Consistent Population Synthesis With Multi-Social Relationships Based on Tensor Decomposition

Peijun Ye, Fenghua Zhu, Samer Sabri, and Fei-Yue Wang

***Abstract*— Social relationships have a strong influence on individual travel behavior and, consequently, on travel demand. However, most current literatures on population synthesis, which is the fundamental building block of disaggregated travel demand forecasting and agent-based traffic simulation, only considers the household impact. This paper makes two contributions in this regard. First, a methodological issue is identified: the existence of multiple social relationships (e.g., a dual set of constraints from social institutions or structures) makes it more difficult to generate a consistent synthetic population, meaning that this population satisfies constraints from more than one type of social organizations. A tensor decomposition method is then proposed to generate a consistent population with multi-social relationships. To our knowledge, this is the first time that this type of methodological issue has been addressed. Our sample-based method constitutes an improvement compared to existing approaches in that it can respect constraints from multiple social organizations without reducing accuracy. A numerical test concerning individual, household, and enterprise, using Chinese national population and economic census data, indicates that the new method can lead to stable and relatively small errors in total. The source code is available from https://github.com/PeijunYe/MulSocPopSyn.git.**

***Index Terms*— Population synthesis, multiple social relationships, tensor decomposition, agent-based simulation.**

I. Introduction

AGENT-BASED simulation has become an indispensable approach to transportation research as well as the study of other complex human systems, as it facilitates system dynamics analysis and management strategy evaluation with much lower cost [1]–[4]. Agent-based models of transportation attempt to forecast travel demand and traffic patterns by simulating the travel mechanisms of relevant human participants. Usually, the underlying assumption of this approach is that travel behavior derives from the conduct of daily activities at the person (or agent) level. This “activity-based micro-simulation” aims to investigate the emergence of systematic traffic phenomena by modeling individual transfer among different places in which daily activities are completed sequentially. Evidently, a synthetic population lies at the foundation in this simulation system. The quality of the synthetic population has a serious impact on the credibility and reliability of the simulation results. To date, this sort of synthetic population has been applied into many representative systems, such as FAMOS [5], CEMDAP [6], ALBATROSS [7], ILUTE [8], [9], TASHA [10], TRANSIMS [11], [12] and so on.

In population synthesis, relationships among individuals are of great significance in modeling travel behavior at a personal level. This is probably because social groupings underlie individual transportation decisions and usually cannot be ignored. Typical social relationships derive from people’s various roles in different organizations like corporations, schools, and perhaps the most representative—households. These connections tie individuals together, and will undoubtedly trigger particular travel behaviors in their daily life. Many micro-simulation models have considered the influence of households. This is because census data, which is typically used as the input of population synthesis, provides statistical information about both individuals and households. However, given the diversity and increasing complexity of social activities, it is essential to consider additional social relationships beyond households. For example, the geographic locations of corporations and schools determine the routine destinations of many affiliated individuals. These individuals may account for a large proportion of urban travel in rush hours. The question here is how to generate a synthetic population that conforms to constraints from all those social relationships. To our knowledge, only a report from the Research Triangle Institute, USA has explicitly concerned this issue [13]. Unfortunately, it does not provide a general method that seeks to handle this problem. This has motivated us to introduce the problem of generating a synthetic population with multi-social relationships and develop a general efficient method for large-scale synthesis. Overall, the major contribution of this paper is twofold: 1) we provide a formal treatment of the problem of multiple social relationships, where the objective is to construct a population that is consistent with all the constraints from not only the household level but also other types of social organizations; 2) we propose a tensor decomposition method to generate such a consistent population with multi-social relationships. This method keeps the correlations among population attributes provided by disaggregate samples and minimizes the inconsistencies between individual and multi-social relationship levels.

Manuscript received January 30, 2019; accepted April 14, 2019. Date of publication June 21, 2019; date of current version May 1, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1004803, in part by the National Natural Science Foundation of China under Grant 61603381, Grant 61533019, and Grant U1811463, in part by the Grant of China Scholarship Council under Grant 201704910284, and in part by the Early Career Development Award of the State Key Laboratory for Management and Control of Complex Systems (SKLMCCS) under Grant 20190207. The Associate Editor for this paper was L. Li. *(Corresponding author: Fei-Yue Wang.)*

P. Ye, F. Zhu, and F.-Y. Wang are with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: peijun.ye@ia.ac.cn).

S. Sabri is with the Computer Science and Computational and Cognitive Neuroscience Lab, University of California at San Diego, San Diego, CA 92093 USA (e-mail: samrsabri@gmail.com).

Digital Object Identifier 10.1109/TITS.2019.2916867

The remainder of the paper is organized as follows. Since most of current studies only consider households and individuals, Section II gives a brief review of the methodology in this field. Section III demonstrates the synthetic population model with multiple social relationships, and includes the details of proposed tensor decomposition method. Section IV presents Chinese nationwide census data that is used in our experiments, and analyzes the results compared with other methods. Finally in Section V, the paper concludes with some additional discussions, and notes potential upcoming work.

II. Problem Statement and Literature Review

The ultimate objective of population synthesis is to acquire an artificial individual-level dataset in full compliance with the statistical characteristics of various input data. The artificial dataset, each record of which reflects one or a group of entities, must be the 'best' estimates of the actual data revealed by available aggregate information. Generally, an individual record is depicted by several personal characteristics according to a specific profile (e.g., gender, residential province or state). The synthetic population data set must conform to the actual statistical marginal (contains only one variable) and partial joint (contains several variables but less than all) frequencies published officially. These frequencies can be converted into distributions by dividing them by total population size. We will therefore use both terms interchangeably. Similarly, social attributes can be also represented by variables such as *Household Type* (such as "census family" or "economic family" defined in Canadian census), *Living Area*, *Type of Residence* (urban, town, or rural), *Number of Members*, *Car Ownership* etc.. Therefore, the synthetic population needs to be consistent with the marginal and partial joint distributions from social organizations as well. To our knowledge, most current literature has either focused on population attributes only, or taken households into account without other relationships [14]–[17]. To limit the scope of this paper, the main methodologies that explicitly consider household attributes are reviewed in the following. For a broader review, we refer the reader to Lenormand and Deffuant [18], Lovelace and Dumont [19], and Ye et al. [20] of existing techniques and their performances.

Currently, two categories of approaches have been developed to tackle household and individual synthesis. The first type—population assignment—begins with the generation of household and individual entity pools. Based on these two pools, individuals and households are matched so that the household can be established. To date, the most representative application comes from Wheaton et al., who used Census Bureau TIGER (Topologically Integrated Geographic Encoding and Referencing) data, SF3 (Summary File 3) and PUMS (Public Use Microdata Sample) of the USA to synthesize the basic dataset of 50 states and the District of Columbia in the year 2000 [13]. For social relationship formulation, they further referred to NCES (National Center for Education Statistics) public and private school data for 2005-2006, STP64 (US Census Bureau Special Tabulation Product 64), InfoUSA Business Counts, SF1 (US Census Bureau Summary File 1), HSIP (Homeland Security Infrastructure Program) database and Group Quarters Age Distributions to construct schools, corporations and group quarters. Gargiulo et al. presented an iterative method to generate statistically realistic population of households [21]. Their method is sample-free which means the generation uses aggregated census data only. Barthelemy and Toint proposed another sample-free household selection method based on entropy maximization and Tabu search, and they adopted it to Belgian synthetic population [22]. Huet et al. created households in French municipalities to study dynamics of labor status and job changes [23]. Ma and Srinivasan proposed a fitness-based approach in 2015. They re-weighted each household sample record by improving the fitness of total constraints [24]. Most recently, Nam Huynh et al. synthesized the population of New South Wales [25]. They assigned the basic population according to the household-individual mapping relations. Anderson and Farooq modeled the population assignment as a k partite-graph problem [26]. Their algorithm dynamically maintains a cost matrix and traverses every node pair to search the optimal solution. Limited by computational resources, such approach may probably not suitable for large-scale scenarios.

The second type—distribution fitting—integrates the individual and household attributes in a unified whole. Such approach focuses on the constituent of each household type, and estimates the joint distribution of the two level attributes. By constructing households based on the distribution, total synthetic population can be realized via generating individual members of every household according to its constituent. Melhuish et al. adopted three linked convergence algorithms to create synthetic households of Australian capital territory by adjusting the household weights [27]. In each round, the procedure involved an overall evaluation of all target variables and a multi-dimensional search for convergence by changing a pair of weights in a positive and/or negative direction. Ballas et al. used a deterministic reweighting approach to generate households of small areas in Great Britain [28]. In their study, the weights were fitted on the basis of British Household Panel Survey, and applied to compensate for non-responding households and those individuals in a responding household who failed to give a full investigation. Guo and Bhat combined household and individual distributions, and designed a recursive procedure to merge marginal tables of each mutual attribute [29]. Their procedure computed the merged distribution by sampling from household and individual distributions. Pritchard used individual samples as the seed and household distributions as constraints to synthesize a population via Iterative Proportional Fitting [9], [30]. Auld and Mohammadian developed another technique to determine how both household- and person- level characteristics can jointly be used as controls when synthesizing populations [31]. They introduced the Baysian method to assemble the constraints from two levels dynamically. Another technique, called Iterative Proportional Updating (IPU), was proposed by Ye et al. [32]. Their algorithm aimed at iteratively adjusting and reallocating weights among households of a certain type until both household- and person- level attributes were matched. Sun et al. synthesized population for Singapore, where he first

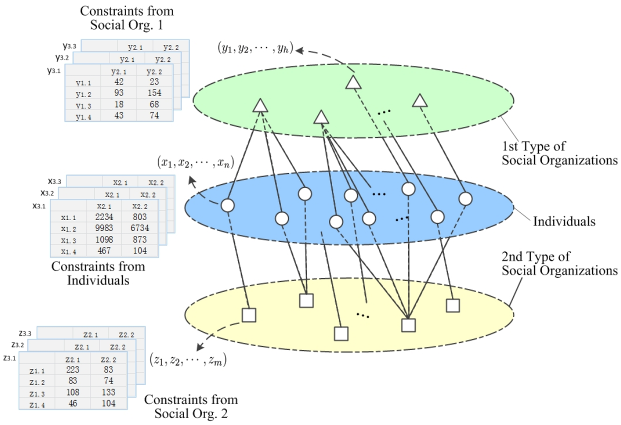


Fig. 1. Synthetic population with multiple social relationships.

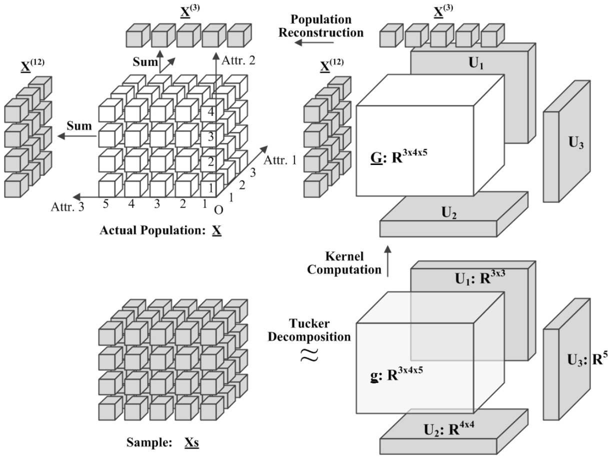


Fig. 2. An example of 3 × 4 × 5 tensor decomposition.

generated the household and then its members [33]. This 'top-down' method was not applicable when dealing with multiple social relationships, since each social level would obtain a separate population set.

In summary, most existing approaches are applied in the household-individual scenarios and are difficult to be extended to more levels, especially for large-scale computation. For the household-individual case, classic methods (such as fitness-based method, combinatorial optimization) are able to improve the fitness of household frequencies, while evaluating the individual fitness by converting the household into individuals according to its member types. However, if controls from two (or more) social relationships included, say the household and enterprise, they may fail to compute the enterprise's fitness since the relations between such two organizations are not explicit. Therefore, to generate an integrated population system with multiple social organizations, we seek to directly assign the person to each type of social entities.

III. Synthetic Population With Multi-Social Relationships

To model multi-social relationships, it is required to expand the current household-individual bi-levels to multiple levels. For simplicity, we concentrate on three levels of constraints: individual and two social relationships (e.g., household and enterprise). Please note that our formulation can easily be expanded to more social constraints as well. Our restricted problem is illustrated in Fig. 1. Let $\underline{X} = (x_1, x_2, \cdots, x_n)$ be the individual attributes we care about in the population (such as *Gender*, *Age*, etc.). These attributes can respectively take on $I_1, I_2, \cdots, I_n \in \mathbb{N}$ possible values. $\underline{Y} = (y_1, y_2, \cdots, y_h)$ and $\underline{Z} = (z_1, z_2, \cdots, z_m)$ are the social relationship attributes of interest (such as *Household Type*, *Household Residential Province*, *Enterprise Type*, *Enterprise Scale*, etc.) with $J_1, J_2, \cdots, J_h$ and $K_1, K_2, \cdots, K_m$ possible values. A distribution over individual attributes will map each possible set of characteristics $(x_1, x_2, \cdots, x_n)$ to a non-negative integer that represents the number of individuals in that type. This frequency distribution is represented by a multi-dimensional array called a tensor, and is denoted by $\underline{X} \in \mathbb{N}^{I_1 \times I_2 \times \cdots \times I_n}$. Similarly, $\underline{Y} \in \mathbb{N}^{J_1 \times J_2 \times \cdots \times J_h}$ and $\underline{Z} \in \mathbb{N}^{K_1 \times K_2 \times \cdots \times K_m}$ are tensors for social relationships. In the rest of this section, we introduce our tensor decomposition method, which involves two steps. In the first step, described in part III-A, individuals and social organizations are synthesized respectively. In the second step, described in part III-B, individuals are assigned to each social organization so that the final synthetic population is achieved.

A. Basic Population and Organization Synthesis

The first step is to generate a synthetic population and social organizations. Basically, for the individual level and each type of organizations, we are facing the same problem. Current methods are categorized into two kinds: sample-based and sample-free ones (see Lenormand and Deffuant [18], and Ye et al. [20]). For the case where disaggregate samples are accessible, we recommend using the sample-based methods because the correlations among attributes can be achieved more accurately via samples. Traditional sample-based methods are Iterative Proportional Fitting (IPF) and Combinatorial Optimization (CO) [12], [34], [35]. However, the former might suffer from zero element problem, whereas the latter does not retain the correlation structure at all. Motivated by this, here we give a new sample-based method to generate basic population.

Generally, when the sample is available, we have two types of data sources. The first one is the sample distribution $\underline{X_s} \in \mathbb{N}^{I_1 \times I_2 \times \cdots \times I_n}$. Same as the unknown actual population denoted by $\underline{X} \in \mathbb{N}^{I_1 \times I_2 \times \cdots \times I_n}$, it is a full dimensional tensor. For example, suppose we have 3 attributes to study ($n = 3$) and they have 3, 4 and 5 categories respectively. The sample and actual population can be represented as $\underline{X_s} \in \mathbb{N}^{3 \times 4 \times 5}$ and $\underline{X} \in \mathbb{N}^{3 \times 4 \times 5}$, which look like the left side of Fig. 2. The gray boxes stand for already known individual frequencies and the white boxes stand for the unknown ones.

The synthesis begins with Tucker decomposition of $\underline{X_s}$:

$$\underline{X_s} = \underline{g} \times_1 U_1 \times_2 \cdots \times_n U_n \tag{1}$$

where $\underline{g} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_n} (r_n \leq I_n)$ is the core tensor and $U_n \in \mathbb{R}^{I_n \times r_n}$ is the factor matrix. $\times_n$ is the n-mode product of the tensor and matrix. For the 3-dimentional example, the bottom right sub-figure in Fig. 2 illustrates the decomposition result. Here we set $r_i = I_i$ in Tucker decomposition, thus the factor matrices are square. The 1-mode product of $\underline{g} \times_1 U_1$ is calculated as: 1) unfold $\underline{g}$ to a 3×20 matrix, denoted as $\underline{g}_{(1)}$ (called the 1-mode unfolding/flattening/matricization). That is arranging the 4 horizontal (3×5) slices one by one; 2) compute $U_1 \cdot \underline{g}_{(1)}$ and then reversely fold the result matrix as a $3 \times 4 \times 5$ tensor. Generally, if a tensor is $r_1 \times \cdots \times r_n$ dimension, and a matrix is $I \times r_m$ dimension, their m-mode product is a $r_1 \times \cdots \times r_{m-1} \times I \times r_{m+1} \times \cdots \times r_n$ tensor. Specifically, if the matrix degenerates to a vector, that is $I = 1$, the m-mode product will be a $r_1 \times \cdots \times r_{m-1} \times r_{m+1} \times \cdots \times r_n$ tensor, which means dimension m collapses.

Mathematically, Tucker decomposition tries to solve the optimization problem:

$$\min_{\underline{\hat{X}_s}} \parallel \underline{X_s} - \underline{\hat{X}_s} \parallel$$

where

$$\underline{\hat{X}_s} = \underline{g} \times_1 U_1 \times_2 \cdots \times_n U_n$$

There are classic algorithms for this problem, such as Higher Order Singular Value Decomposition (HOSVD) and Higher Order Orthogonal Iteration (HOOI) [36]. To improve the accuracy, we set $r_n = I_n$, which means each factor matrix U_n is square and $\underline{g}$ has the same size of $\underline{X_s}$ (as the bottom right part of Fig. 2). In addition, to guarantee each frequency positive, we add a group of constraints as

$$\underline{g} \geq 0, \quad U_i \geq 0 \ (i = 1, \cdots, n)$$

Tucker decomposition can be viewed as a high dimensional Principle Component Analysis (PCA). Factor matrices are the principle components. Similarly, the actual population tensor can be also decomposed as

$$\underline{X} = \underline{G} \times_1 V_1 \times_2 \cdots \times_n V_n$$

$\underline{G}$ is in the same size with $\underline{g}$. We use U_n as the approximation of V_n and the above formula is converted into

$$\underline{X} = \underline{G} \times_1 U_1 \times_2 \cdots \times_n U_n \tag{2}$$

Since U_n are already known, the problem is to estimate the core tensor $\underline{G}$, under certain constraints (as the top right part of Fig. 2).

The second type of data source is marginal/partial joint distributions from actual population. Such marginal and partial joint distributions are obtained by summing up the original tensor in other unrelated dimensions. As the top left sub-figure in Fig. 2 shows, there are two constraints: one marginal distribution $\underline{X}^{(3)}$ and one partial joint distribution $\underline{X}^{(12)}$. $\underline{X}^{(3)}$ is achieved by aggregating the full $3\times4\times5$ tensor in dimension 1 and 2. $\underline{X}^{(12)}$ is achieved only from the aggregation in dimension 3. The marginal and partial joint constraints can be written as

$$\begin{aligned} \underline{X}^{(3)} &= \underline{X} \times_2 e_2 \times_1 e_1 = e_1 \cdot (e_2 \cdot \underline{X}_{(2)})_{(1)} \\ \underline{X}^{(12)} &= \underline{X} \times_3 e_3 = e_3 \cdot \underline{X}_{(3)} \end{aligned}$$

where $e_n = (1, \cdots, 1) \in \mathbb{N}^{1 \times I_n}$ is an I_n-dimensional vector of ones. Substitute Eq. (2) into the above constraints (here $n = 3$), we have

$$\begin{aligned} \underline{X}^{(3)} &= (\underline{G} \times_1 U_1 \times_2 U_2 \times_3 U_3) \times_2 e_2 \times_1 e_1 \\ \underline{X}^{(12)} &= (\underline{G} \times_1 U_1 \times_2 U_2 \times_3 U_3) \times_3 e_3 \end{aligned}$$

This can be re-written as

$$\begin{aligned} \underline{X}^{(3)} &= \underline{G} \times_1 (e_1 \cdot U_1) \times_2 (e_2 \cdot U_2) \times_3 U_3 \\ \underline{X}^{(12)} &= \underline{G} \times_1 U_1 \times_2 U_2 \times_3 (e_3 \cdot U_3) \end{aligned}$$

Therefore

$$\begin{aligned} \underline{G} \times_1 (e_1 \cdot U_1) \times_2 (e_2 \cdot U_2) &= \underline{X}^{(3)} \times_3 U_3^{\dagger} \\ \underline{G} \times_3 (e_3 \cdot U_3) &= \underline{X}^{(12)} \times_1 U_1^{\dagger} \times_2 U_2^{\dagger} \end{aligned} \tag{3}$$

where $U_n^{\dagger}$ is Moore-Penrose (MP) pseudo inverse of U_n. Note that in Eq. (3), $(e_n \cdot U_n)$ is a vector for $n = 1, 2, 3$, and our task is to determine $\underline{G} \geq 0$ with all the other items known. In this 3-dimensional example, it means we need to solve an under-determined system with $3 \times 4 \times 5 = 60$ variables and $5+3\times4 = 17$ equations. Many optimization algorithms can be exploited to complete such task. Here we use Gradient Decent to do this. Define the total error as

$$\begin{aligned} J = \sum \Big[\underline{G} \times_1 (e_1 \cdot U_1) \times_2 (e_2 \cdot U_2) - \underline{X}^{(3)} \times_3 U_3^{\dagger}\Big]^2 \\ + \Big[\underline{G} \times_3 (e_3 \cdot U_3) - \underline{X}^{(12)} \times_1 U_1^{\dagger} \times_2 U_2^{\dagger}\Big]^2 \end{aligned} \tag{4}$$

On the right of equation sign, the two items inside square brackets are 5×1 matrix and 3×4 matrix, same sizes as the constraints $\underline{X}^{(3)}$ and $\underline{X}^{(12)}$. The summation and square signs mean each element of the matrix is squared and then summed together. Thus Eq. (4) sums $5 \times 1 + 3 \times 4 = 17$ quadratic items to compute the total error. Using the sample decomposition kernel $\underline{g}$ as an initial solution, Gradient Decent algorithm will minimize the total error and finally achieve an estimation of $\underline{G}$. Substitute the estimation into Eq. (2), we will get the population distribution.

In a general case, there may be more than two marginal and partial joint constraints, and Eq. (4) will have more quadratic items as well. In addition, we may set $r_n < I_n$ so that factor matrix U_n is no longer square. This means we use principal components with lower ranks to approximate the original tensor. General pseudo code is provided in Table 1. The tensor decomposition method is also applicable for other social organizations' synthesis. If the disaggregate sample of such type of organizations is not available, the initial solution can be set arbitrarily or from other sample-free methods.

B. Population Assignment

After basic population and social organization synthesized, the second step is to assign each individual to specific organizations. The objective is to establish the assignment probability to each organization type. To avoid unreasonable assignment (means the assignment probability equals to zero), we need to establish the mapping relations between social organizations and individuals. Two sources of information can be exploited. One is called data based assumption. If the

TABLE I
INDIVIDUAL AND HOUSEHOLD ATTRIBUTES

Attr.	Values	Num. of Values
Individual Attributes		
Gender	male, female	2
Residential Province	Beijing, Tianjin, ...	31
Residence Type	city, town, rural	3
HH Type	family, collective HH	2
Age Interval	0-5, ..., 96-100, ≥100	21
Household Attributes		
HH Type	family, collective HH	2
Residential Province	Beijing, Tianjin, ...	31
Residence Type	city, town, rural	3
Mem. Num.	1, ..., 9, ≥10	10
Elder Num.	0, 1, 2, ≥3	4
Enterprise Attributes		
Res. Prov.	Beijing, Tianjin, ...	31
Enter. Type	Corp., Indust. Unit, None	3
Enter. Scale	≤7, 8-19, ..., ≥10000	10

Algorithm 1 TensorDecomPopSyn ($\underline{X_s}$, $\underline{X}^{con}$)

```
Input:
    $\underline{X_s}$, $I_1 \times I_2 \times \cdots \times I_n$ sample tensor;
    $\underline{X}^{con}$, lower dimensional constraint tensors, represented as
    $\underline{X}^{(l_1,l_2\cdots,l_c)}$, $\{l_1, l_2, \cdots, l_c\} \subset \{I_1, I_2, ?, I_n\}$;
Output:
    Population Tensor.
$[g, U_1, \cdots, U_n] \leftarrow Tucker\_Decom(\underline{X_s}, I_1, I_2, ?, I_n)$;
$CoreCons \leftarrow \{\}$;
for each constraint $\underline{X}^{(l_1,l_2\cdots,l_c)}$ do
    $con(l_1, l_2, \cdots, l_c) \leftarrow \underline{X}^{(1,2,\cdots,m)} \times_1 U_1^{\dagger} \times_2 \cdots \times_m U_m^{\dagger}$; /*
    $U_m^{\dagger}$ is Moore-Penrose pseudo inverse of $U_m$ */
    $CoreCons \leftarrow CoreCons \cup con(l_1, l_2, \cdots, l_c)$;
end for
$\underline{G} \leftarrow GradDecent(\underline{g}, U_1, \cdots, U_n, CoreCons, error)$;
return $\underline{G} \times_1 U_1 \times_2 \cdots \times_n U_n$.
```

input data from different levels have overlapped attributes (such as the locations in both individual and enterprise levels, as discussed later in this paper), or some member features (such as members' age structure from the organization survey), we can establish some mapping relations based on reasonable assumptions. For instance, if individual and enterprise levels both contain locations, we can safely assume that in a coarse-grained level, employed person works in an enterprise that has the same location as his residence. The overlapped attributes can lead to a conditional assignment probability as

$$\begin{aligned}
&P(\underline{X}, \underline{Y} \mid \underline{X}) \\
&= P(\underline{Y} \mid \underline{X}) \\
&= P(y_1, \cdots, y_u, y_{u+1}, \cdots, y_h \mid x_1, \cdots, x_u, x_{u+1}, \cdots, x_n) \\
&= P(y_1, \cdots, y_u, y_{u+1}, \cdots, y_h \mid x_1, \cdots, x_u) \\
&= P(y_1, \cdots, y_u, y_{u+1}, \cdots, y_h \mid y_1, \cdots, y_u) \\
&= P(y_{u+1}, \cdots, y_h \mid y_1, \cdots, y_u) \\
&= \frac{IndNum(y_1, \cdots, y_u, y_{u+1}, \cdots, y_h)}{IndNum(y_1, \cdots, y_u)}
\end{aligned}$$

where

$$\begin{aligned}
\underline{X} &= (x_1, \cdots, x_u, x_{u+1}, \cdots, x_n) \\
\underline{Y} &= (y_1, \cdots, y_u, y_{u+1}, \cdots, y_h)
\end{aligned}$$

are individual and social organization attributes respectively, and $x_i = y_i (i = 1, \cdots, u)$ are the overlapped attributes. $P(\underline{X}, \underline{Y} \mid \underline{X})$ stands for the probability that the individual of type X is assigned to the organization of type Y. It is computed from the individual number required by eligible organizations. Considering the denominator is possibly zero, the assignment probability is

$$\begin{aligned}
&P(\underline{X}, \underline{Y} \mid \underline{X}) \\
&= \begin{cases} \frac{IndNum(y_1,\cdots,y_u,\cdots,y_h)}{IndNum(y_1,\cdots,y_u)} & \text{if } IndNum(y_1, \cdots, y_u) > 0; \\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{5}$$

where

$$\begin{aligned}
&IndNum(y_1, \cdots, y_u) \\
&\quad = \sum_{y_{u+1},\cdots,y_h} IndNum(y_1, \cdots, y_u, y_{u+1}, \cdots, y_h)
\end{aligned}$$

Note that overlapped attributes do not always exist, and the assumptions should be made reasonable. For organization member features such as the age structure mentioned before, the assignment probability can be also computed by Eq. (5). In this case, the numerator is the number of individual X required by organization Y, and the denominator is the total number of individual X required by the whole organizations. In essence, the problem is converted into the overlapped attribute case.

The second source of information for establishing the mapping relations is implicit heuristic rule. This type of heuristics is set according to the rules that are not reflected by the input data. For example, normal non-single family has a couple as its members. This is not mandatory, but we usually give priority to assemble such family. General pseudo code of assignment is provided in Table 2. The loop from line 5 to line 20 sequentially assigns the randomly generated individual to each level organization. The assignment sequence relies on the dependence of different social relationships.

IV. NUMERICAL TEST AND APPLICATION

To validate the proposed method, experiments of Chinese national population synthesis are conducted. We consider two social relationships: household and enterprise. However, the methodology can be easily applied in a more general case. As one of the most populous countries in the world, Chinese population structure is highly complex. Dealing with such a case is very representative in application. Individual and household constraints come from the census results, which directly reflect basic features of target population. Enterprise constraints come from national economic investigation. Experiment results are compared with another two sample-based methods, IPF and CO. This section will briefly

Algorithm 2 PopAssigment (Pop, Orgs, Heurs)

```
Input:
    Pop, basic population;
    Orgs, set of social organizations with $S$ levels, represented as $\{org_1, \cdots, org_S\}$;
    Heurs, heuristics;
Output:
    Population with Social Relationships.
$AssPop \leftarrow \{\}$;
while Pop is not empty do
    $Ind(x_1, \cdots, x_n) \leftarrow GetRandInd(Pop)$;
    $OrgCand \leftarrow \{\}$;
    for each $org_i$ in a given sequence /*The sequence (if has) is determined by the assignment dependencies of different levels*/ do
        for each $Y = (y_1, \cdots, y_h)$ in $org_i$ do
            $IndNum(y_1, \cdots, y_h) \leftarrow ComMemNum(org_i, Ind(x_1, \cdots, x_n), Heurs)$;
            /*Compute required number of Ind by organization $Y$ in level $org_i$*/
            $OrgCand \leftarrow OrgCand \cup \{(y_1, \cdots, y_h), IndNum(y_1, \cdots, y_h)\}$;
        end for
        if OrgCand is not empty then
            $ObjOrg \leftarrow GetRandOrg(OrgCand)$;
            /* Get a random organization according the distribution by normalizing OrgCand*/
            $Ind \leftarrow LinkOrg(Ind, ObjOrg)$;
            /*Assign Ind to ObjOrg */
            $ObjOrg \leftarrow LinkInd(ObjOrg, Ind)$;
            if the member number of ObjOrg reaches its maximum then
                $or_i \leftarrow org_i \backslash ObjOrg$;
            end if
        else
            $Ind \leftarrow LinkOrg(Ind, null)$;
            /* There is no suitable organizations */
        end if
    end for
    $AssPop \leftarrow AssPop \cup Ind$;
    $Pop \leftarrow Pop \backslash Ind$;
end while
return $AssPop$.
```

introduce the data sources, followed by the evaluation results and computational performance.

A. Data Source and Evaluation Criterion

The first sort of input data is statistical marginal/partial joint frequencies from the 5-th national census. As the census is conducted by the government and provides a lot of personal features, these distributions partially indicate the detailed structure of the target population. The marginal frequencies are published in the form of partial cross-classification tables, each of which contains a part of attributes [37]. In the national census, two kinds of questionnaires are used. One is called Short Table which involves several basic characteristics, whereas the other is Long Table which not only has all the content of the short one but also includes additional detailed features like migration pattern, educational level, economic status, marriage and family, etc.. About 9.5% households are surveyed by Long Table, while the rest are recorded by Short Table. Statistical results from the two kinds of tables both include constraints at household and individual levels. Thus Short Table reveals total nationwide population frequencies, while Long Table only covers a small part of the households and population. Table I lists the attributes to study.

TABLE II

MARGINAL DISTRIBUTIONS USED AS INPUT AND EVALUATION CRITERIA

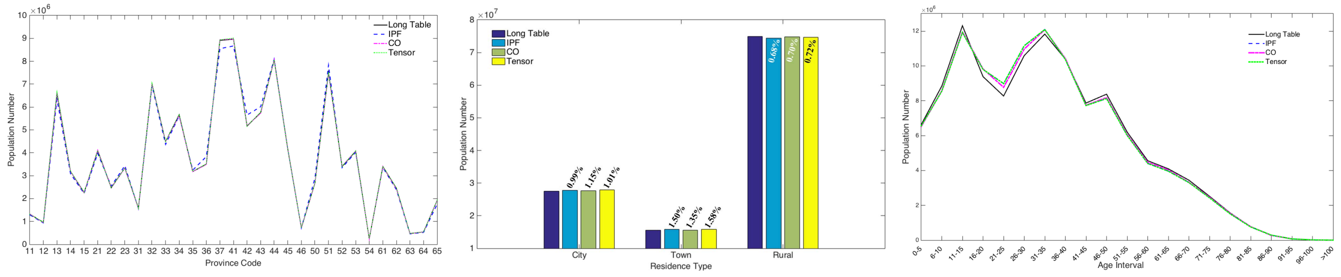
Input Marginal Distributions	
Attr. of Distribution	Type and Level
Gender×Res. Prov.× Res. Type×HH Type	Short Table, Ind.
Gender×Res. Prov.× Res. Type×Age Inter.	Short Table, Ind.
Res. Prov.×Res. Type×HH Type	Short Table, HH.
Res. Prov.×Res. Type×Mem. Num.	Short Table, HH.
Res. Prov.×Res. Type×Elder Num.	Short Table, HH.
Res. Prov.×Enter. Type	Short Table, Enter.
Enter. Type×Enter. Scale	Short Table, Enter.
Criteria for Evaluation	
Attr. of Distribution	Type and Level
Gender×Res. Prov.× Res. Type×HH Type	Long Table, Ind.
Gender×Res. Type×Age Inter.	Long Table, Ind.
Res. Prov.×Res. Type×HH Type	Long Table, HH.
Res. Prov.×Enter. Type	Short Table, Enter.
Enter. Type×Enter. Scale	Short Table, Enter.

The second sort of data is the national economic investigation results, still in the form of cross-classification tables [38]. This investigation reveals the total frequencies of economic entities. Note that the economic investigation does not have the short and long tables as census. Rather, its results are all from the whole target population.

The third sort of input data is the disaggregate population sample, covering both individual and household attributes. The sample comes from the 5-th national census as well, and it is a small part of records from Long Table. The sample consists of $1, 180, 111$ records, each of which gives detailed information of a particular individual (with private information omitted). According to the final census data, our sample accounts for 0.95% of the whole national population.

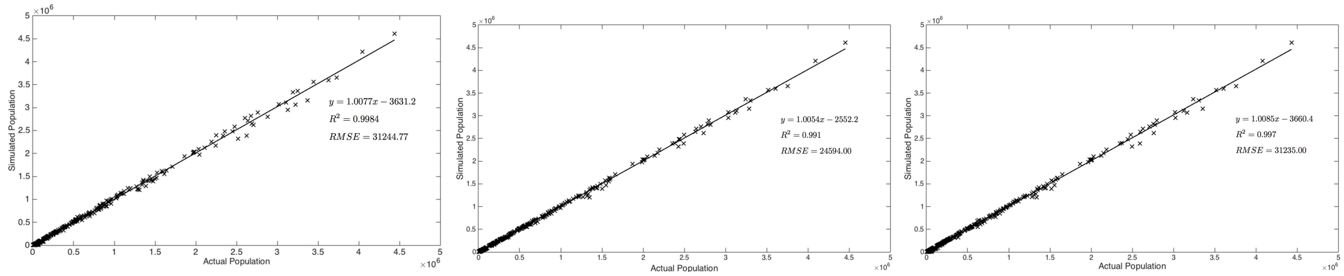
In our experiments, the target population contains $1, 242, 612, 226$ persons, $351, 233, 698$ households and $11, 992, 297$ enterprises. Disaggregate samples and Short Tables are used as inputs while Long Tables are treated as evaluation criteria (Table II). In order to keep our evaluation as objective as possible, a proportion of populations from the total synthetic database are stochastically extracted according to the Long Table scale and compared with the given Long Table results at each level. This is because the Long Table census results come from the actual population directly and are suitable to serve as a benchmark.

In population assignment, two heuristics are set to control the plausible organization's generation. One is to check whether the member number meets the organization's limit after each person assigned. This rule is applicable to both



(a) (b) (c)

Fig. 3. Population frequencies: (a) Residential province (b) Residence type (c) Age interval.



(a) (b) (c)

Fig. 4. Synthetic population partial joint distributions: (a) IPF (b) CO (c) Tensor decomposition.

household and enterprise. Assignment to household also needs to check the elder (age≥65) number. Only adult (aged between 20 and 65) is possibly assigned to enterprise (also controlled by its employee number). When the organization is "full", it means that organization has been completely generated. Another heuristic aims to endow each family with a couple. This is not mandatory but could generate most normal households.

B. Results

The final synthetic population is composed of three databases: individuals, households and enterprises. Some record samples are listed in Table III. As can be seen, there are two males and one female. Their ages are in the 65-69, 60-64 and 35-39 intervals. All the three people belong to the household with ID 320322. The household is recorded in Synthetic Household table (the ID is slightly different with its province code ahead), and it has three members including one elder person. The Enter. ID of the female is -1, which means she is retired and not affiliated with any enterprise. The other two people are employed and affiliated with enterprise 7 and enterprise 29 respectively. The enterprise 7 is recorded in Synthetic Enterprise table.

Fig. 3 shows the aggregated individual frequencies in *Residential Province*, *Residence Type* and *Age Interval*. The average errors are 3.23% (IPF), 0.46% (CO), 0.27% (Tensor) in Residential Province level, respectively, and are 8.28% (IPF), 5.67% (CO), 3.55% (Tensor) in Age Interval. Also, the errors of *Residence Type* are about 1% (marked in the second sub-figure). The three methods can achieve good results for 1-dimensional individual indicators. Fig. 4 presents partial joint distributions of individual numbers. This figure draws the two partial criteria listed in Table II together. So each sub-figure includes 498 data points, concerning "$Gender \times Res.\ Prov. \times Res.\ Type \times HH\ Type$" and "$Gender \times Res.\ Type \times Age\ Inter.$". As can be seen, CO performs a little better than the other two. But IPF and tensor-based methods are also able to reconstruct the target population well.

TABLE III
RECORDS OF SYNTHETIC POPULATION

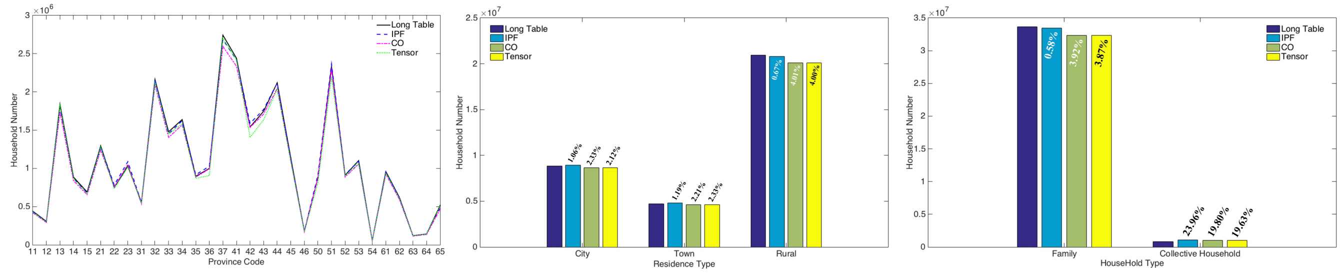
Synthetic Individual

Ind. ID	1100000067	1100000082	1100000099
Gender	Female	Male	Male
Res. Prov.	Beijing	Beijing	Beijing
Res. Type	City	City	City
Age Interval	65-69	60-64	35-39
HH Type	Family	Family	Family
HH ID	320322	320322	320322
Enter. ID	-1	7	29

Synthetic Household & Enterprise

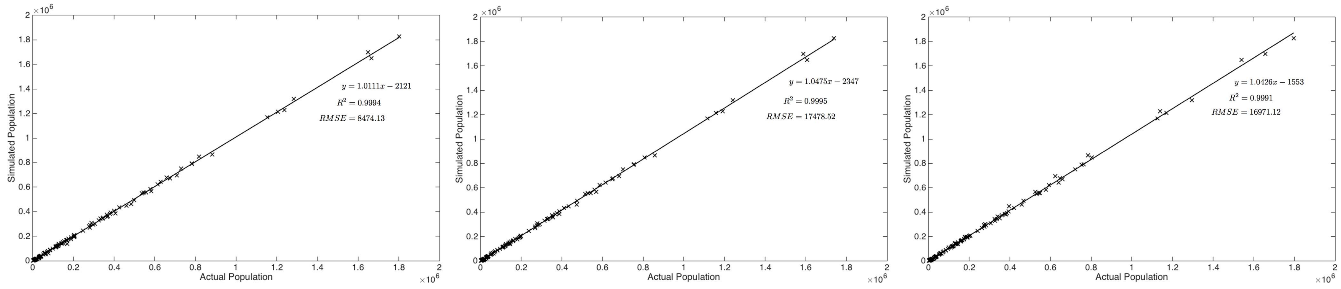
HH ID	1100320322	Enter. ID	1100000007
Res. Prov.	Beijng	Res. Prov.	Beijing
Res. Type	City	Enter. Type	Corporation
HH Type	Family	Enter. Scale	8-19
Member Num.	3	Employee Num.	9
Elder Num.	1		

Similar to individual level, synthetic households are evaluated from marginal and partial joint distributions as well. Fig. 5 gives the marginal results of three metrics. Obviously, the errors from households are larger than individuals. It is mainly caused by the smaller household number compared with individuals. This trend can be clearly seen specifically in the "Collective Household" in the third sub-figure, where the three errors are all around 20%. For the partial joint distributions, Long Table only contains one criterion table in household level, as shown in Table II. Thus there are only 186 data points from $Res.\ Prov. \times Res.\ Type \times HH\ Type$. Fig. 6 indicates that all results are linearly correlated in general, since the goodness of fitting is near 1. However,



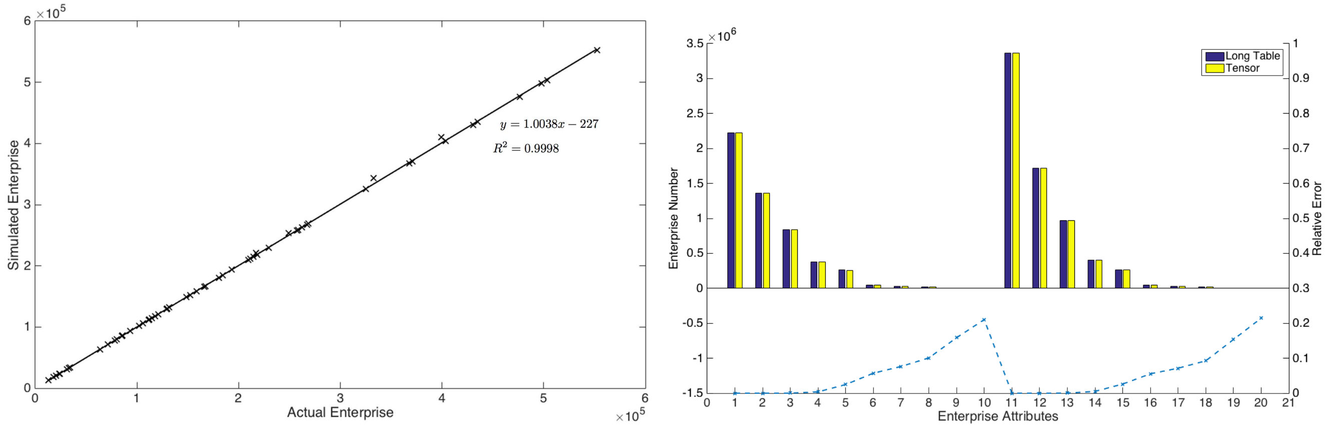
(a) (b) (c)

Fig. 5. Household frequencies: (a) Residential province (b) Residence type (c) Household type.



(a) (b) (c)

Fig. 6. Synthetic household partial joint distributions: (a) IPF (b) CO (c) Tensor decomposition.



(a) (b)

Fig. 7. Synthetic enterprise partial joint distributions: (a) Residential province*enterprise type (b) Enterprise type*enterprise scale.

our tensor-based method gets smaller deviations than the others. This is reflected by the smaller constant of the fitted linear equation, which represents systematic error of the algorithm.

The evaluation of synthetic enterprises uses original short table, due to the unavailability of similar Long Tables. In addition, since our disaggregate sample does not include enterprise information, IPF and CO sample-based algorithms cannot be applied in this level. Fig. 7 gives the results of two partial joint distributions. From the figure, tensor-based algorithm can reconstruct the constraints well, as the systematic deviation is only 227 for the first criterion. In the second sub-figure, error line is associated with the right longitude axis. We can see that the relative error increases when absolute frequency gets smaller. This indicates the algorithm has worse performance on minor groups. However, the largest error still remains about 20%, which can be accepted for low-frequency groups.

TABLE IV
COMPUTATIONAL PERFORMANCE

Method	Dis. Com.	Pop. Assign.	
		Min	Max
IPF	132 sec	38,502 sec	1,675,177 sec
CO	14 day 9 hour 10 min		
Tensor Dec.	5 hour 9 min		

C. Computational Performance

Overall, the computational cost can be calculated in two phases. To accelerate our experiments, synthesis is conducted by province (we will discus the reason in next section). Table IV presents the statistical averaged rum time of Java-based program. The computational environment is composed of two identical machines. Their configuration are Java1.8 SE (x64), Windows 7 (x64), AMD Opteron(tm) Processor 6320

(12 Cores, 2.79 GHz) and 12 GB RAM.. The three methods are compared in distribution computing, where CO requires about two weeks. This is because the fitness improvement is quite slow. In the assignment stage, the minimum time comes from the Tibet province, which has 2.6 million people. The maximum time comes from the Henan province, which has about 91 million people.

V. Conclusion And Discussion

Most current population synthesis methodologies only concentrate on individual and household. This paper addresses a general case with multiple social organizations, and proposes a new tensor decomposition method to solve the problem. To our knowledge, it is the first time that this methodological issue has been addressed. Using Chinese national population data source, the proposed method is tested and compared with other two sample-based ones—IPF and CO. Results indicate that our method can simultaneously deal with multiple social organizations and achieve the same accuracy of IPF and CO.

Generally, when the multi-level assignments involve dependencies, considering each level as an explicit layer (as shown in Fig. 1) is advantageous. For instance, in most municipalities of China, the teenager whose parent (usually the head of his household) has a local registration can enroll the nearest school in priority. Other immigrant children will be randomly assigned after such enrollment. Therefore, we need to investigate his family members before assigning one to school. This can be only achieved after his assignment to household completed. Even though the personal attributes, household attributes and family member list can be stored in one record to merge the household and individual levels, the data structure would be much more complicated especially when the dependencies involve many social organizations. So it is more convenient to treat the multi-level social organizations separately. The sequence of assignments to different levels is also determined by such dependency, as encoded in Alg. 2.

In Alg. 2, the main "while" loop assigns synthetic individual into each type of organizations one by one. The algorithm scans the synthetic populations as well as organizations to dynamically choose eligible candidates. When the scale is large, as the Chinese scenario that involves billions of people, such scanning may take a long time. This is unacceptable in real applications. One solution is to only maintain joint distributions of individuals and organizations, and dynamically draw candidates during iteration. For instance, if the current individual to be assigned is in type X (also can be dynamically drawn from joint distribution) and one eligible enterprise type is Y. The number of X required by Y is

$$IndNum(X) = IndNum(X \mid Y) \cdot EnterNum(Y)$$

$IndNum(X \mid Y)$ means the individual number of X in each enterprise of the type Y. During iteration, organization candidates can be dynamically drawn according to normalization of the above frequencies for each type Y, and complete the assignment. When the assigned organization is "full", it is deleted from memory and saved into database. In the worst case, the algorithm only needs to maintain one organization entity for each type, rather than the whole organization set. This improvement can reduce the memory cost extensively.

Another approach to accelerate the computation of Alg. 2 is parallel computing. Such operation is effective provided that assignments among different individual groups and different levels are independent. As assumed in our case where *Residential Province (RP)* is the conditional attribute between individual and enterprise, the assignment probability is

$$P(y \mid RP = Beij.) \\ = \begin{cases} \frac{IndNum(y,RP=Beij.)}{IndNum(RP=Beij.)} & \text{if } IndNum(RP = Beij.) > 0 \\ 0 & \text{otherwise} \end{cases}$$

That is to say, any person with *RP=Beijing* can only be assigned to an organization with the same *RP* value. Thus we can deploy such assignment in an independent CPU or core without influencing other provinces. Note that this parallel deployment is applicable only when all types of organizations can be partitioned. If there is at least one exception, we will unavoidably face the data synchronization.

References

[1] H. Zhao, S. Tang, and Y. Lv, "Generating artificial popullations for traffic microsimulation," *IEEE Intell. Transp. Syst. Mag.*, vol. 1, no. 3, pp. 22–28, Nov. 2009.

[2] Y. Ou, S. Tang, and F.-Y. Wang, "Computational experiments for studying impacts of land use on traffic systems," in *Proc. IEEE 13th Int. Conf. Intell. Transp. Syst. (ITSC)*, Funchal, Portugal, Sep. 2010, pp. 1813–1818.

[3] P. Ye and D. Wen, "A study of destination selection model based on link flows," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 1, pp. 428–437, Mar. 2013.

[4] P. J. Ye, X. Wang, C. Chen, Y. Lin, and F.-Y. Wang, "Hybrid agent modeling in population simulation: Current approaches and future directions," *J. Artif. Soc. Social Simul.*, vol. 19, no. 1, pp. 1–20, Jan. 2016.

[5] R. M. Pendyala, R. Kitamura, A. Kikuchi, T. Yamamoto, and S. Fujii, "Florida activity mobility simulator: Overview and preliminary validation results," *Transp. Res. Rec.*, vol. 1921, no. 1, pp. 123–130, 2005.

[6] C. R. Bhat, J. Y. Guo, S. Srinivasan, and A. Sivakumar, "Comprehensive econometric microsimulator for daily activity-travel patterns," *Transp. Res. Rec.*, vol. 1894, no. 1, pp. 57–66, 2004.

[7] T. A. Arentze and H. J. P. Timmermans, "A learning-based transportation oriented simulation system," *Transp. Res. B, Methodol.*, vol. 38, no. 7, pp. 613–633, Aug. 2004.

[8] P. Salvini and E. J. Miller, "ILUTE: An operational prototype of a comprehensive microsimulation model of urban systems," *Netw. Spatial Econ.*, vol. 5, no. 2, pp. 217–234, Jun. 2005.

[9] D. R. Pritchard, "Synthesizing agents and relationships for land use/transportation modelling," M.S. thesis, Dept. Civil Eng., Univ. Toronto, Toronto, ON, Canada, 2008.

[10] M. J. Roorda, E. J. Miller, and K. M. N. Habib, "Validation of TASHA: A 24-h activity scheduling microsimulation model," *Transp. Res. A, Policy Pract.*, vol. 42, no. 2, pp. 360–375, Feb. 2008.

[11] L. Smith, R. Beckman, and K. Baggerly, "TRANSIMS: Transportation analysis and simulation system," Los Alamos Nat. Lab., Los Alamos, NM, USA, Tech. Rep. LA-UR-95-1641, 1995.

[12] R. J. Beckman, K. A. Baggerly, and M. D. McKay, "Creating synthetic baseline populations," *Transp. Res. A, Policy Pract.*, vol. 30, no. 6, pp. 415–429, Nov. 1996.

[13] W. D. Wheaton *et al.*, "Synthesized population databases: A US geospatial database for agent-based models," RTI Int., Research Triangle Park, NC, USA, Tech. Rep. MR-0010-0905, 2009.

[14] L. Sun and A. Erath, "A Bayesian network approach for population synthesis," *Transp. Res. C, Emerg. Technol.*, vol. 61, pp. 49–62, Dec. 2015.

[15] I. Saadi, A. Mustafa, J. Teller, B. Farooq, and M. Cools, "Hidden Markov model-based population synthesis," *Transp. Res. B, Methodol.*, vol. 90, pp. 1–21, Aug. 2016.

[16] D. Casati, K. Müller, P. J. Fourie, A. Erath, and K. W. Axhausen, "Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking," *Transp. Res. Rec.*, vol. 2493, no. 1, pp. 107–116, Jan. 2015.

[17] J. E. Abraham, K. J. Stefan, and J. D. Hunt, "Population synthesis using combinatorial optimization at multiple levels," in *Proc. Transp. Res. Board 91st Annu. Meeting*, Washington DC, USA, Jan. 2012.

[18] M. Lenormand and G. Deffuant, "Generating a synthetic population of individuals in Households: Sample-free vs sample-based methods," *J. Artif. Soc. Social Simul.*, vol. 16, no. 4, pp. 1–12, Oct. 2013.

[19] R. Lovelace and M. Dumont, *Spatial Microsimulation With R*. Boca Raton, FL, USA: CRC Press, 2016.

[20] P. Ye, X. Hu, Y. Yuan, and F.-Y. Wang, "Population synthesis based on joint distribution inference without disaggregate samples," *J. Artif. Soc. Social Simul.*, vol. 20, no. 4, p. 16, Jan. 2017.

[21] F. Gargiulo, S. Ternes, S. Huet, and G. Deffuant, "An iterative approach for generating statistically realistic populations of households," *PLoS ONE*, vol. 5, no. 1, Jan. 2010, Art. no. e8828.

[22] J. Barthelemy and P. L. Toint, "Synthetic population generation without a sample," *Transp. Sci.*, vol. 47, no. 2, pp. 131–294, May 2013.

[23] S. Huet, M. Lenormand, G. Deffuant and F. Gargiulo, "Parameterisation of Individual Working Dynamics," in *Empirical Agent-Based Modelling—Challenges and Solutions: The Characterisation and Parameterisation of Empirical Agent-Based Models*, vol. 1, A. Smajgl and O. Barreteau, Eds. New York, NY, USA: Springer, 2014, pp. 133–169.

[24] L. Ma and S. Srinivasan, "Synthetic population generation with multilevel controls: A fitness-based synthesis approach and validations," *Comput. Aided Civil Infrastruct. Eng.*, vol. 30, no. 2, pp. 135–150, Feb. 2015.

[25] N. Huynh, J. Barthelemy, and P. Perez, "A heuristic combinatorial optimization approach to synthesizing a population for agent based modeling purposes," *J. Artif. Soc. Social Simul.*, vol. 19, no. 4, p. 11, 2016.

[26] P. Anderson and B. Farooq, "A generalized partite-graph method for transportation data association," *Transp. Res. C, Emerg. Technol.*, vol. 76, pp. 150–169, Mar. 2017.

[27] T. Melhuish, M. Blake, and S. Day, "An evaluation of synthetic household populations for census collection districts created using optimization techniques," *Australas. J. Regional Stud.*, vol. 8, no. 3, p. 369, 2002.

[28] D. Ballas, G. Clarke, D. Dorling, H. Eyre, B. Thomas, and D. Rossiter, "SimBritain: A spatial microsimulation approach to population dynamics," *Population, Space Place*, vol. 11, no. 1, pp. 13–34, Jan. /Feb. 2005.

[29] J. Y. Guo and C. R. Bhat, "Population synthesis for microsimulating travel behavior," *Transp. Res. Rec.*, vol. 2014, no. 1, pp. 92–101, 2007.

[30] D. R. Pritchard and E. J. Miller, "Advances in population synthesis: Fitting many attributes per agent and fitting to household and person margins simultaneously," *Transportation*, vol. 39, no. 3, pp. 685–704, May 2012.

[31] J. Auld and A. Mohammadian, "Efficient methodology for generating synthetic populations with multiple control levels," *Transp. Res. Rec.*, vol. 2175, no. 1, pp. 138–147, 2010.

[32] X. Ye, K. Konduri, R. M. Pendyala, B. Sana, and P. Waddell, "A methodology to match distributions of both household and person attributes in the generation of synthetic populations," in *Proc. 88th Transp. Res. Board Annu. Meeting Transp. Res. Board*, Jan. 2009, pp. 1–24.

[33] L. Sun, A. Erath, and M. Cai, "A hierarchical mixture modeling framework for population synthesis," *Transp. Res. B, Methodol.*, vol. 114, pp. 199–212, Aug. 2018.

[34] W. E. Deming and F. F. Stephan, "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known," *Ann. Math. Statist.*, vol. 11, no. 4, pp. 427–444, Dec. 1940.

[35] P. Williamson, M. Birkin, and P. H. Rees, "The estimation of population microdata by using data from small area statistics and samples of anonymised records," *Environ. Planning*, vol. 30, no. 5, pp. 785–816, May 1998.

[36] S. Rabanser, O. Shchur, and S. Gunnemann, "Introduction to tensor decompositions and their applications in machine learning," 2017, *arXiv:1711.10781*. [Online]. Available: https://arxiv.org/pdf/1711.10781.pdf

[37] National Bureau of Statistics of the People's Republic of China. *The 5-th National Census Data*. Accessed: Jan. 15, 2019. [Online]. Available: http://www.stats.gov.cn/tjsj/pcsj/rkpc/5rp/index.htm

[38] National Bureau of Statistics of the People's Republic of China. *The 1st Economic Investigation Data*. Accessed: Jan. 15, 2019. [Online]. Available: http://www.stats.gov.cn/tjsj/pcsj/jjpc/1jp/indexch.htm

Peijun Ye received the Ph.D. degree from the University of Chinese Academy of Sciences in 2013. He was a Visiting Scholar with the Department of Cognitive Science, University of California at San Diego, from 2017 to 2018. He is currently an Assistant Researcher with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. His research interests mainly focus on multi-agent systems, social simulation, artificial intelligence, and intelligent transportation systems. He is an Invited Reviewer of the *Journal of Artificial Societies and Social Simulation, Knowledge and Information Systems*, and the IEEE International Conference on Intelligent Transportation Systems (2011–2019). He serves as an Associate Editor for the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, the IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, and the Secretary for the Social and Economic Computing ACM Chapter.

Fenghua Zhu received the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2008. He is currently an Associate Professor with the State Key Laboratory for Management and Control of Complex Systems, Chinese Academy of Sciences, China. His research interests are artificial transportation systems and parallel transportation management systems.

Samer Sabri received the B.S./B.A. degree in computer science and ethics, politics, and economics from Yale University in 2013. He is currently pursuing the M.Sc. degree in computer science with the University of California at San Diego (UCSD). From 2014 to 2016, he was a Business Analyst at McKinsey & Company, a global management consulting firm, where he focused on the technology and healthcare sectors. His research interests include computational psychiatry and psychosis, Bayesian models of cognitive processing and agent-based models in sociology and economics, and the ethics of machine learning and AI.

Fei-Yue Wang received the Ph.D. degree in computer and systems engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990. He joined the University of Arizona, Tucson, AZ, USA, in 1990, and became a Professor and the Director of the Robotics and Automation Lab and the Program in Advanced Research for Complex Systems. In 1999, he founded the Intelligent Control and Systems Engineering Center, Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, under the support of the Outstanding Overseas Chinese Talents Program from the State Planning Council and the 100 Talent Program from CAS. In 2002, he was appointed as the Director of the Key Laboratory of Complex Systems and Intelligence Science, CAS. In 2011, he became the State Specially Appointed Expert and the Director of the State Key Laboratory of Management and Control for Complex Systems. His current research interests include methods and applications for parallel systems, social computing, and knowledge automation.