A method to model population-wide social networks for large scale activity-travel micro-simulations

Conference Paper

Author(s): Arentze, Theo A.; Kowald, Matthias; Axhausen, Kay W.

Publication date: 2012

Permanent link: https://doi.org/10.3929/ethz-b-000038374

Rights / license: In Copyright - Non-Commercial Use Permitted
A Method to Model Population-Wide Social Networks for Large Scale Activity-Travel Micro-Simulation

Theo A. Arentze*
Faculty of Architecture, Building and Planning
Urban Planning Group
Eindhoven University of Technology
The Netherlands
E-mail: t.a.arentze@bwk.tue.nl

Matthias Kowald
Institute for Transport Planning and Systems (IVT) ETH Zurich
8093 Zurich
Switzerland
E-mail: kowald@ivt.baug.ethz.ch

Kay W. Axhausen
Institute for Transport Planning and Systems (IVT) ETH Zurich
8093 Zurich
Switzerland
E-mail: axhausen@ivt.baug.ethz.ch

* Corresponding author

Submission date: July, 2011
Word count: 7288 + 4 tables and 2 figures

Submitted for presentation only to
Transportation Research Board
91st Annual Meeting, January 2012
Washington, D.C.
ABSTRACT

Social-leisure activities account for an important and increasing segment of travel in modern societies. Yet, these activities are least understood in current activity-based models of travel demand. In this paper we propose a model to generate population-wide social networks that in the context of large-scale micro-simulation of travel demand provide a basis for modeling social interactions. The proposed model consists of a friendship formation model formulated in the RUM framework, and a component to simulate the network in a population. We show how the friendship model can be estimated by loglikelihood methods on observations of personal networks. In an application to the Swiss context, we demonstrate the estimation and ability of the model to reproduce relevant characteristics of networks, including for the first time simultaneously geographic distance, attribute similarity (homophily), size of personal networks (degree distribution) and clustering (transitivity). We conclude that the model, in combination with current methods to generate synthetic populations, offers a basis to model social-leisure activities and associated travel in more rigorous behavioral ways than previously possible.

Keywords: synthesis, social network, activity-based modeling, travel demand modeling, social-leisure activities, micro-simulation.
1. INTRODUCTION

Micro-simulation is a well-established technique in travel demand modeling and has gained importance with the emergence of the activity-based approach in this field. In activity-based models, the unit of prediction is broadened from trips to complete daily activity-travel patterns of persons. Micro-simulation is the only tractable way to represent fully the observed heterogeneity in behavior on this level. Over the last twenty years, large-scale micro-simulation models have been developed and recently several systems have reached the stage of maturity and are making the transition to practice. Examples of such models are Albatross (Arentze and Timmermans, 2004), Matsim (Charipar and Nagel, 2005; Meister et al., 2005), Cemdap (Bhat et al., 2004), Famos (Pendyala et al., 2005) and Tasha (Roorda et al., 2008).

Social activities are still receiving limited systematic attention in these systems. Traditionally, attention is focused on modeling of subsistence and maintenance activities, such as work, school and shopping. Prediction of social activities is often still based on simple drawing from observed distributions (of timing and location) in activity data (Axhausen 2008). Such statistical approaches offer no guarantee that generalizable characteristics of behavior are captured. Yet, social activities account for a large and increasing proportion of travel demand in modern societies. In the Netherlands, for example, more than 15% of all trips and more than 20% of the total volume of passenger kilometers are made for social visits alone. Furthermore, most leisure activities are social in nature, as they tend to be conducted jointly with friends or family.

Explicit representation of the social network of a simulated population would provide a way to model social activities in a more satisfactory behavioral manner. Although several authors stressed the relevance of this, work in the area of modeling social networks is still scarce in transportation. Only recently social networks attracted the attention of transportation researchers. The first studies in this emerging tradition focused on data collection instruments and analysis (Carrasco and Miller 2006; Frei and Axhausen, 2007; Schwane, 2008; Axhausen, 2008; Molin et al., 2007; Carrasco et al., 2008; Van den Berg et al., 2009). Furthermore, several studies have begun to explore the use of agent-based models to simulate social activities in time and space (Arentze and Timmermans, 2008; Ronald et al., 2008; 2009) and econometric frameworks (e.g., Dugundji and Walker 2005, Paez and Scott 2007, Goetzke 2008). However, these models assume that the social network for a simulated population is given. Hackney and Axhausen (2006), Marchal and Nagel (2006) and Illenberger et al. (2009) are the first to explore the application of network generation models for social travel simulation. As the authors admit, however, the network models used are either small scale or leave important (spatial or topological) characteristics of networks out of consideration.

The purpose of the present study is to develop a method to generate population-wide social networks for use in traffic micro-simulation systems. We build on earlier work where we developed a model to generate friendship networks (Arentze et al., 2011). This model fits the requirements of large-scale travel demand modeling, but ignores an important tendency of real social networks, namely clustering. The intended contribution here is to extend this model such that generated networks accurately represent this tendency. Furthermore, we estimate and test the model using a new large dataset of personal networks collected in Switzerland. We focus on (non-family-based) friendship relationships which, besides family, constitute a major category of social life. We emphasize that generation of a social network is only a first step in predicting social activities and travel. The social interactions and related trips emerging from a social network remains to be modeled. An explicit representation of the network provides a behavioral basis for this.
2. RELATED WORK

Modeling of (social) networks has a long tradition in disciplines such as sociology, physics and mathematics. A common objective of approaches that have been developed is to reproduce known characteristics of networks such as homophily, reciprocity and transitivity. Homophily refers to the phenomenon that similar people are attracted to each other (Byrne, 1971; McPherson et al., 2001; Bidart and Degenne, 2005). Reciprocity is a tendency that social relationships are symmetric, i.e. that if A is a friend of B, then B is also a friend of A; this is the norm in friendship networks. Transitivity is the phenomenon that if a relationship exists between A and B, and between B and C, then there is an increased probability that also a relationship exists between A and C (Bidart and Degenne, 2005). A consequence of this latter tendency is that networks display a larger degree of clustering than can be explained by homophily alone (Louch, 2000). Apart from these topological characteristics, geographic distance has a strong influence, in the sense that the probability of a friendship relationship strongly decreases with geographic distance (Mok et al., 2007; Frei and Axhausen, 2007).

Exponential random graph (ERG) models represent a dominant approach in this area (for a review, see Robins et al., 2007). These models predict in a probabilistic fashion the social relationships among a given set of actors as the structure of a graph based on an exponential function of possible configurations. Different models emerge from assumptions about dependencies between social ties that lead to certain substructures in the graph. In the most basic form, these substructures concern reciprocal and non-reciprocal relationships between pairs (dyads) or triples (triads) of nodes. Furthermore, node attributes can be taken into account so that, for example, homophily tendencies can be incorporated (Robins et al, 2001). The likelihood of observing certain structures is determined by parameters that are estimated based on network data. Depending on the model formulation, the estimation of parameters is not an easily solved problem. Maximum likelihood is the general principle on which estimation methods are based. However, standard likelihood estimations soon become intractable for larger networks. Markov Chain Monte Carlo methods are gaining interest as a technique to solve estimation problems for some of the existing models. This still is an active field of research (Robins et al., 2007).

ERG models do not intend to model the actual choice behavior of individuals (or firms). So-called actor-based models use a different approach that does have this intention. Specific actor-based models have been developed for testing specific hypotheses or deriving implications of a specific social theory. Recently, a more unified framework is emerging known as stochastic actor-based models. Snijders et al. (2010) give a review of this field. These models are interesting here as they include an error theory so that they are estimable on data. The models focus on behavior of actors and network dynamics. Parameters of the model are estimated based on longitudinal data of network evolution of some studied community (e.g., a school class). Although the approach is appealing, models are in an early stage of development and, especially, issues pertaining to model estimation need to be solved before they can be applied to large-scale longitudinal data sets.

The existing approaches briefly reviewed above are not readily applicable to travel demand modeling. Generally, they are focused on relatively isolated communities, such as classrooms or workplaces. In a national or regional population, however, geographic distance and place are important factors. An attempt to model social networks explicitly in geographic space is Illenberger et al. (2009). The model they propose is able to handle large populations and reproduce spatial characteristics of networks. Being in an early stage of development, their model does, however, not incorporate tendencies of homophily and transitivity. The model proposed by Arentze et al. (2011) considers both geographic distance and social distance (homophily). They model friendship formation as decisions of agents within a...
random-utility-maximization (RUM) framework. Since RUM is also the standard framework for modeling activity-travel choices, their friendship model can be integrated seamlessly in the larger systems. Their model does, however, not consider transitivity (the common-friend factor) implying that the networks generated do not display clustering.

In the present paper we extend the framework proposed in earlier work (Arentze et al., 2011) with the aim to generate networks that are accurate in terms of observed degree of clustering as well as degree distribution (the distributions of number of friends), distance distributions of social ties and homophily tendencies.

### 3. MODEL

In this section we first briefly introduce the model developed in earlier work (3.1). We then consider clustering in social networks (3.2) and the proposed extensions to incorporate this tendency in the model (3.3-3.6).

#### 3.1. Basic model

The core part of the earlier model predicts for any two random individuals in a studied population the probability that a friendship exists between the two persons. To take into account the three major known characteristics in friendship ties, the authors proposed the following utility function:

\[
U_{ij} = V_{ij}^Q + V_{ij}^D + V_{ij}^C + \varepsilon_{ij}
\]  

where \( U_{ij} \) is the utility of a tie between persons \( i \) and \( j \), and \( V_{ij}^Q, V_{ij}^D, V_{ij}^C \) are structural utility components related to homophily (Q), geographic distance (D) and existence of common friends (C), and \( \varepsilon_{ij} \) is a stochastic utility component (i.e., error term). In line with the known reciprocity in friendship relations, the model assumes that all social ties are symmetric, i.e. that if \( i \) is a friend of \( j \), then \( j \) is also a friend of \( i \). The authors stress that utility does not imply that only rational considerations of costs and benefits play a role. Rather the authors consider a statistical interpretation of the term as equally valid. In a statistical interpretation, the utility represents a tendency or probability (e.g., having a common friend increases the probability of a social tie). Whichever interpretation is used, the model assumes that the utility of a tie is the same for the involved persons \( i \) and \( j \) (\( U_{ij} = U_{ji} \)). In reality perceptions of utility may differ. It is possible to refine the utility function to represent perception differences.

Apart from preferences, opportunities have an influence on the probability that a friendship is present. Maintenance of a friendship requires time (and possibly expenditure) and, hence, competes with other activities for limited time (and possibly limited income) (Hummon, 2000). Opportunity costs may differ between the persons involved and are represented in the model as a threshold utility value. The model assumes that a social tie is present only if the utility exceeds the threshold values of both persons. This means that the highest threshold across the two persons determines the probability of a link, given its utility. Hence, the probability of forming a tie is defined as:

\[
P(i \leftrightarrow j) = \Pr(U_{ij} > \max[u_i^o, u_j^o])
\]
Where, as before, $U_{ij}$ is the utility of the tie (defined by Eq. 1) and $u_i^o$ and $u_j^o$ are threshold values for person $i$ and person $j$, respectively.

It is important to note that the thresholds do not only represent opportunity costs. In the model, the threshold is a parameter that an agent sets also depending on his opportunities to meet other people in settings where social contacts may result. With higher frequency of relevant opportunities, an agent should set the threshold higher in order to limit the growth of his network in line with his time budget constraints and needs for social contact. Vice versa, the smaller the pool of possible friends given the reach of mobility tools of the agent, the lower the threshold should be for the agent.

The second component of the model is concerned with creating (artificial) meetings between individuals (agents). Although behavioral models could be considered, the current model assumes a straight-forward procedure, where every agent encounters every other agent once. In every encounter a decision is made whether a friendship is present based on the model given by equation (1) and (2). Thus, in a population of $N$ agents, there are $N (N - 1) / 2$ encounters and a corresponding number of decisions. The result of this process is a population-wide social network that shows degrees of homophily, spatial proximity, transitivity and density depending on relative values of the utility components $V_{ij}^Q$, $V_{ij}^D$, $V_{ij}^C$ and $u_i^o$. We emphasize that encounters in the model are artificial; they do not represent actual meetings but rather a possible meeting. Thus, in effect, the probability defined by Eq. (2) is about the joint events of actually having met and become friends.

The model has strong similarities with the model proposed by Mayer and Puller (2008). The friendship model the latter authors propose includes similar terms and is applied in a similar process of simulating interactions. Important differences are, however, that their model is not based in the RUM framework and considers relatively small communities, such as students on a university campus, where geographic factors do not play a role.

The model we described in this section has only partly been developed. As the earlier authors admit, the common-friend factor $V_{ij}^C$ is not handled properly. In the presence of friendships, conditions for existence of new friendships change to the extent that common friends are treated. The authors indicate that the linear process of checking for friendships should be replaced by an iterative procedure, but they do not develop this. Since the mechanism is deficient, the networks generated in the earlier work do not display the degree of clustering that we see in real networks.

![FIGURE 1 Illustration of transitivity in networks](image-url)
3.2. Clustering in social networks

Empirical studies indicate that clustering is a pronounced characteristic of social networks in general. A commonly used method to measure the degree of clustering (transitivity) in social networks is based on the following index:

$$C = \frac{3 \times \text{number of triangles}}{\text{number of triples}}$$ \hspace{1cm} (3)

A triple is a combination of two edges in a network that are interconnected by a node. A triangle is a triple that is closed, i.e. a triple where each node is connected to each other node. Figure 1 shows an arbitrary example. In this network, 4 – 6 – 5 is an example of an open triple and 6 – 8 – 7 is an example of a triangle, i.e. a triple that is closed by edge 6 – 7. Since there are in total 3 triangles and 20 triples in this network, the clustering index for this network equals $C = \frac{9}{20}$. This indicates that 9/20th of all pairs that have a common friend, are friends.

Real social networks show an above-chance degree of clustering. For example, Mayer and Puller (2008) find cluster coefficients of 0.17 and 0.27 among students’ Facebook contacts at two universities in the US. In the dataset of leisure contacts of individuals used in the present study (described below), we find a cluster coefficient of 0.206. These numbers clearly show the significance of clustering in social networks and, hence, the relevance of incorporating this tendency in a model of social-network generation.

3.3. Modeling transitivity

The modeling of transitivity within the framework outlined above has consequences for both the friendship model and the encounter-simulation process. We start with the friendship model.

A first notion is that the presence of a common friend increases the probability that the two individuals meet (e.g., at the birthday party of the common friend). Since the probability of meeting (in a meaningful setting) is an element of the threshold, we model the threshold as a function of the presence of common friends as follows:

$$u_{ij} = \begin{cases} u_i^o & \text{if } C_{ij} = 0 \\ u_i^o - \theta & \text{otherwise} \end{cases} \hspace{1cm} (4)$$

where $u_{ij}$ is the threshold for individual $i$ regarding $j$, $u_i^o$ is the basic value of the threshold, $C_{ij}$ is a binary variable indicating whether or not $i$ and $j$ have a common friend and $\theta \geq 0$ is a parameter representing the increased probability of a (real) meeting. In addition to a threshold effect, we may hypothesize that there is also a utility effect. A common friend offers an opportunity to engage in joint activities involving three (or more) persons which may add value to a friendship especially in the domain of leisure activities. Although this is certainly plausible, in the present model we conveniently assume that there is no utility effect or that the effect can be ignored. Thus, we define:

$$V_{ij}^c = 0 \hspace{1cm} (5)$$

Thus, the model assumes that transitivity is solely the result of an increased probability that two persons meet when they have a common friend.
To account for the common friend factor, we suggest the following extended process of simulating friendship ties:

Round 1  Determine primary ties by deciding for each pair of individuals whether a social tie exists between them

Round 2  Determine secondary ties by deciding for each pair of individuals who are not yet connected to each other and who have at least one common friend through primary ties, whether a social tie exists between them

In this procedure, each pair of individuals is processed twice in two rounds: in the first round to determine whether a tie exists between them without and in the second round with the support of a common friend if they have one. It is worth noting that Rule (7) states that at least one common friend must exist: the number of common friends is not considered relevant for secondary-tie decisions. A difference between primary and secondary ties is that for the former the higher value of the threshold applies ($u_i^o$) and for the latter the lower value ($u_i^o - \theta$).

In theory, additional rounds of friendship-formation decisions could be considered by adding the follow step:

Round 3  If secondary ties have been created in the last round, upgrade all secondary ties to primary ties and repeat Round 2.

This third step may lead to detection of additional ties as adding secondary ties may create new triples and, hence, new opportunities for friendships supported by a common friend. This added step would be useful for a model that intends to represent dynamics of social-tie formation. However, we are not interested in modeling a time trajectory, but rather to generate a network for a given moment in time that represent an initial state of a population in a micro-simulation. For that purpose the two-round model suffices and is preferred as it is robust and saves computation time.

### 3.4. Friendship formation prediction

Assuming that the error term, $\varepsilon_{ij}$, in Eq. (1) is Gumbel distributed, a (binary) logit model can be used to derive probabilities:

$$p_{ij} = \frac{\exp(V_{ij})}{\exp(\max[u_i^o , u_j^o ] - \theta ) + \exp(V_{ij})}$$

where $V_{ij} = V_{ij}^Q + V_{ij}^C$ ($V_{ij}^C = 0$) is the structural utility component and, as before, $u_i^o$ is the threshold of a person $i$ and $\theta$ is the threshold reduction parameter which applies only for secondary ties (i.e., $\theta = 0$ for primary ties). To derive probabilities of secondary ties, the following correction is needed:

$$p_{ij}^{r2} = (p_{ij}^2 - p_{ij}^1) / (1 - p_{ij}^1)$$

where $p_{ij}^1$ is the probability of a primary tie, and $p_{ij}^2$ and $p_{ij}^{r2}$ are the probabilities of a secondary tie before and after correction ($p_{ij}^1$ and $p_{ij}^2$ are defined by the logit model – Eq. 9).
The correction takes into account the condition which holds when we calculate this probability, namely that \(i\) and \(j\) are not connected by a primary tie. Given this condition, we know that the utility does not exceed the high threshold (for a primary tie). Hence, the probability that the utility exceeds the low threshold given that it does not exceed the high threshold equals \( (p_{ij}^2 - p_{ij}^1)/(1 - p_{ij}^1) \)

### 3.5. Scalability

The populations in the class of micro-simulation systems we consider can be as large as hundreds of thousands or even millions of agents. Computation time therefore is an issue. Fortunately, for such large populations an exhaustive enumeration of pairs of individuals is not needed. Consider the following enumeration method:

\[
I = \sum \sum_j 1 = N(N-1)/2y, \quad i = 1, 2, \ldots, N; \quad j = i + y, i + 2y, i + 3y, \ldots, N
\]  

where \(N\) is size of the population, \(I\) is number of pairs evaluated and \(y\) is some chosen step size. In this formula, the case of complete enumeration of pairs is represented by step size \(y = 1\), i.e. every individual \(i\) is paired with every next individual \(j\) in a population list. If step size \(y = 2\) is chosen, every \(i\) is paired with every second next individual, reducing the total number of evaluations, \(I\), by half. If the population is large enough and the threshold \(u^o\) is adapted properly, the statistical properties of the resulting networks should be the same. In general, the larger the population is, the larger step size \(y\) can be chosen without affecting the result (in terms of relevant statistical properties of the networks). Thus, for small populations the complexity of the model is of the order of magnitude of \(N^2\), but for larger populations it becomes less. For large populations the computation time scales linearly with the size of the population, i.e. has a complexity of \(aN\) where \(a\) is some given minimum of meetings that need to be evaluated for each agent.

### 3.6. Likelihood function

In this section, we develop a likelihood function for estimating unknown parameters in the utility and threshold functions of the friendship formation model. The likelihood function we propose assumes observations of social ties that can be obtained from existing survey instruments. Specifically, we assume data of personal networks of a random sample of individuals from a studied population. We assume that the social ties within (and between) personal networks constitute independent observations. Although there might be unobserved correlations between observations of a same ego (an individual in the sample), this assumption is not critical; the validation of the model will be based on networks generated (below). A complication is that existing datasets do not differentiate between primary and secondary ties, i.e. ties that were formed with and ties that were formed without support of a common friend. Therefore, we will formulate a likelihood function for unlabeled ties and next discuss the implications for estimability of the model.

The likelihood of a sample of ego-centric networks can be defined as:

\[
L_i = \left( \prod_{j \neq i} p_{ij} \right) \left( \prod_{j \neq i} (1 - p_{ij}) \right)
\]
where \( J_i \) is the observed set of friends of \( i \), \( J_i^\complement \) is the complementary set of all other individuals in the population and \( p_{ij} \) is the predicted probability of a tie between persons \( i \) and \( j \) by the model. The first term on the RHS represents the joint probability of all positive decisions and the second term the joint probability of all negative decisions concerning person \( i \) and all other persons of the population.

Although theoretically correct, this likelihood function is not practical. It requires full census data of a population to compute the function. Even if such data were available, estimation will not be tractable given its scale. A way to deal with this is to use a sampling method in a similar way as is standard practice in applications of discrete choice models where the choice sets are large. For the present case, a sampling method can be defined as follows:

\[
L_i = \left( \prod_{j \in J_i} p_{ij} \right) \prod_{s \in S_i} \left( 1 - p_{is} \right)^{1/h}
\]

(13)

where \( S \) is a random sample of persons from the (full) population and \( h \) is the sample fraction. Each person \( s \) in the sample stands for \( 1/h \) (similar) persons in the population, so that declining a tie with a person \( s \) stands for \( 1/h \) times doing the same with other persons with probability \( (1 - p_{is}) \). Provided that the sample is large enough, the approximation of the likelihood should be sufficiently accurate.

We can use Eq. (9) to predict probabilities \( p_{ij} \) and \( p_{is} \) in Eq. (13) assuming the setting \( \theta = 0 \). In terms of the model, this setting means that all observed ties are considered as primary ties. Obviously, this will lead to biased estimates of threshold, \( u^0 \) (or, more precisely, estimates of a threshold function for any person \( i \)). Since part of the ties will be secondary the threshold will be underestimated. A further shortcoming is that we do not obtain an estimation of \( \theta \). However, the drawbacks are less problematic as it may seem. Given the assumption that the utility function is the same for both types of ties, utility function parameters should still be unbiased. In sum, the likelihood function and data allows us to estimate utility function parameters, but does not allow us to estimate the threshold parameters \( u^0 \) and \( \theta \) without bias.

Thresholds, however, can be estimated with further data, given that we have appropriate estimates for the utility function. For a given utility function, the value of the basic threshold \( u^0 \) primarily determines the average size of personal networks, whereas the value of the threshold-reduction \( \theta \) primarily determines the degree of transitivity. This suggests that the thresholds can be calibrated based on the average size of personal networks, on the one hand, and the degree of transitivity of the whole population-wide network, on the other. Therefore, the method we propose is to estimate the utility function and a preliminary value of the basic threshold based on observed personal networks, and next estimate the two threshold parameters by means of (manual) calibration, whereby the preliminary estimate of the basic threshold parameter provides a good starting point.

4. ILLUSTRATION

4.1. Sample data

The personal network data used for estimating the model originates from a survey conducted recently in Switzerland. The survey uses snowball sampling, which is a method of sampling where individuals who participate in the survey are asked to identify one or more members of their network. At the moment of this study the survey was still going on. The sample used includes a total of 265 individuals (egos) that resulted from 4 rounds of snowballing from an initial sample of 40 individuals in the Kanton Zürich.
The survey asked respondents (egos) to report their leisure-related social contacts (alters) as well as to fill out a sociogram indicating their membership in a clique undertaking joint activities. For a detailed description of the survey method and resulting sample readers are referred to Kowald et al. (Forthcoming).

The average network size across respondents is 22.0 contacts with a standard deviation of 13.4 contacts. Given the illustration purpose, we only consider gender and age as attributes of egos and alters. In terms of age, we use a 5-way classification: younger than 23 years, 24–37 years, 38–50 years, 51–65 years, 65+ years. We measure distances between homes as straight line distances based on the X, Y coordinates of the home addresses. Table 1 shows some statistics of the sample of ties after removing outliers and imputing missing values. Ties over longer distances than 800 km are identified as outliers and are removed from the dataset. Missing values are imputed, where possible by drawing from distributions of differences between egos and alters for the attribute under concern as observed in the total sample.

A consequence of the snowball sampling method is that the sample of egos is not a purely random sample of the population. This means that an assumption of the likelihood function (observations of ties are independent of each other) is not met and, furthermore, that the statistics of the personal networks may not be fully representative for the population at large. Nevertheless the dataset is very rich and still serves the present purpose of illustrating the model. The possibility of biases in estimation results and statistics, however, should be kept in mind in the interpretation of the results below.

<table>
<thead>
<tr>
<th>Segment</th>
<th>N ties</th>
<th>Same age ratio</th>
<th>Same gender Ratio</th>
<th>Distance [km]</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ego-alter ties</td>
<td>6271</td>
<td>0.501</td>
<td>0.649</td>
<td>25.4</td>
<td>65.1</td>
<td></td>
</tr>
<tr>
<td>Alter-alter</td>
<td>4897</td>
<td>0.414</td>
<td>0.576</td>
<td>28.0</td>
<td>81.3</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>11168</td>
<td>0.463</td>
<td>0.617</td>
<td>26.6</td>
<td>72.7</td>
<td></td>
</tr>
</tbody>
</table>

As it appears, there is a clear tendency of homophily in terms of both age and gender: 46.3% of the ties involve persons from the same age class and 61.7% of the ties involve persons of the same gender (Table 1). Same-age and same-gender ratios appear to be somewhat smaller than average for alter-alter links. Average distance equals 26.6 km with a standard deviation of 72.7 km. Average distances are slightly larger than average for alter-alter links. Furthermore, the socio-gram data allows us to calculate the cluster index given by Eq. (3). The cluster index across all ego-networks is 0.206 meaning that 20.6 % of all alter pairs are friends (participate in leisure activities together). This can be taken as a valid estimate of the cluster index of the population-wide network at large.

4.2. Population data

Data of the population serves two purposes here: 1) in the estimation stage, to obtain a sample of negative friendship decisions to calculate the likelihood function (S in Eq. 13) and 2) in the prediction stage, to generate a population-wide network based on the estimated friendship formation model. The population of Switzerland was considered here as the target population given the fact that the sample (mainly) originates from this area. For the illustration purpose,
we used the Swiss micro-census (travel survey) of 2005 to obtain population data (BFS and ARE, 2007). This dataset includes 33,390 individuals. Arbitrarily, we took a weighted random fraction of 10% of this sample, using person weights provided with this survey as weights. As a result, our synthetic population for this illustration counts 3301 persons. By using weights the resulting sample should be representative in terms of person attributes for the full population. Although this sample is small relative to the full population of Switzerland, simulations showed that increasing the sample (e.g., doubling the size) did not have any effect in the later application on characteristics such as transitivity, degree distribution and so on, provided that the threshold for primary ties is properly adjusted (the larger the population the higher the threshold needs to be).

4.3. Estimation results

The estimation is based on observations of ego-alter ties. For calculating the log likelihood function (logarithm of the function defined in Eq. 13), a random sample, $S$, of 100 persons was drawn for each observed ego from the population of 3,301 persons. Tests on simulated data showed that a sample of this size suffices to estimate parameters accurately and with sufficient power (t-values). Table 2 represents the estimation results. Only a constant parameter is estimated for the threshold function. Modeling the threshold function as a function of socio-demographic and spatial variables is relevant, but is not considered here in this illustration. Several transformations of the distance variable were tried. A logarithmic transformation maximizes the goodness-of-fit and was used in the final specification (shown in the table) (See also Illenberger et al., 2011). Based on gender and age-class, similarity measures were defined in a straightforward way. In case of gender, similarity is defined by a single binary variable (yes or no same gender). In case of age, there are five possible outcomes of a relationship, ranging from same age class to 4 classes difference. Four dummy variables were included using 1-class difference as the base level, to encode this variable. As a null model we consider a specification where utility parameters are set to zero and the corresponding threshold is set according to a base probability for a tie (which results in a threshold value of 4 utility units). Relative to the null model, the rho-square for the estimated model equals 0.378 which indicates a relatively high goodness-of-fit. All parameters are strongly significant with signs as expected. We see relatively strong effects of same-gender (0.725 utils) and same-age-class (0.918 utils compared to 1 class difference). Utility of a tie decreases strongly with an increase of difference in age.

<table>
<thead>
<tr>
<th>TABLE 2 Estimation results of a basic model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null model</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>Log distance</td>
</tr>
<tr>
<td>Gender: Same</td>
</tr>
<tr>
<td>Age: Same class</td>
</tr>
<tr>
<td>Age: 2 classes difference</td>
</tr>
<tr>
<td>Age: 3 classes difference</td>
</tr>
<tr>
<td>Age: 4 classes difference</td>
</tr>
<tr>
<td>Threshold</td>
</tr>
<tr>
<td>LL</td>
</tr>
<tr>
<td>Rho-square</td>
</tr>
</tbody>
</table>
TABLE 3 Generated ties: statistics

<table>
<thead>
<tr>
<th>Segment</th>
<th>N link</th>
<th>Same Age ratio</th>
<th>Same gender ratio</th>
<th>Distance [km]</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary ties</td>
<td>19451</td>
<td>0.446</td>
<td>0.655</td>
<td>29.5</td>
<td>50.7</td>
<td></td>
</tr>
<tr>
<td>Secondary ties</td>
<td>16842</td>
<td>0.410</td>
<td>0.628</td>
<td>10.4</td>
<td>21.2</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>36293</td>
<td>0.429</td>
<td>0.643</td>
<td>20.7</td>
<td>39.9</td>
<td></td>
</tr>
<tr>
<td>All (observed)</td>
<td></td>
<td>0.463</td>
<td>0.617</td>
<td>26.6</td>
<td>72.7</td>
<td></td>
</tr>
</tbody>
</table>

TABLE 4 Characteristics of generated personal networks

<table>
<thead>
<tr>
<th>Segment</th>
<th>Network size</th>
<th>Transitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Primary ties</td>
<td>11.5</td>
<td>6.0</td>
</tr>
<tr>
<td>Secondary ties</td>
<td>10.0</td>
<td>11.0</td>
</tr>
<tr>
<td>All</td>
<td>21.5</td>
<td>16.3</td>
</tr>
<tr>
<td>All (observed)</td>
<td>22.0</td>
<td>13.4</td>
</tr>
</tbody>
</table>

4.4. Generating a population-wide social network

The model was used to generate a synthetic social network for the sample population (3301 persons). The thresholds are calibrated (manually) such that mean personal network size and transitivity index match the observed values (22.0 persons and 20.6 % respectively). Tables 3-4 show statistics of the generated whole social network with calibrated threshold parameters of \( u^0 = 1.735 \) and \( u^0 - \theta = -0.10 \). The tables show statistics for primary and secondary ties separately as well as for all ties together. The last row shows the statistics for the observed case for comparison. Since the distinction between primary and secondary ties cannot be made for the observed case, these latter numbers refer to the total set. On the outset, we should point to an imperfection in this comparison: unlike predicted ties, observed ties are taken from a sample of individuals that may not be fully representative for the population due to sampling error. Therefore, we can expect to find differences even if the model is accurate.

Table 3 shows statistics related to homophily and distance relationships. The model reproduces the observed age and gender similarity fairly accurately. The same-age ratio and same-gender ratio are of the same order of magnitude as the observed ratios. In terms of distance, we do see a difference: both the average distance and standard deviation of predicted ties are somewhat smaller than the observed ones. The difference is caused by the fact that the model underpredicts very large distances (more than 250 km) which are also rare in the observed set but nevertheless have quite a strong influence on the average and standard deviation. This is due to the log transformation, which penalizes long distance ties (See Illenberger et al., 2011). Another difference that contributes to this is that predicted ties are confined to fall within the borders of Switzerland, whereas reported ties are not. Figure 2a shows, that the model fairly accurately reproduces the distribution of ties across age-class difference categories. Figure 2b, which shows the distribution of ties across distance categories, confirms the earlier observation that the model underpredicts the largest distance category. Apart from that, the distance distribution reasonably closely matches the observed one.
FIGURE 2 Predicted versus observed network characteristics

a) Age difference

b) Distance distribution

c) Degree distribution
Table 4 focuses on characteristics of personal networks. The statistics shown are based on a random sample of 200 individuals from the (synthetic) population, mimicking the type of observations we have in the actual survey. The table shows that the average network size and transitivity for the predicted networks closely match the average network size and transitivity of the observed networks. This is the result of calibration, but nevertheless shows that the model is able to reproduce these characteristics successfully. The generated networks contain on average approximately an equal number of primary and secondary ties. The transitivity among primary ties is very low (0.048), as expected. So, according to the model, the common-friend mechanism is predominantly responsible for the level of transitivity that the network displays. The standard deviation of network sizes is somewhat overpredicted by the model (16.3 versus 13.4 persons) (Figure 2c). This overprediction might be caused by a tendency where large primary networks (result of the first round) grow more strongly in size by the formation of secondary ties than small networks do. In other words, the socially richer become richer and the poorer stay poor. This suggests that the bias can be addressed by network size dependent thresholds. However, the bias is small.

5. CONCLUSIONS AND DISCUSSION

We introduced a model to generate population-wide social networks for application in large-scale micro-simulation of travel demand. The model consists of a friendship-tie model based in the RUM framework, and a component to simulate encounters between individuals in a population. We showed how the friendship tie model can be estimated by standard log likelihood methods based on observations of personal social networks. An application on a synthetic population of Switzerland indicated that the model is able for the first time to reproduce characteristics of social networks that are important for travel-demand modeling, including geographic distance of ties, attribute similarity of ties (homophily), size of personal networks (degree distribution) and clustering of networks (transitivity). Theory and first findings suggest that the model is scalable to the large populations of the order of magnitude of hundreds of thousands or even millions of agents we are dealing with in micro-simulation systems such as MATSIM and Albatross. A very useful property of the model is that it supports sampling methods: essential properties of a generated network are maintained when an arbitrary fraction of a population instead of the full population is simulated. Key to this property is that the scale of the population (the size of the fraction) can be taken into account by adjustment of a single threshold parameter.

Arguably, the model opens a way to improve existing travel-demand models significantly. The modeling of social-leisure activities, which account for an important segment of travel in modern societies, lacks a behavioral basis in current frameworks. The model proposed in this study has the potential to change this situation; it is able to generate in a statistically correct way the connections between individuals in a population that play an important role in their decisions when, where and with whom to participate in what kinds of social-leisure activities. Methods to generate synthetic populations for micro-simulation have been on the research agenda for many years in transportation. Development of models to generate social networks, however, is still a relatively new area. Advancement in this new area can provide a means to make the move from synthetic populations where individuals are isolated (or at best organized into households) to populations where they are interconnected.

The proposed model presents a step forward in this direction. Several problems remain for future research. First, the present model does not disentangle factors that influence the evaluation of a friendship and factors that influence the probability of a meeting. Implicitly the procedure of pairing every individual with every other individual assumes uniform
chances of meeting. It is interesting to explore more behaviorally oriented models of meetings instead. An aspect of this is related to geographic space. Factors such as urban density, which vary spatially, may influence opportunities of meetings. In terms of the model this means that geographic place factors should be incorporated in threshold functions. Another aspect of this concerns the interaction between individuals’ activities, on the one hand, and formation of friends, on the other. For example, school and work places tend to provide settings where people meet in meaningful ways for starting friendships. Hence, there is a correlation between these activities and personal networks (Arentze and Timmermans, 2008). Finally, the household context deserves attention. Although through the mechanism of common friends the personal networks between members of a same household are correlated in the present model, the special character of ties between members of a same household deserve more attention.

REFERENCES


*ArbeitsberichteVerkehrs und Raumplanung* 439, IVT, ETH Zürich, Zürich.


Kowald, M. and K.W. Axhausen (Forthcoming) Focusing on Connected Personal Networks: Selected Results from a Snowball Sample, Environment and Planning A.

Louch H, 2000, “Personal network integration: transitivity and homophily in strong-tie relations” *Social Networks* 22 45-64.


