */

do while ScaleLevel < _scalelevelcell
  begin
    /* Link A→B */
    Construct the macro-micro link;

    /* Link B→C */
    Construct the micro link;

    /* Link C→D */
    Construct the micro-macro link;
    Obtained expected macro outcomes;

    /* Node D */
    Output expected aggregate out-migration and in-migration;
    Output difference between observed & expected macro outcomes;
    Obtain migration impact on population distribution;

    /* Link D→A */
    Adjust population level of all origins by net migration;
    Adjust population level of all destinations by net migration;
    Adjust population level in the macro-micro link;
    Adjust population density in the macro-micro link;
    ScaleLevel ← ScaleLevel + _scaleincrementlevel;
    t ← t + 1; /* next evolution */
  end; /* while loop */
Synthesize migration impacts from various ScaleLevel;
Assess migration impact on population redistribution;
Obtain evolution of population redistribution;
/* end of migration impact analysis */
```

In brief, findings from the simulation of immigration impact on population redistribution are as follows: the macro regularities (Link 4) suggest that areas with high immigration concentration are associated with net out-migration of internal migrants, while findings from the micro model (Link 2) on the basis of explanatory variables (Link 1) suggest that immigration exhibits significant supplementary effect for the professionals, but shows tiny or insignificant pushing effect for blue-collar labor. As a result, the micro-model approach turns out to be inconsistent with the observed macro pattern.

However, immigration impact analysis, based on the micro-macro link derived from transformation rules (Link 3) and the endogenization link (Link 5), suggests that the simulation results resemble observed pattern of native labor "flight", mostly native-born migrants with less education, from major immigration "port of entry" on both sides of Taiwan and the United States. In addition, it further suggests the mechanism of immigration impact: the observed native-born worker “flights” from high immigration concentration areas essentially are not mainly triggered by immigration pushing force, instead it is mainly resulted from the negative impact of immigration on the in-migration that outweighs the corresponding negative impact on out-migration for native labor.

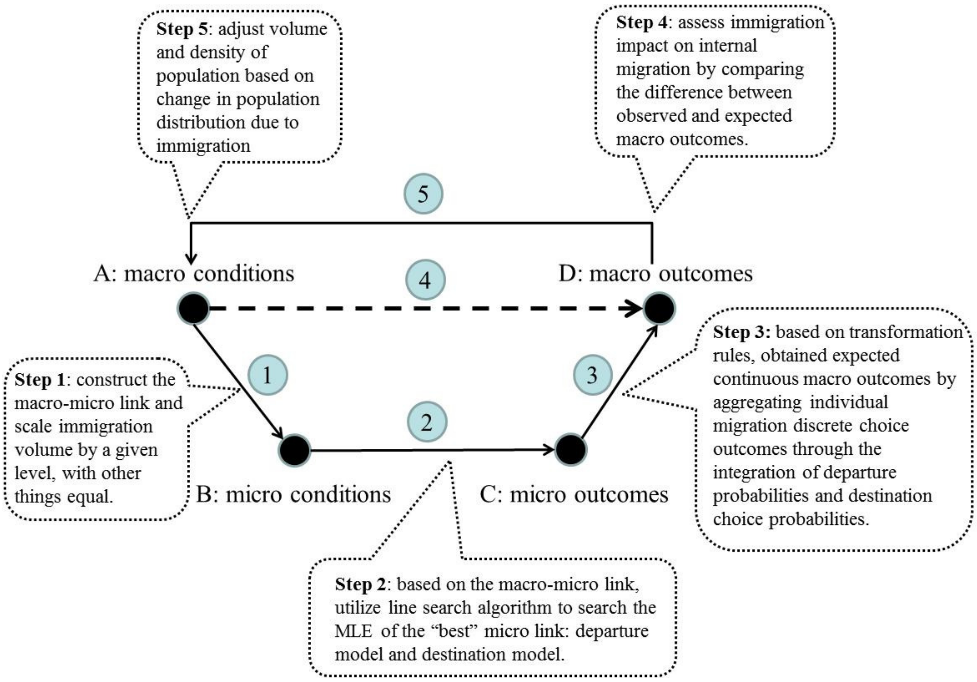

Step 5: adjust volume and density of population based on change in population distribution due to immigration
Step 4: assess immigration impact on internal migration by comparing the difference between observed and expected macro outcomes.
5
A: macro conditions
4
D: macro outcomes
Step 1: construct the macro-micro link and scale immigration volume by a given level, with other things equal.
1
3
Step 3: based on transformation rules, obtained expected continuous macro outcomes by aggregating individual migration discrete choice outcomes through the integration of departure probabilities and destination choice probabilities.
2
B: micro conditions
C: micro outcomes
Step 2: based on the macro-micro link, utilize line search algorithm to search the MLE of the “best” micro link: departure model and destination model.

Fig. 3. (1) the revised Coleman’s Scheme with *the endogenization link* from node D to node A being added and (2) steps of simulation designed for immigration impact analysis.

4.2 Digital Infrastructure

Because exact line search algorithm that is applied to MLE parameter estimation of micro modeling is very computing-intensive, computer hardware and software application tools used in study will matter when data for estimation grow in size. To enhance computing performance, main components of the computer for study are appropriately overclocked, including CPU, I/O bus, and DRAM. The hardware infrastructure in the research is a i7-3960x-based workstation with 64GB DDR3 1600 memory. To increase computing efficiency, its storage adopts RAID0 with two SATAT3 256GB OCZ Vertex3 MaxIOPS SSDs (W:550Mb/sec, R: 950Mb/sec with I/O bus overclocking and disk cache enabled). In addition, I also use moderate hardware acceleration method (i.e. overclocking of CPU/DRAM/IO bus) to shorten computing time of model parameter MLE estimation and simulation.

Since making full utilization of computer internal memory will boost computing performance and shorten total computing time substantially, the computing environments, including OS and application software, are all x64-based. As for software infrastructure, the OS is Win7 Enterprise x64, programming language is Embarcadero RAD Studio XE2 x64, and APs are SAS 9.3 x64 and ArcGIS 10. Using dynamic memory allocation method to control computer internal memory, I wrote an IMSL-like library in Pascal long time ago which allows me to do generalized matrix manipulation and numerical optimization using line search algorithm as long as having enough internal memory. Dynamic memory allocation method is implemented by utilizing pointers to request a block of internal memory from the OS for data storage and in-memory computing. The allocated size of internal memory will vary by the size of data which is loaded into memory to accelerate computing. When a computing task is finished, the requested memory will be disposed and becomes available for other applications. To save time, I recompile it using Embarcadero RAD Studio XE2 x64 and it functions well and can fully utilize my computer's internal memory. SAS 9.3 x64 is very stable and easy to use and can take full advantage of internal memory and all cores of CPU. ESRI ArcGIS 10 is used to visualize computing and simulation results.

5. Concluding Remarks and Discussion

The paper at first specifies the undocumented detailed processes of data integration, construction of theory-, evidence-based, and data-driven modeling, development of analytics, computing processes, and design and scenarios of a simulation in [4]. More importantly, it aims to share experiences and thoughts while conducting this research from the perspective of a social science researcher.

First of all, micro-macro modeling based on rational choice theory has long been criticized for its weak links with empirical research. To a certain extent, the transformation rules proposed in the study might have contributed to strengthen these weak links and help development of analytics without the need to sacrifice the complexity of micro-foundations. In the past two decades, simulation method such as the known agent-based modeling (ABM) has emerged as a new alternative in exploring the nature of a complex system, e.g., see [14][15][16]. From the perspectives of analytic social scientists, simulation of ABM adopts a strategy of "from top to bottom and back again", while simulation in the study adopts that of "from bottom to top and back again".

The computing of simulation on immigration impact was at first implemented with ordinary digital hardware and software settings. With fast fall in hardward prices and gradual availability of 64-bit OS and applications in the past four years, the research benefits a lot from utilizing high-end personal digital infrastruture that was not affortable for or accessible to an ordinary social scientist. Using high-end personal digital infrastructure, the research finds that in-memory computing and moderate overcolocking of CPUs and I/O bus are ideal solution in accelarating social computing and simulation. However, computing starts becoming a serious issue as the micro data sets grow in size or/and spatial-temporal unit under research becomes smaller. Under these circumstances, the author gradually acknowleges the importance of Grids/Clouds/HPC computing [17][18], but finds some barriers that need to overcome from the perspective of social computing.

The first challenge arises from computing complexity of model building and development of analytics under the conditions of without sacrificing micro-foundation complexity, as micro data sets grow in size. Take computing tasks of [4] as an example, it has a size of population about 23 million and an annual migration rate of 9 percent (that is, 2.07 million migrants). If geographic units are its 23 counties, then data for the destination model estimation requires 45.5 million (=2.07million*22) records and the size of data is about 6GB in ASCII format. The total area of Taiwan is about 32,000km^2. In light of growing availability of huge micro data sets which contain precise individual geographic location information in Taiwan, if the analytics adopt a geographic unit with a "resolution" of 100m*100m, then 3.2million geographic units are required for the destination model estimation, suggesting that we need to generate 6,623,998 million (=2.07million*(3.2million-1)) records. Because it is 145,582 times the size of using traditional county for analysis, the required data size will jump to about 853TB (=6GB*145,582/1024).

Under the aforementioned situation, the computing associated with parameter estimation of micro modeling becomes a big challenge. A traditional way to mitigate this sort of computing challenge is to sacrifice the complexity of micro-foundations through a way like deleting most explanatory variables from the micro modeling process. Nevertheless, such way of sacrificing micro-foundation complexity has a serious drawback that analytics and simulation derived from this sort of micro modeling will become less informative and even pointless. In this regard, Grids/Clouds/HPC computing inevitably must be taken into consideration for advanced social computing.

Given Grids/Clouds/HPC computing could be a good alternative, another big challenge arises if we want to maintain the complexity of micro-foundation using big micro data sets to conduct micro modeling. This challenge lies in the fact that social events and outcomes are barely independent to each other, suggesting that the computing task of estimating optimal micro model can't be done by partitioning the big micro data sets to smaller ones and then sending them to distributed computing system to do computing and analysis independently. In short, because components of a socioeconomic system are interdependent, Grid computing may not work well as data for computing of social modeling grow in size unless we sacrifice the complexity of micro model.

During conducting the research [4], the author ever tried the possibility of using other possible alternatives, i.e., Cloud computing and HPC, in an attempt to conduct immigration impact simulation based on the aforementioned big micro data. To the author's limit knowledge, the author finds that commercial Cloud computing like MS Azure could be taken into consideration, but is too expensive to use for ordinary researcher in social sciences. Unlike natural and life sciences that have a long history of using HPC for advanced computing, using HPC for social science study is barely seen. Take HPC in Taiwan as an example, the author finds that the main barrier of utilizing HPC is not due to the problem of accessibility and affordability. Rather, it is mainly due to the lack of package and middleware developed for social computing.

In terms of collaboration, the author has two suggestions from the perspective of social sciences. First, we need to seek a balance between commercial software and open source tools. Commercial software allows us to manipulate computing facility more easily than open source tools. However, it is pricey to use commercial software that in turn tends to hamper collaboration. Open source tools are great for social computing, but the author finds that researchers in social sciences generally are not familiar with UNIX environments. Using UNIX-based open source tools tends to hamper collaboration with researchers in social sciences.

In the end, the author summarizes personal thoughts regarding challenges and possible mitigation methods from the perspective of social scientist. First, seeking a balance between using commercial and open soure software will serve as a key dimension of developing theory- and evidence-based modelling and data-driven predictive analytics. Second, seeking for mutual understanding between social scientists and computing experts will serve as a gateway to overcome interdependence of social computing tasks that are not implementable within distributed computing/storage system. Third, if we acknowledge the importance of utilizing emerging big social data to improve data-driven predictive analytics, computing experts are expected to contribute to develop friendly middleware of Grids/Clouds/HPC for social sciences. Finally, to enhance computing literacy among social scientists and to promote social literacy of computing experts are the greatest challenge. Overcoming such challenge will serves as the gateway toward multi-disciplinary collaboration.

References

1. T.C. Schelling, 1978, *Micromotives and Macrobehavior. Norton, New York*
2. J.S. Coleman, 1987, *Microfoundations and macrosocial behavior. In: Alexander, J.C., Giesen, B., Münch, R., Smelser, N.J.(eds.) The Micro-Macro Link, pp. 153-173. University of California Press, Berkeley, CA*
3. W. Raub, V. Buskens, M.A.L.M. Van Assen, 2011, *Micro-Macro Links and Microfoundations in Sociology. In: The Journal of Mathematical Sociology,* vol. 35, pp.1-25

4 J.P. Lin, 2013, *"Are Native "Flights" from Immigration "Port of Entry" Pushed by Immigrants? Evidence from Taiwan" in Fong, Chiang, and Denton (eds) Immigrant Adaptation in Multi-Ethnic Societies , New York: Routledge*

5. A., Diem-Land, 2008. *Spooky Physics: A Brief Introduction to the Einstein-Bohr Debate. MSAC*
6. J.S. Coleman, 1990, *Foundations of Social Theory. Belknap Press of Harvard University, Cambridge, MA*
7. M. Olson, 1965, *The Logic of Collective Action (2nd ed.). Harvard University Press, Cambridge, MA*
8. C.F. Camerer, 2003, *Behavioral Game Theory: Experiments in Strategic Interaction. Russell Sage, New York, NY*
9. J.P. Lin, 2012, *Tradition and Progress: Taiwan's Evolving Migration Reality. Migration Information Source, Migration Policy Institute, Washington D.C.* http://www.migrationinformation.org/Profiles/display.cfm?id=877
10. D. Mcfadden, 1974, *Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka, P. (ed.) Frontiers in Econometrics. Academic Press, New York*
11. M. Ben-Akiva, S. Lerman, 1985, *Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press, Cambridge, MA*
12. R. Fletcher, 2000, *Practical Methods of Optimization, John Wiley & Sons*
13. J. Nocedal, S. Wright, 2006, *Numerical Optimization(2nd ed.). Springer*
14. N. Gilbert, P. Terna, 2000, *How to Build and Use Agent-based Models in Social Science. Mind & Society, vol 1(1), pp. 57-72*
15. S.H. Chen, L. Jain, and C.C. Tai, 2005, *Computational Economics: A Perspective from Computational Intelligence, Computational Intelligence and its Application Series, (series editors: Lakhmi Jain and Colin Fyfe), Idea Group Publishing*
16. F. Frank Liu, S. Lin, J.Y. You, Y.T Chen, J.L. Sun, 2011, *Form internal validation to sensitivity test: How grid computing facilitates the construction of an agent-based simulation in social sciences*. Proceedings of Science

17. S. Lin, E. Yen, 2010, *Data Driven e-Science: Use Cases and Successful Applications of Distributed Computing Infrastructures*, Springer.

18. V. Alex, J. Kmunicek, L. Matyska, M. Paganoni, and S. Lin, 2010, *The EUAsiaGrid Roadmap: Paths to a Sustainable, Persistent e-Infrastructure in the Asia-Pacific Region. In: Simon C. Lin and Eric Yen (eds.) Data-Driven e-Science: Use Cases and Successful Applications of Distributed Computing Infrastructures.* Proceedings of the 2010 International Symposium on Grid Computing. Springer