Connected Population Synthesis for Urban Simulation

Danqing Zhang\textsuperscript{a}, Junyu Cao\textsuperscript{b}, Sid Feygin\textsuperscript{a}, Dounan Tang\textsuperscript{c}, Alexei Pozdnoukhov\textsuperscript{a,c}

\textsuperscript{a}Civil Systems Engineering, UC Berkeley
\textsuperscript{b}Industrial Engineering Operations Research, UC Berkeley
\textsuperscript{c}Institute of Transportation Studies, UC Berkeley

Abstract

Urban modeling, including agent-based modeling of the coupled transportation and land use evolution, requires detailed information on each of the agents that represent the population in the region of a study. Traditional ways of obtaining this information include household surveys based on individual travel diaries. The surveys data provide a rich set of features, but they are limited in sampling size, geographical scope and frequency of updates. Moreover, they lack detail on inter-personal connections that give rise to social influences driving choice processes at a range of time scales. While manual surveying techniques are limited in their ability to collect social network data at scale, digital records of inter-personal communications provide an abundance of social networking information. This work proposes an algorithmic procedure that makes use of both traditional survey data as well as digital records of networking and human behaviours in generating connected synthetic populations for urban simulation.

Keywords: population synthesis, cellular data, Bayesian networks, structural learning, mixed integer programming, exponential random graph model, agent-based modeling, urban simulation

Email addresses: danqing0703@berkeley.edu (Danqing Zhang), jycao@berkeley.edu (Junyu Cao), sid.feygin@berkeley.edu (Sid Feygin), dounan.tang@berkeley.edu (Dounan Tang), alexeip@berkeley.edu (Alexei Pozdnoukhov)
1. Introduction

Understanding how urban systems operate and evolve through modeling individuals’ urban daily activities has been a major focus of transportation planners, urban planners and geographers. To address operational needs in planning and policy decisions, reliable agent-based land use and urban transportation micro-simulation frameworks such as TRANUS [1], UrbanSIM [52], ILUTE [40], MATSim [3], MEPLAN [10] are becoming popular. The models implemented via micro-simulations require detailed information on each of the agents that represent the population in the region of study. Traditional ways of obtaining this information include using community survey data, or travel surveys based on individual or households travel diaries. These datasets provide a rich set of features, but are limited in sampling size, geographical scope and frequency of updates. The reasons for the limited availability of such detailed disaggregated data to researchers range from the lack of technical means or resources for surveying to personal information protection requirements and related data security and privacy concerns.

Inter-personal communication networks are not captured by the traditional data collection methods. It limits the development of the next generation of models that would appropriately integrate social effects with spatial-temporal information so as to better capture urban systems dynamics. To enable agent-based simulation in the presence of social influence effects, a synthetic connected population should be generated as input. However, current state-of-art population synthesis models fail to generate social networks information since household surveys do not capture this aspect. Alternatively, the prevalence of mobile phones provides a new data source for simulation. Pervasive sensing by telecom companies and location-based service providers generates large scale geolocated communication datasets in which timestamped locations of users are recorded whenever calls are placed or messages are sent. An example of such data is network carrier mobile phone usage logs, such as Call Detail Record (CDR) data. While manual surveying techniques are limited in their ability to collect social network data at scale, digital records of CDR provide an abundance information on spatial patterns of social networks [53].

In this paper, we propose an integrated connected population synthesis methodology that incorporates survey-based household level data with digital records of social networking. It brings together several data components, reproducing the marginal and joint distributions of individual and household level socio-economic characteristics, a geographical pattern of the observed
community structure, and the statistics of the observed social networks.

The proposed population synthesis methodology includes the following steps, presented graphically in Figure 1. First, the household data are used to reproduce the socio-economic characteristics within a generated synthetic population. Next, the structure of the social network in the region is inferred from available network data, and applied to connect the members of the generated households into a synthetic social network that follows the key structural properties of the observed one. The proposed methodology acknowledges the limitations of the data availability, as household and social network data are typically available from two separate sources with no implicit way to identify individuals present in both. Therefore, the sequence of methods that we introduce is aimed to model and reproduce key statistical characteristics of the connected population. The methods involved at the population modeling and synthesis are:

- **Step 1. Bayesian Networks**: Composition and socio-economic characteristics of the synthetic households are generated based on Bayesian network parameters estimated from a typical household survey data.
(such as the American Community Survey);

- **Step 2. Community allocation:** Integer programming problem of community assignment is solved with the Lagrangian Relaxation Method in order to enrich the simulated population with community membership;

- **Step 3. ERGM learning and simulation:** parameters of an Exponential Random Graph Model are calibrated on the available social network data and applied to simulate social connections between the members of the synthetic population.

The rest of the paper is organized as follows. Section 2 presents the methods involved at each step, placing them into the context of the state-of-the-art and discussing the advantages and limitations of both. The following Sections 3 and 4 illustrate the application of the methods on real data, simulating a connected population in the San Francisco Bay Area in California, US. Finally, Section 5 concludes the paper with a discussion of the achieved results and outlines the directions for future work.

2. Methods

2.1. Step 1: Bayesian Networks

Based on the vast amount of literature on population synthesis, we choose to follow the Bayesian networks approach similar to [51]. The Bayesian network approach is a generic formalism aimed at modeling the joint distribution of \( X \) from data. It consists of the two main steps: (1) structure learning to define the Bayesian network structure \( \mathcal{G} \) that describes the conditional independence of the random variables, and (2) parameter learning to learn a conditional distribution of random variables given this fixed directed acyclic graph (DAG) structure \( \mathcal{G} \).

In the context of population synthesis, the objectives of this method is to infer the multivariate probability distribution \( P(X) \) of socio-economic parameters of households based on observed data. Socio-economic parameters \( X \in \mathbb{R}^d \) and \( x_1, x_2, ..., x_d \) are typically composed of \( d \) discrete random variables representing the available information on both households and individuals, collected via surveying, and regarded as the complete set of observations \( D = \{(x_{1t}^t, x_{2t}^t, ..., x_{dt}^t), t \in [1,n]\} \).
Traditional population synthesis methods encompass two main directions: (1) Iterative Proportional Fitting (IPF) related models that focus on fitting a contingency table constructed from the micro samples to meet marginal distribution constraints from aggregated census data, along with the extensions of IPF models that aim to meet marginal distribution constraints of both individuals and households characteristics \[56, 6, 58\]; and, (2) statistical models that use MCMC (Monte Carlo Markov Chain) to sample a vector of socio-economic characteristics of each individual sequentially so that it captures the observed conditional relationships between the variables \[14\].

As a method of replicating existing sample data, IPF-related methods are sensitive to data quality and sample size. MCMC-based approach, however, also suffers from the drawback that it is hard to obtain the specified conditional distributions especially when one deals with many variables of interest. As pointed out by \[51\], Bayesian network approach is powerful in characterizing the underlying joint distribution, outperforming IPF and MCMC, and avoiding over-fitting the data at the same time.

2.1.1. Structural learning and parameter estimation

The Bayesian network structure learning algorithms can be grouped in two categories: (1) constrained-based algorithms where dependencies are set using domain knowledge and the resulting models are interpreted as causal models; (2) score-based models algorithms where we select the graph structure \(G\) that results in the highest score following some accepted scoring criterion. In the present work, we use the algorithms from the second category.

For each graph structure \(G\), we factorize the joint distribution \(P(X)\) as the product of conditional distributions \(P(X_i|X_{\pi_i})\) where \(X_{\pi_i}\) represents the parent nodes of \(X_i\) given graph \(G\).

\[ P(X; G) = \prod_{i=1}^{d} P(X_i|X_{\pi_i}) \]  

Then we can rewrite the log likelihood of the data given model parameters \(\theta\) and the graph structure \(G\) in question as:

\[ l(D; \theta, G) = \sum_{t=1}^{n} \log(P(x_1^t, ..., x_d^t; \theta, G)) \]

\[ = \sum_{t=1}^{n} \sum_{i=1}^{d} \log(P(x_i^t|x_{\pi_i}^t; \theta)), \]
where $d = \dim(\mathcal{G})$ is the number of parameters (a table of conditional probabilities in case of the discrete variables), and $n$ is the number of observations in $D$. Based on the definition of log likelihood, one can get the probability parameters $\hat{\theta}$ based on maximum likelihood estimation. One can then consider the maximum value of log likelihood for the graph structure $\mathcal{G}$, given estimated parameters $\hat{\theta}$ as $l(D; \hat{\theta}, \mathcal{G})$. Each candidate Bayesian network can be assigned a score, and a final model can be selected based on the graph structure that results in the highest score. There are different network scores that can be used, among which the (Bayesian information criterion) and AIC (Akaike information criterion)

$$BIC(\mathcal{G}) = l(D; \hat{\theta}, \mathcal{G}) - \frac{d}{2} \log(n),$$

$$AIC(\mathcal{G}) = l(D; \hat{\theta}, \mathcal{G}) - d,$$

are the most common choices. It is impossible to search over all possible graph structures $\mathcal{G}$ since the number of candidate graph structures increase super-exponentially with the number of variables [39]. The admissible set of structures can be defined using domain knowledge, and the selection process then optimized with some heuristic algorithms such as hill-climbing or tabu search. In this work, we followed [51] to use tabu search algorithm and AIC score for Bayesian networks structural learning. We further refer the reader to [22], [23], and [7] for the detailed survey on Bayesian networks.

A specific Bayesian network model is estimated for every type of a household, following a descriptive classification accepted in a region of study. Household tables are produced with an IPF method to match the marginals of household types within the total population. Population synthesis includes sampling $N$ entries from the $P(X)$, specific to the household composition dependencies captured by the structure of the graph $G$ for the given type of the household. It results in a complete set of $N$ individuals within $H$ households with detailed socio-economic characteristics. Section 4.1 below illustrates the use of the method in the context of typically available survey data.

2.2. Step 2: Community Assignment

As justified in the introduction, community membership is one of the key features that we would like to reproduce in the connected population synthesis.
We consider community to be defined as a group of individuals that possess stronger ties within the group as compared to the connections emanating to the outside of the group. Community formation its structure is strongly related to its function. In relation to identifying the function, it is however not trivial distinguishing social influence from homophily \cite{32,9,45}, which is defined as the tendency for individuals with similar characteristics and behaviors to form clusters in social networks. Spatial aspects of community formation is also a long standing problem. Urban sociology studies have identified communities based on neighborhoods and school districts, as well as forming as a result of social activities that bring various groups of residents together \cite{20}.

Therefore, we acknowledge both spatial proximity and homophily (i.e. a tendency to bond with peers similar in terms of the socio-economic parameters) as two main properties that needs to be included in the community assignment step of population synthesis. However our precise goal is not to model the causal reason of communities formation in a given area, but to provide a method that transfers the observed structure onto a synthetic population. The step we propose is a prescriptive assignment of community membership, that is then taken into account at the actual social connections synthesis step using the model described in Section 2.3.

To assign individual to communities, we formulate the community assignment as an integer programming problem. Without loss of generality, we consider spatial locations to be known at the resolution of a spatial zone according to the system defined by a population census. A random location within a zone can be assigned to each household, or a location following the exact address can be specified if available from a real estate or a cadaster dataset.

\subsection{Notation}

Assume a synthetic population of the total of $N$ people to be grouped into a total of $K$ communities. Suppose $F \in R^{N \times N}$ is the feature distance matrix between individuals. It can be defined as a distance based on the vector of socio-economic characteristics to capture homophily relationships, as well as extended (or replaced) with a similarity measure available from the data at hand and the goals of the analysis specific to the function carried by the modeled community membership. An appropriate model for $F$ is therefore application specific and should be governed by the type of the simulation the synthetic population is intended to serve. We provide an example in the
experimental section below.

We further introduce $c_j$ and $Z_{ij}$ as decision variables. For convenience of the assignment formulation, we define $c_j = 1$ if the $j^{th}$ person is the social center of his community. This representation does not carry a functional meaning but defines a convenient way to enumerate the groups. Similarly, $Z_{ij}$ is an assignment matrix, with $Z_{ij} = 1$ if the $i^{th}$ person belongs to the community centered around $c_j$.

Table 1: Notation

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{ij}$</td>
<td>the feature distance between individual $i$ and $j$</td>
</tr>
<tr>
<td>$Z_{ij}$</td>
<td>individual $i$ belongs to the community whose center is $j$</td>
</tr>
<tr>
<td>$m$</td>
<td>the smallest size of the community</td>
</tr>
<tr>
<td>$M$</td>
<td>the largest size of the community</td>
</tr>
<tr>
<td>$c_j$</td>
<td>if $j^{th}$ individual is the center of the community</td>
</tr>
</tbody>
</table>

2.2.2. Community Assignment Method

In the presented approach, we formulate an optimization problem that represents the following process of community assignment. Assume we first select $K$ individuals serving as centroids of the communities, and assign other individuals to these centroids. Assuming that members of the same community tend to have similar features defined and available in $F$, we relax this assumption to the one that people tend to have similar features as their community centroids, justifying the re-assignment step that maximizes the objective function. An individual can belong to different communities, but he/she should be a part of at least one community. Finally, the size of each community is bounded by the lower and upper bound detected from data or specified otherwise. This set of assumptions corresponds to the below optimization problem for the community assignment step (refer to Table 1).
for notation).

\[
\begin{align*}
\text{minimize} & \quad \sum_{i}^{N} \sum_{j}^{N} F_{ij} \ast Z_{ij} \\
\text{subject to} & \quad \sum_{j=1}^{N} Z_{ij} = 1, \forall i; \\
& \quad \sum_{i=1}^{N} Z_{ij} \leq M \ast c_{j}, \forall j; \\
& \quad \sum_{i=1}^{N} Z_{ij} \geq m \ast c_{j}, \forall j; \\
& \quad c_{j} = 1 \text{ or } 0, \forall j; \\
& \quad Z_{ij} = 1 \text{ or } 0, \forall i, j.
\end{align*}
\]

To efficiently solve the assignment problem we relax the condition \(\sum_{j=1}^{N} Z_{ij} = 1\), using the Lagrangian Relaxation Lower Bound (LLBP) method that results in an formulation defined in Equation (5):

\[
\begin{align*}
\text{minimize} & \quad \sum_{i}^{N} \sum_{j}^{N} (F_{ij} + \lambda_{i}) \ast Z_{ij} - \sum_{i=1}^{N} \lambda_{i} \\
\text{subject to} & \quad \sum_{i=1}^{N} Z_{ij} \leq M \ast c_{j}, \forall j \\
& \quad \sum_{i=1}^{N} Z_{ij} \geq m \ast c_{j}, \forall j; \\
& \quad c_{j} = 1 \text{ or } 0, \forall j; \\
& \quad Z_{ij} = 1 \text{ or } 0, \forall i, j.
\end{align*}
\]
This problem is separable, so we write the sub-problem as $j$-problem:

$$\text{minimize}_{c,Z} \sum_{i}^{N} (F_{ij} + \lambda_i) * Z_{ij} - \lambda_i$$

subject to

$$\sum_{i=1}^{N} Z_{ij} \leq M * c_j, \forall j$$
$$\sum_{i=1}^{N} Z_{ij} \geq m * c_j, \forall j$$
$$c_j = 1 \text{ or } 0, \forall j$$
$$Z_{ij} = 1 \text{ or } 0, \forall i, j$$

This problem can be solved numerically using Algorithm 1. If $c_j = 1$, one should pick at least $m$ and at most $M Z_{ij}$ to be 1, that is, if $c_j$ is a community center, then at least $m$ people and at most $M$ people should belong to this community. Then $F_{ij} + \lambda_i$ are computed for each $i$ and ranked. Let $N_c$ denote the number of negative coefficients. If $N_c \leq m$, then we sum the first $m$ smallest coefficients. If $m < N_c \leq M$, then we sum the first $N_c$ smallest coefficients. If $N_c > M$, then we sum the first $M$ coefficients. Let $S$ denote this sum. Since we want to minimize the objective function, set $c_j = 1$ if $S < 0$ and set $c_j = 0$ otherwise. For any $\lambda$, what we calculate for Lagrangian relaxation is a lower bound, which is denoted by LB. To update $\lambda$, we also need to compute the upper bound, which is denoted by UB. Let $L = \{i| \sum_{j} Z_{ij} = 0\}$. Randomly pick $\lceil |L|/M \rceil$ points as centers $C$. Then we set each point $i \in L$ to a center $j \in C$ until each community has $m$ people according to the feature distance order. Then one continues to set the remaining points $i \in L$ to center $j \in C$ until the number reaches $M$. When all members are allocated, one finds a feasible solution and thus an upper bound for this problem. Lagrange multipliers $\lambda_i$ are then updated as follows:

$$\lambda_{i}^{t+1} = \lambda_{i}^{t} + \Delta^{t}(\sum_{j=1}^{N} Z_{ij} - 1)$$

with the step size $\Delta^{t} = \frac{\alpha(UB-LB)}{\sum_{i}(\sum_{j} Z_{ij} - 1)^2}$.

The result of the algorithm is the community assignment matrix $Z_{ij}$ and membership labels $c_j$ that match the required spatial structure and sizes of the observed communities. It will be used in social link synthesis step as described below.

2.3. Step 3: ERGM learning and social network simulation

The objective of the final step is to generate social connections for the simulated population data enriched with community assignment and household
Algorithm 1 Community Assignment Lagrangian Relaxation Algorithm

Initialize: \( \lambda = 0, \alpha = 0.01 \)

while \( \frac{UB - LB}{LB} > \epsilon \) do

Solve LLBP given \( \lambda \) values to get \( LB, c_j, Z_{ij} \)

Find a feasible solution to get \( UB \)

Calculate the subgradients \( G_i = \sum_{j=1}^{N} Z_{ij} - 1 \)

Calculate the step size \( \Delta t = \frac{\alpha (UB - LB)}{\sum_{i=1}^{n} G_i^2} \)

Update \( \lambda_i = \lambda_i + \Delta t G_i \)

end

locations. For this purpose, we adopt an Exponential Random Graph Model (ERGM) to learn the parameters from available social network data, and use the learned parameters to create the social connections between the synthetic population members. This step concludes the generation of the complete connected population enabling urban simulation with social extensions.

2.3.1. Methodology review of ERGM models

Among the large amount of literature on the existing models of social networks, one cornerstone approach is the exponential random graph models. It is a set of models that assume the probability of the existence of certain graph structure with the corresponding adjacency matrix \( a \) as belonging to an exponential family

\[
Pr(A = a) = \exp(\theta^T s(a) - \psi(\theta)),
\]

where \( \theta \) is the vector of unknown parameters, \( s(a) \) is the vector of sufficient statistics computed on the adjacency matrix \( a \), such as the counts of subgraphs, like triangles and k-stars, and \( \psi \) is the normalizing constant. It is therefore focused on deriving probabilistic models of graphs that match the statistics of the structural properties of the observed networks.

The work on Exponential Random Graphs Models can be classified into two groups concerning the statistical independence or dependence of links.

Type I ERGM: Links are independent. The sufficient statistics in this case is the total number of links. In the simplest case introduced by [13], one assumes all network edges are random variables that follow the same distribution. The model can be parameterized with more variables when more
information on the node properties is available. For example, the classical stochastic block model [50], [33] incorporates community assignment. It assumes that the probability of existence of an edge between two nodes within the same community and the probability of existence of an edge between two nodes from different communities are different. Some other models [54] use ERGM with the assumption that the edge probability between any two nodes depends on the characteristics difference between them. Variables such as age, gender [21], as well as spatial distance [54] can be considered as factors defining the edge probabilities. Such models incorporate given homophily assumptions and are used extensively in social sciences [38], [49], [37].

**Type II ERGM: Links are dependent.** Link independence, resulting in Bernoulli and dyadic dependence structures, are unrealistic assumptions in many circumstances, both empirically and theoretically. To address this shortcoming, models which assume dependence of links were proposed. The classical model is a Markov Random Graphs Model [17], where the number of triangles and number of k-stars are used as sufficient statistics, so that two edges are supposed to be conditionally dependent, given the values of all other ties. A major difficulty in Type II ERGM model inference is how to evaluate $\psi(\theta)$, as there is no feasible analytic method for approximating $\psi(\theta)$ for large networks [36]. Various Monte Carlo schemes have been proposed to approximate $\psi(\theta)$, but two fundamental difficulties of this type of ERGM model remain. First, the estimation result using Markov Random Graph Models is not robust, and Monte Carlo methods converge to ERGM model with link independence within mixing time. Second, Markov Random Graph Model is not scalable enough to handle large network inputs, which means that the parameter estimation in practice turns to be computationally slow if at all possible.

2.3.2. **Community-distance ERGM**

Given the limitations of conventional ERGMs mentioned above, we propose the “community-distance” ERGM model where we make the following assumptions: (1) links are independent; (2) people who are characteristically close to each other are more likely to get connected; (3) people in the same community are more likely to get connected with each other.

We therefore aim to infer the parameters $\theta$ based on observation of adjacency matrix $A$, characteristics distance matrix $F$ and community labels $C$ from available social networking data. The learned parameter $\theta$ are then used to generate links for the simulated population, given the characteristics
distance matrix $\tilde{F}$ and community assignment $C$ where $C_i = j$ indicates that the community label of individual $i$ is $j$.

The probability of the social graph in our model is:

$$ Pr(A = a | X, F) = \frac{1}{K} \exp(\left[ \theta_1 \theta_2 \theta_3 \right] \begin{bmatrix} \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij} \\ \sum_{\{(i,j):C_i=C_j\}} A_{ij} \\ \sum_{i=1}^{N} \sum_{j=1}^{N} F_{ij} A_{ij} \end{bmatrix} ) \quad (7) $$

where $\sum_{\{(i,j):C_i=C_j\}} A_{ij}$ represents the number of edges which connect two nodes belonging to the same community, and $\sum_{i=1}^{N} \sum_{j=1}^{N} F_{ij} A_{ij}$ represents the sum of the distances between pairs of nodes connected by an edge. To transform this problem into standard logistic regression problem, we introduce a random variable $Y_{ij}$, with $Y_{ij} = 1$ when $A_{ij} = 1$, $Y_{ij} = -1$ when $A_{ij} = 0$. Model parameters can then be estimated based on maximum likelihood, given by

$$ l(A = a; \theta) = \log(P(A = a | X, F)) = \sum_{i=1}^{N} \sum_{j=1}^{N} \log(P(A_{ij} = a_{ij})) = \sum_{i=1}^{N} \sum_{j=1}^{N} \log(\frac{1}{1 + e^{-y_{ij}s_{ij}}}) \quad (8) $$

where

$$ s_{ij} = \begin{cases} \theta_1 + \theta_2 + \theta_3 F_{ij}; & \text{if } C_i = C_j \\ \theta_1 + \theta_3 F_{ij}; & \text{otherwise} \end{cases} \quad (9) $$

Unlike Type II ERGM models where links are dependent, this model specification does not require MCMC procedure to generate possible graph structures, and maximum likelihood estimation provides numerically stable and robust parameter estimation $\hat{\theta}$. The social network for the synthetic population $\tilde{A}$ can then be simulated given the characteristics distance matrix $\tilde{F}$ and community assignment $\tilde{C}$, with

$$ Pr(\tilde{A}_{ij} = 1) = \frac{1}{1 + e^{-\tilde{s}_{ij}}}, \quad (10) $$
where
\[
\tilde{s}_{ij} = \begin{cases} 
\hat{\theta}_1 + \hat{\theta}_2 + \hat{\theta}_3 \tilde{F}_{ij}; & \text{if } \tilde{C}_i = \tilde{C}_j \\
\hat{\theta}_1 + \hat{\theta}_3 \tilde{F}_{ij}; & \text{otherwise.}
\end{cases}
\] (11)

The following Sections of the paper illustrate the use of the introduced methodology, describing the practicalities of the applications of methods to the particularities of available data.

3. Data

Model specifications for an application of the described methods depend on the availability of data. This section describes two typical data sources that are used to illustrate the synthesis of individuals for a given type of the households, and a corresponding social network of the required community structure for this synthetic population. Due to the privacy protection regulations, it is unlikely that the two components of data will become readily available in a way enabling matching the users across two sources, motivating the development of the population synthesis methodology presented in this paper.

3.1. American Community Survey

The first data source we utilized in the experiments part of this paper is a typical household survey data, known as the Public Use Microdata Sample (PUMS) of the American Community Survey (ACS). It contains multiple socio-economic parameters of the household members for a sample of the households in the region. This sample size typically ranges from 1% to 10%, and is maintained by regional governments and agencies. Apart from this micro sample data set, the aggregated marginal distribution of the population totals at the block group and census tract levels is also provided.

In the presented work, the region of study encompasses the San Francisco Bay Area: a 7000 square-mi region spanning the nine counties under the jurisdiction of the Association of Bay Area Governments. According to updated information from Metropolitan Transportation Commission and the Association of Bay Area Governments, there are 7 million people residing in nine counties and 101 cities. There are in total 1588 census tracts and 72 Public Use Microdata Areas. The data available from the ACS database carries the records for 439525 people from 132311 households, which corresponds to the sampling rate of 6.1%.

By the ACS classification, households are divided into several different types based on the household size and structure. In this paper, we use 2-people
households as an example. There are a total of 23895 2-people households in the PUMS of the study area.

3.2. Social Network Data

Social network data sources range from small scale samples collected within research projects to massive repositories of users data kept by online social network companies and telecom operators. One source of the latter type is known as the Call Detail Records (CDR), which is a standardised format of call logs collected by the operators of cell phone networks. Below we give a brief introduction to CDRs data, and then elaborate on how we construct the social networks from these records.

Collected by cellular network operators for billing purposes, CDR datasets contain several features that have helped fuel the burgeoning field of computational social science. Each record describes a communication event on the cellular network. It contains Universally Unique Identifiers (UUIDs) representing the anonymized calling (or texting) and receiving individuals, the time of initiation, its length (if it was not a text), and the unique identifiers of the cell towers at the outgoing and incoming locations (see Figure 2 for an example).

3.2.1. Data preparation for network generation

To construct a social network using CDR data certain pre-processing operations are often necessary in order to filter out spurious calls, marketing calls, and other interactions not necessarily indicative of social contact [5, 28, 34, 35]. Regarding the construction of the CDR network there are multiple ways to represent edges between two individuals. Key choices depend on the desired network representation to be weighted/unweighted and directed/undirected. For the purpose of population synthesis, we choose to construct the social network as unweighted and undirected. We add an edge between individuals $i$ and $j$ if individual $i$ called/received a call or send/received a message from individual $j$, and it has been reciprocated within the time span of the dataset. In the sample available to us from a
telecom operator, the resulting network representation consists of 1,321,765 edges and 343,299 nodes. This way of constructing the CDR network is based on Section 2.1 of [4] and Section 2.1.1 of [31]. No personally identifiable information is available due to privacy protection, making it impossible to match the users to the community survey sample, or to assign the users with the exact set of the socio-economic parameters. Instead, the key objective of the proposed methodology is to infer the structural properties of the social networks in order to reproduce them in its synthetic version.

3.2.2. Home and work locations

The first characteristic that is required for social network synthesis is the spatial spread of the detected communities. It defines the decay of the edge probability with distance between nodes as well as the geographical boundaries between the communities that are known to be different from the administrative or other artificially defined divisions [28]. A set of “anchor” locations such as home and work are required in order to be able to define characteristic spatial structure of communities that we use at the assignment algorithm described above in Section 2.2. There are mainly two popular ways for home and work detection: first, a Gaussian mixture model is adopted to model locations centered around home and work [8], second, “home" is defined as the location where the user spends more than 50% of her/his time during night hours, similarly, “work" is defined as the location where the user spends more than 50% of her/his time during day hours [27]. Alternatively, “home" is defined as the location where the user spends most of her/his time during the night hours [57]. In this paper, we adopt the similar approach as in [57] that we define the most frequently visited locations as “home" or “work", and we define 6pm-8am as night hours and 8am-6pm as day hours.

3.2.3. Community Detection

The second characteristic of interest is the community structure of the social network. To obtain the largest community size $M$ and the smallest community size $m$ for the improved two-stage Lagrangian relaxation method, we need to conduct community detection on Call Detail Records social networks to obtain these features.

Within the variety of state-of-the-art methods of community detection, two main approaches are modularity-based methods and conductance-based methods [26], with modularity and conductance correspondingly serving as the criteria for community detection. The majority of studies involving
CDRs for community detection use modularity due to their focus on intra-community size and homogeneity. For example, a common approach is to use the well-established Louvain method [5], as exemplified by [53] and [24], with extensions such as fast hierarchical agglomerative clustering based on the modularity metric [18], [46]. However, modularity-based methods fail to capture small-scale community patterns [30] typical for small social groups at the scale of households that we are interested in. Thus we propose to use conductance-based methods [26], particularly the local graph clustering [16], [48], for finding realistic upper and lower bound for community size required for the assignment step described in Section 2.2. We refer the reader to [15], [2] for the detailed survey of conductance-based methods.

4. Experiments

We provide an illustrative experiment by generating a synthetic connected population within the San Francisco Bay Area, California. The results at each step of the methodology are described in detail for the county of Napa, that consists of 108 census tracts, and contained a total of 48'876 households with 131'556 residents according to the US Census Bureau.

4.1. Household synthesis

A specific Bayesian network model has to be produced for each household category, calibrated from micro-census sample, and applied to synthesize the required number of households of each type following the available aggregate numbers of household types in the area of interest.

For simplicity of the illustration, we report on the results of simulating the 2-people households. Out of the total set of observations with 500 variables contained in the PUMS, we restrict the example to illustrate the results for 23895 2-people households. The latter consist of the head of the household (the householder) and a spouse or a domestic partner. For the set of socio-economic parameters, we choose the variables that are of interest to urban planners and transportation researchers, that are: SEX (sex of householder), SEX_S (sex of householder spouse), AGE_index (age level of householder), AGE_S_index (age level of householder spouse), PINCP_index (income level of householder), PINCP_S_index (income level of household spouse), VEH (number of vehicles owned by this household). We factorized raw input data from PUMS to categorical variables based on the corresponding ACS.
Figure 3: The learned structure of the Bayesian network for 2-people households.

age and income level bins so that it can be matched with aggregated marginal distributions in full population synthesis.

SynthPop API \([56]\) provides a convenient tool for dealing with ACS data.

4.1.1. Bayesian Network Structure and Parameter Learning

Within each type of a household, the dependency structure between the chosen variables has to be defined. This problem is known as the structure learning. By applying the tabu search methods \([51]\), we have found out that the estimated model structure is not robust when we bootstrap or use multiple data subsets. This behaviour was not reported in \([51]\). To overcome the undesirable outcomes and constrain the dependencies based on the domain knowledge and common-sense relationships, we defined a “whitelist”. The whitelist defines a set of relationships in structural learning procedure that must be preserved and are guaranteed to be present in the final graph. The whitelist used in our experiments included the gender/income dependence, which means the arrow, \(\text{SEX} \rightarrow \text{PINCP\_index}\), is whitelisted. Figure 3 presents an estimated model structure for 2-people households.

Parameter learning for the fixed structure of the Bayesian network is straightforward. We used an implementation of \([44]\) for the parameter learning and sampling from the final model for population synthesis. To keep the presentation concise, we report on the representative results of the parameter estimation without providing a full set of tables for conditional distributions between pairs of variables (shown in Figure 3). Figure 3 shows the joint
distribution of owner income and spouse age of PUMA and simulated data from Bayesian network, and Figure 4 shows the joint distribution of owner age and spouse age of PUMA and simulated data from Bayesian network. It illustrates that the Bayesian network captures the joint distribution of variables. One can clearly observe that the pairs of variables are not independent, and the dependency structures between the variables household members is preserved.
4.1.2. Synthesis with Bayesian networks and marginal distributions

Marginal distribution or joint distribution of two or more variables, which come from aggregated census data, can be used as an evidence to control the quality of sampled populations from the inferred Bayesian network [51]. However, it is instead desirable to match the observed marginal distributions precisely. Although one can get access to all aggregated feature variables for different geographical zones, in Bayesian network models, it is only possible to precisely match the distribution of the mother node. That is because based on the conditional distribution theorem, once the distribution of mother nodes and the conditional probability table are known, the distribution of child nodes are uniquely defined [22], [23], [7]. Thus it is impossible to satisfy the conditional distributions while satisfying the marginal distributions of all variables at the same time. A theoretically justified method for the latter is an open research problem. An acceptable practical strategy is to keep the Bayesian network as simple as possible, only capturing the dependencies between variables with strong statistical significance in addition to the white-listed ones. In our application, we have observed the best results while maintaining the marginal distribution of the sex of the head of a household (mother node “sex” in Figure 3) according to the gender distribution from the aggregated census data.

4.2. Community Assignment

At this step, it is required to obtain the necessary parameters for community assignment, including the $F_{ij}$ matrix for the simulated population of the households, as well as the lower bound $m$ and upper bound $M$ of community size, based on the community detection results on available social network data set.

4.2.1. Community detection result

Structural properties of the social network in the presented application were detected from the available CDR data described in Section 3. We have used the parallel local graph clustering method [48] implemented in Ligra [47] for the community detection. Figure 6 shows the histogram of the communities sizes of community detection result using local graph clustering. The communities were found to be clustered geographically, however keeping the different characteristic spatial scale of their spread withing the high and low population density areas. Home and work locations were detected with an algorithm described above, and assigned a location uniformly at random.
within a census tract. Spatial maps presented in this work only contain information on the synthetic population. We have found the sizes range for $m$ and $M$ to be 30 and 50 respectively. We then aimed at reproducing the spatial proximity and geographical boundaries within the detected community and the synthetic population, using the pairwise distances between the locations of the synthetic households as elements of $F_{ij}$. Having introduced a generic method of assignment communities to reproduce the observed spatial structure, we leave the detailed study on modelling $F_{ij}$ and the relative importance of social versus spatial factors to further research.

4.2.2. Community Assignment Experiments

With derived values of $m$, $M$ and $F_{ij}$ we then perform community assignment. Since the large-scale assignment problem is quite time-consuming to solve, we implement the community assignment using CPLEX solver [11] and parallel computing paradigm known as Message Passing Interface (MPI) [19]. We randomly partition the input data into K groups, each with 300 individuals, and use CPLEX solver for each sub-problem, coordinated by MPI. We have controlled the running time of CPLEX by setting the relative MPI gap tolerance to 20%. Because we are not interested in finding the exact minimum value of the original problem, we can tolerate an approximate assignment. CPLEX with MPI procedure was found to provide a sufficiently small value for objective function while meeting all the constraints. We found
that it is also much faster than solving original large mixed integer program.

Figure 7 illustrates the assigned communities color-coded on a map overlay, illustrating the spatial spread of the communities. One can notice the characteristic structure of more spatially homogeneous communities in the rural areas and higher overlapping ones in the suburban zones and the county capital town as observed in previous studies [29].

4.3. Step 3: ERGM learning & simulation

In this step, we apply the community-distance ERGM model implemented using CVX solver. We learn model parameters based on a subsample of a CDR network, and generate links for synthetic individuals. We then visualize the communities on a map overlay, and explore the network statistics of the simulated connected population.

4.3.1. ERGM learning

In our experiments we tried to adopt the type II ERGM model with dependent links, using the current state-of-art R ERGM implementations: (1) block exponential random graph model from “blkergm” package [55], and (2) exponential random graph model with local dependence [43] from “hergm”
model\textsuperscript{[42]}. However both models were found to be not scalable to efficiently deal with the sampled CDR networks. Based on this finding that there is no existing scalable parameter learning methods for Markov Random Graph Model that could efficiently handle large networks \textsuperscript{[41]}, we propose to use the model where edges are independent. However, even with the state-of-art R \texttt{ergm} packages \textsuperscript{[25]} for the type I ERGM models, the estimation result is still not satisfiable because the implementation involves Monte Carlo MLE estimates. Here we report on the results achieved by maximizing the convex maximum likelihood function of the community-distance ERGM model. We have implemented it using CVXPY \textsuperscript{[12]}.  

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Estimate</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>-0.8803</td>
<td>0.00334</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.1182</td>
<td>0.82880</td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>-14.5637</td>
<td>&lt; 1e-04</td>
</tr>
</tbody>
</table>

Based on Equation 10 and Equation 11, and estimation result of Table 2, we observe that the probability of a connection decreases with distance. The probability of two individuals getting connected when they are in the same group is bigger than when they are not in the same group, since we have $\frac{1}{1+e^{-(\theta_1+\theta_3*D_{ij}+\theta_2)}} > \frac{1}{1+e^{-(\theta_1+\theta_3*D_{ij})}}$ because $\theta_2 > 0$, and it corresponds to the increasing trend of sigmoid function. This indicates that people who are in the same community are more likely to get connected.  

4.3.2. ERGM simulation  

With parameters learned from community-distance ERGM learning step, we are able to simulate the social connections for the simulated households. We illustrate and report on the properties of the simulated social networks. Figure 8 shows the degree rank plot of simulated social networks. As one can see the network is sparse. Figure 9 shows the histogram of the distance between two individuals connected by an edge. One can see that the closer two individuals are, the more likely they get connected, following the expected behavior from the model specification and estimation results. Figure 10 shows a geographical visualization of simulated social networks. We also attempt to illustrate that people who are in the same community are
more likely to get connected (given the same distance). We provide Figure 10 (right), which is the visualization of 10 communities within the simulated social network. As one can see from Figure 10 (right), there are people living far away that are connected, but they are most likely to be in the same community in accordance with the parameter estimation results.

5. Conclusion and Future Work

This proposed framework for connected population synthesis is applicable to cities or metropolitan regions where data availability allows estimating the component models. The framework utilizes both traditional data sources such as household survey and census data (such as an ACS PUMS (public micro sample) and ACS aggregated census data in the context of the United States), and social network information that can be available for the region.
form cellular records or social media data. We implemented the proposed methods in code using state-of-art R packages \cite{25,44}, and optimization toolboxes CPLEX \cite{11} and CVX \cite{12}, applying an MPI parallel computing approach \cite{19} to guarantee scalability. The code we developed in this work is available at \url{https://github.com/DanqingZ/CPS_TRC}.

The presented work has faced several limitations that we are expecting to address in future research. Particularly, we expect new developments to emerge along the lines of:

- Amending the Bayesian Networks modeling step with an advanced method that simultaneously allows: (1) fitting the observed marginal distributions by constructing a constrained optimization problem where the objective function is the data likelihood under a given network structure, and marginal distributions are the constraints; (2) extending the current work to generate hierarchical structures of the social networks;

- Generating more realistic home locations based on third-party real estate data;

- Replacing ERGM with more scalable social network generation method that can both handle large input networks and account for link dependence, particularly incorporating the effect of the households structure;
• Thoroughly applying the synthetic population within a larger-scale agent-modelling framework.

A practical application of the proposed methods has demonstrated its usefulness. We presented an example illustrating the application of the approach to simulating a connected population for the Napa county, California, describing the modelling choices we made. The results have shown that the simulated connected population successfully captures a pattern from household survey data and transfers the observed community structure onto a simulated population. We believe this framework presents a starting point for connected population synthesis research.

6. Acknowledgement

This work was partially funded by a gift from AT&T. Support from the State of California Department of Transportation (CalTrans) through UCCONNECT faculty research grant program, agreement 65A0529, is also acknowledged. We thank Siyu Chen from Didi Labs for assistance with data processing.

References


[58] Yi Zhu and Joseph Ferreira. “Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation”. In: Transportation Research Record: Journal of the Transportation Research Board 2429 (2014), pp. 168–177.