An agent-based random-utility-maximization model to generate social networks with transitivity in geographic space
Arentze, T.A.; Kowald, M.; Axhausen, K.W.

Published in:
Social Networks

DOI:
10.1016/j.socnet.2013.05.002

Published: 01/01/2013

Document Version
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the author's version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Citation for published version (APA):
An agent-based random-utility-maximization model to generate social networks with transitivity in geographic space

Theo A. Arentze a,∗, Matthias Kowald b, Kay W. Axhausen b

a Faculty of the Built environment, Urban Planning Group, Eindhoven University of Technology, The Netherlands
b Institute for Transport Planning and Systems (IVT), ETH Zurich, 8093 Zurich, Switzerland

ARTICLE INFO

Keywords:
Social network generation
Random utility maximization
Spatial choice modeling
Agent-based modeling
Likelihood estimation

ABSTRACT

Stochastic actor-based approaches receive increasing interest in the generation of social networks for simulation in time and space. Existing models however cannot be readily integrated in agent-based models that assume random-utility-maximizing behavior of agents. We propose an agent-based model to generate social networks explicitly in geographic space which is formulated in the random-utility-maximizing (RUM) framework. The proposed model consists of a friendship formation mechanism and a component to simulate social encounters in a population. We show how transitivity can be incorporated in both components and how the model can be estimated based on data of personal networks using likelihood estimation. In an application to the Swiss context, we demonstrate the estimation and ability of the model to reproduce relevant characteristics of networks, such as geographic proximity, attribute similarity (homophily), size of personal networks (degree distribution) and clustering (transitivity). We conclude that the proposed social-network model fits seamlessly in existing large-scale micro-simulation systems which assume RUM behavior of agents.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

The generation of social networks is increasingly receiving attention in agent-based modeling in various domains of human behavior. The structure of a social network determines to an important extent how information, opinions, norms and behavior-change spread through a community or a population. Traditionally, work on social-network modeling has considered actors (individuals) within the confined spatial settings of a local community, such as a school classroom, a workplace or a neighborhood, where the geographic distance between the actors is not a relevant factor (Mok et al., 2007). Increasingly, however, there is an interest in the spatial modeling of social networks involving a larger population where geographical distance between agents is an important determinant of the likelihood that a social tie exists (Moon and Carley, 2007; Barrett et al., 2009; Butts et al., 2012). An example of an area where spatial modeling is highly relevant is the simulation of epidemic spread of diseases (Carley et al., 2006). Another field of application, which is relatively new, concerns the simulation of individuals' travel behavior and resulting traffic flows in a metropolitan area or entire country for transport planning. In modern societies a large and increasing proportion of individuals' travel is motivated by social and leisure purposes and, therefore, is influenced by the spatial characteristics of family and friendship relationships (Axhausen, 2008). The use of micro-simulation to predict traffic flows in time and space has a long tradition in transport planning and large-scale spatial models have emerged from this work (e.g., Arentze and Timmermans, 2004; Meister et al., 2005). Only recently, transportation researchers have acknowledged the importance of simulating the social network as part of the model-system to make further progress (Hackney and Axhausen, 2006; Carrasco et al., 2008; Barrett et al., 2009; Dugundji et al., 2011). Hence, prediction of social activities and related travel is another area of social-network models that is gaining interest.

An influential paradigm in the modeling of human choice behavior is based on the concept of random utility maximization (RUM). Since the seventies when McFadden (1974) derived the discrete choice model and corresponding likelihood estimation method, the RUM model is the standard model in consumer and travel behavior research (Ben-Akiva and Lerman, 1985; McFadden, 1986; Train, 2009). The assumption is that choice behavior is governed by an objective to maximize utility within constraints of available resources (i.e., time and money budgets) and cognition (i.e., limited information and mental effort). To acknowledge the fact that the modeler/analyst has only limited information about attributes of choice alternatives and decision maker, the utility function includes a random component (error term) so that choice behavior, from

* Corresponding author. Tel.: +31 40 2472283.
E-mail addresses: t.aarentze@bwh.tue.nl, t.aarentze@tue.nl (T.A. Arentze), kowald@ivt.baue.ethz.ch (M. Kowald), axhausen@ivt.baue.ethz.ch (K.W. Axhausen).
0378-8733/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.socnet.2013.05.002
the point of view of the modeler/analyst, is stochastic. Thus, an appealing property of the model is that it provides an error theory which means that parameters can be estimated based on choice data. Closed-form formulation of the model leading to the family of logit models, furthermore, makes the approach suitable for application in large-scale micro-simulation. Probably, this combination of theoretical, statistical and computational appeal explains that the model has become the standard in agent-based systems in transport and spatial planning.

Integration of RUM behavior of agents in models of social network generation is limited to date. Many network generation models in the literature are focused on reproducing known global characteristics of social networks such as path length, clustering coefficient and degree distribution without the intention to model choice behavior of actors explicitly (Wasserman and Robins, 2005). The small-world algorithm (Watts and Strogatz, 1998) and preferential attachment algorithm (Barabasi and Albert, 1999) are key representatives of this approach. A statistical approach to modeling social networks is the exponential random graph (ERG) approach (for an overview see Robins et al., 2007; Lusher et al., 2012). 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. The primary purpose of ERG models is to reproduce known tendencies such as reciprocity, transitivity and homophily in social networks. Also in case of ERG models, the focus is not on modeling the choice behavior of actors but rather on more direct modeling of the social structures this behavior may produce.

More recently, so-called stochastic actor-based models of network dynamics have been introduced which do have the intention to model behavior of actors explicitly (for a review of this field see Snijders et al., 2010). In these models, actors make decisions based on an objective function reflecting individuals’ preferences for reciprocity, homophily and transitivity in the formation of social ties. These models include an error theory that allows estimation of parameters based on social-network data. However, they cannot be readily integrated in simulation systems that assume RUM behavior of actors, as objective functions are not formulated in that framework. Also, model formulations are non-spatial as attention to date has focused on confined communities (e.g., a classroom) where geographic distance does not play a role. Furthermore, empirical estimation of the models on longitudinal data is not straightforward and needs further development before the models are fully operational for application to large populations (Snijders et al., 2010). Utility-based approaches have received some attention outside the above-mentioned framework (Hummon, 2000; Doreian, 2006). Stochastic dynamic networks also received attention in the network formation game literature (Jackson, 2008; Goyal, 2007). However, models stemming from this work are focused on testing of specific social theories or equilibrium structures in small-scale networks; for prediction of large social networks the models have limited value.

The purpose of the present study is to develop a model of social network generation that is consistent with the RUM model of human choice behavior, is estimable based on social network data and considers large populations in geographic space. We build on earlier work (Arentze et al., 2012), where we introduced a RUM model of friendship selection and a method to generate population-wide social networks based on this model. It was shown that this method based on RUM behavior of agents can reproduce homophily, degree distributions and effects of geographical distance in social networks. However, the model ignores an important element of behavior that leads to transitivity in networks. The intended contribution of the present study is to extend this model such that generated networks also accurately represent clustering in social networks. Furthermore, we develop a new maximum-likelihood method to estimate parameters of the friendship-formation model based on data of personal social networks of a sample of individuals. We apply the model to a large new dataset of personal networks that was recently collected in Switzerland (Kowald and Axhausen, 2012). Our model is static: the aim of the model is to generate the social network of a population at a given moment in time. In that way our approach is complementary to dynamic actor-based approaches which assume an initial network as given and focus on the simulation of changes of a network over time.

The remainder of the paper is structured as follows. In the next section, we present the proposed model and likelihood estimation method. In the section that follows, we describe the results of an empirical application of the model conducted to illustrate the approach. We conclude the paper by summarizing the major conclusions and discussing remaining problems for future research.

2. Model

In this section we first briefly introduce the basic RUM model of friendship formation which we developed in earlier work (Arentze et al., 2012) (Section 2.1). We then introduce the proposed extension to incorporate transitivity in the model (Section 2.2) and discuss computational demands of the model (Section 2.3). Finally, we develop a likelihood method to estimate the model based on personal network data (Section 2.4).

2.1. Basic model

The model assumes that the probability of a friendship between two persons depends on an evaluation of utility the two individuals expect to gain from the relationship. It is stressed that the concept of utility does not imply that only rational considerations of costs and benefits play a role. Individuals may seek various benefits in a friendship relationship such as opportunities to exchange information or receive support, to engage in joint activities and to satisfy basic social needs of affection and accompaniment (Wellman and Wortley, 1990). According to the model, an expected utility regarding all these dimensions depends on three attributes of a relationship leading to the following utility function:

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

where \( U_{ij} \) is the utility of a tie between persons \( i \) and \( j \), and \( V_{ij}^Q, V_{ij}^D \) and \( V_{ij}^C \) are structural utility components related to homophily (Q), geographic distance (D) and presence of common friends (C), and \( \epsilon_{ij} \) is a stochastic utility component (i.e., error term). The error term takes into account that not all attributes are observed by the modeler and, consequently, that the behavior is stochastic from the modeler point of view. The structural components reflect well-known tendencies in friendship networks indicating that the likelihood of a tie increases with attribute similarity between persons (homophily), decreases with geographic distance and increases when persons have friends in common (transitivity). In line with a general finding that reciprocity is the norm in friendship relations (Ilkink and van Tilburg, 1999), 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 \( i \). For reasons of parsimony, the present model assumes that the utility of a tie is the same for the involved persons \( i \) and \( j \) (i.e., \( U_{ij} = U_{ji} \)). In reality perceptions of utility may differ. It is possible to refine the utility function to represent perception differences.
Apart from benefits, costs evaluations play a role in the probability that a friendship evolves between two persons (Hummon, 2000; Doreian, 2006). Maintenance of a friendship requires time (and possibly expenditure) and, hence, competes with other activities for limited time (and possibly limited income). The opportunity costs may differ between the persons involved and are represented in the model as a threshold utility value (Arentze and Timmermans, 2009). The model assumes that a social tie evolves from encounters between two persons 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 can be defined as:

$$P(i \leftrightarrow j) = Pr(U_{ij} > \max\{u_i^0, u_j^0\})$$  

(2)

where, as before, $U_{ij}$ is the utility of the tie (defined by Eq. (1)) and $u_i^0$ and $u_j^0$ are threshold values for person $i$ and person $j$, respectively.

It is important to note that the threshold does not only represent opportunity costs. The threshold also captures opportunities a person has of meeting other people in settings where social contacts may result (Mollenhorst et al., 2008). Keeping all else equal, the more opportunities the environment and activities of a person provide the higher the threshold of the person needs to be in order to keep the number of friends in line with the person’s time-budget constraints and needs of social contact. Vice versa, the smaller the pool of possible friends given the reach of mobility tools of the individual, the lower the threshold should be for the person.

Eqs. (1) and (2) define the probability that a friendship evolves between persons given that they meet each other. In the simulation system, a complementary component of the model is concerned with creating (artificial) meetings between individuals (agents). Although behavioral models could be considered, the current model assumes a procedure where every agent encounters every other agent once. In every encounter a decision is made whether a friendship is formed based on the model given by Eqs. (1) and (2). In a population of $N$ agents, there are $N \cdot (N - 1)/2$ encounters and a corresponding number of decisions. The result of this process is a population-wide social network that shows tendencies of homophily, spatial proximity, transitivity and density depending on relative values of the utility components $V_i^D$, $V_j^D$, $V_{ij}^S$ and $u_{ij}^0$. We emphasize that encounters in the model are artificial; they do not represent actual meetings but rather possible meetings. This means that in effect the probability defined by Eq. (2) is related to the joint events of meeting and becoming 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 a spatially confined community – students on a university campus –, where geographic distance between the actors does not play a role.

2.2. Extension to model transitivity

Given the geographic locations and attributes of agents, the model described above can be used to generate a social network. A complication is, however, that the common-friend factor $V_{ij}^S$ introduces a problem of endogeneity. In order to generate a network, the presence of common friends needs to be known, which can be known only after a network has been created. Arentze et al. (2012) acknowledge that a linear process of checking for friendships therefore should be replaced by an iterative procedure, but they did not develop this. Since a mechanism of transitivity is lacking, networks generated by this basic model do not display the degree of clustering that we see in real networks.

The concept of transitivity is well-defined in the literature as reflected in methods to measure clustering in social networks. A commonly used method to measure the degree of clustering (transitivity) is based on the following index:

$$C = \frac{3 \times \text{number of triangles}}{\text{number of triples}}$$

(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. Fig. 1 shows an arbitrary example. In this network, subgraph 4-6-5 is an example of an open triple and subgraph 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 $C = 9/20$. 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.

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

In theory, the presence of a common friend has an influence on the utility the persons derive from a friendship as well as the probability that they meet and have an opportunity to develop a friendship. In terms of utility, a common friend may create benefits due to the opportunity it creates for persons to engage in joint activities involving three (or more) friends which may add value to a friendship. Apart from joint activities, having common friends may contribute to positive feelings of similarity and shared experience. As for meeting probability, a common friend can ensure that persons meet in joint activities (e.g., going to a concert) or events such as a birthday party.

In terms of the model, these effects can be represented respectively through the utility component ($V_{ij}^S$) and threshold component ($u_{ij}^0$). Practically, however, since the two effects do not relate to specific attributes the modeler can observe, it is difficult to disentangle them in a likelihood estimation and, therefore, we propose to model the two effects combined in the threshold function, as follows:

$$u_{ij} = \begin{cases} u_{ij}^0 & \text{if } C_{ij} = 0 \\ u_{ij}^0 - \theta & \text{otherwise} \end{cases}$$

(4)

where $u_{ij}$ is the threshold for individual $i$ regarding $j$, $u_{ij}^0$ 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 joint effect of increase in utility of a friendship and opportunity to meet caused by a common friend.

Having extended the threshold function, the issue of endogeneity requires attention in the friendship formation simulation process. We suggest the following extended process to deal with endogeneity:

Step 1 Create a primary tie between each pair of individuals where the utility of a tie exceeds the higher threshold level of both persons \((\theta > 0)\).

Step 2 Create a secondary tie between each pair of individuals who are not yet connected to each other and who have at least one common friend through primary ties and where the utility exceeds the lower threshold level of both persons \((\theta > 0)\).

In this procedure, each pair of individuals is evaluated in two stages of tie formation. In a first round (Step 1), friendship formation without support of common friends is considered. Ties formed in this phase are called primary ties. In the second round (Step 2), friendships are formed with support of common friends. Friendships that develop in this second phase are called secondary ties. Formally, the difference between the rounds is that for primary ties the higher value of the threshold is applied \((u^{ij}_0)\) and for secondary ties the lower value \((u^{ij}_0 - \theta)\). It is worth noting that the criterion for lowering the threshold of secondary ties is that at least one common friend exists; the exact number of common friends is not considered relevant for secondary-tie decisions.

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

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

This third step may lead to formation 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 here in the modeling of a time trajectory. Rather the purpose of the model is to generate a network for a given moment in time that represents an initial state of a population in a micro-simulation. For that purpose the two-step model suffices and is preferred as it is robust and saves computation time.

Assuming that the error term, \(\epsilon_{ij}\), in Eq. (1) is Gumbel distributed, a binary logit model defines the probabilities of primary and secondary ties. For primary ties the probabilities are given by McFadden (1988):

\[
p_{ij}^{(1)} = \frac{\exp(V_{ij})}{\exp(u^{ij}_0) + \exp(V_{ij})}
\]

where \(V_{ij} = v^{ij}_y + v^{ij}_y (V_{ij}^C = 0)\) is the structural utility component and \(u^{ij}_0\) is the maximum threshold across persons \(i\) and \(j\) where the threshold reduction parameter is zero \((\theta = 0)\). Given the probabilities, Monte-Carlo simulation is used to make the decisions. For secondary ties the probabilities are conditional on the knowledge that no primary tie exists between the pairs of persons considered. The unconditional probability is given by the ordinary binary logit model:

\[
p_{ij}^{(2)} = \frac{\exp(V_{ij})}{\exp(u^{ij}_0) + \exp(V_{ij})}
\]

where \(u^{ij}_0\) is the maximum threshold across persons \(i\) and \(j\) after threshold reduction \((\theta > 0)\). The conditional probability given that the utility does not exceed the high threshold, \(u^{ij}_0\), is defined as:

\[
p_{ij}^{(2)}(C_{ij} = 0) = \frac{p_{ij}^{(1)} - p_{ij}^{(2)}}{1 - p_{ij}^{(2)}}
\]

The probability defined in the denominator relates to the combined condition that the utility is smaller than the high threshold (no primary tie) and larger than the low threshold (a secondary tie). The numerator represents the total probability that no primary tie exists. Monte-Carlo simulation is used to generate decisions for formation of secondary ties.

2.3. Computational demands

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. The total number of times a pair of agents needs to be evaluated is given by:

\[
l = l_1 + l_2
\]

where \(l_1\) is the number of pairs to be evaluated for primary ties (Step 1) and \(l_2\) is the number of pairs to be evaluated for secondary ties (Step 2). The second set of pairs is a relatively small fraction of the total number of pairs. If \(n\) is the average degree (number of ties per person) and \(N\) is the size of the population, then this set is roughly of the order of magnitude of \(l_2 \approx N \cdot n \cdot (n - 1)/2\) (since agents on average have around \(n \cdot (n - 1)/2\) pairs of friends). On the other hand, the number of pairs to be evaluated for primary ties equals the total number of possible pairs in the population, which is \(l_1 = N \cdot N = N^2\). Fortunately, for large \(N\) an exhaustive enumeration of pairs is not needed. Note that the size of the threshold parameter \((u^{ij}_0)\) represents, among others, the opportunities agents have to form ties with others. If the number of arranged meetings for an agent is decreased, the threshold of the agent can be decreased to account for the reduced opportunities and keep the expected number of ties the same. Thus, if each agent meets only a fraction of all other agents and thresholds of agents are properly adjusted, then the statistical properties of networks of agents should stay the same. Obviously, the size of the fraction cannot be reduced without limit. In order to achieve correct statistical properties of the resulting network, each agent in the simulation has to be able to meet other agents that in terms of distance and socio-economic attributes are representative for the population. It is not known a priori what this minimum sample size is. One needs to test empirically whether the sample used was big enough by comparing the statistical properties of a simulated network to known properties.

There are several ways in which sampling can be conducted. A possible method is systematic sampling based on the following encounter-enumeration method for each agent \(i\):

\[
j = i + y, i + 2y, i + 3y, \ldots, N
\]

where \(y\) is some chosen step size. Complete encounter-enumeration is given by the case where step size is \(y = 1\), i.e., where every agent \(i\) meets every next agent \(j\) in line. On the other hand, if step size \(y = 2\), every \(i\) meets every second next agent in the population, reducing the total number of evaluations by half, and so on. In general, in this sampling method, every \(i\) meets every \(y\)th next agent so that for a given value of \(y\) the number of evaluations equals:

\[
l_1 = \frac{N \cdot (N - 1)}{2y}
\]

The larger the population is, the larger the value one can choose for step size \(y\) without affecting the result (in terms of relevant
2.4. Likelihood estimation

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. In reality, correlations between error terms of observations of a same ego (an individual in the sample) may exist due to unobserved attributes of an ego. An error-component form of the utility function (Eq. (11)) could be formulated and estimated in a mixed logit framework (Train, 2009). At present, however, we maintain the independence assumption in order to derive a basic framework for estimation and leave the extension for future research. A complication we should deal with is that existing datasets do not differentiate between primary and secondary ties. 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 personal networks can be defined as:

$$L_i = \prod_{j \in J_i} p_{ij} \left( \prod_{j' \in J_i^-} (1 - p_{ij}) \right)$$  

(14)

where $J_i$ is the observed set of friends of $i$, $J_i^-$ is the complementary set of all other individuals in the population and $p_{ij}$ is the likelihood of a tie between persons $i$ and $j$ predicted by the model. The first term on the right-hand-side of the equation represents the joint probability of all positive decisions and the second term the joint probability of all negative decisions concerning ties of person $i$ with 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 (McFadden, 1978). For the present case, a sampling method can be defined as follows:

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

(15)

where $S$ is a random sample of persons from the (full) population and $h$ is the sampling 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 for other persons with probability $(1 - p_{is})$. Provided that the sample is large enough, the approximation of the likelihood should be sufficiently accurate.

Eq. (8) can be used to predict probabilities $p_{ij}$ and $p_{is}$ in Eq. (15) 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, $\theta$. Since part of the ties will be secondary, the threshold will be underestimated. Furthermore, in this way, we do not obtain an estimation of $\theta$. However, these drawbacks are less problematic than they 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 do allow us to estimate utility function parameters, but do not allow us to estimate the threshold parameters $\theta$ 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, $\theta$, 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 calibration, whereby the preliminary estimate of the basic threshold parameter provides a good starting point.

3. Illustration

In this section, we describe an empirical application to illustrate the model. The application concerns the generation of a friendship network for the population of Switzerland. We first describe the datasets used in this application and next the results of estimation and application of the model to generate the network.

3.1. 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. The sample used includes a total of 265 individuals 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 and Axhausen (2012).

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 and 65+ years. We measure distances between homes as straight line distances based on the X-and-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. In 11.5% of the persons (ego or alter) there is a missing age value, in 6.2% a missing gender value and in 19.8% the exact location could not be identified. These missing values shares are typically found in personal-network surveys of this type (Kowald et al., 2012). Missing values are imputed where possible by drawing from distributions of ties conditional upon available information of the ego or alter for

<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 ties</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>11,168</td>
<td>0.463</td>
<td>0.617</td>
<td>26.6</td>
<td>72.7</td>
<td></td>
</tr>
</tbody>
</table>
the attribute under concern as observed in the total sample. In this way, distance and homophily relationships are preserved and not affected by the imputation.

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 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.

As it appears, there is a clear tendency of homophily in terms of both age and gender: 46% of the ties involve persons from the same age class and 62% 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 27 km with a standard deviation of 73 km. Average distances are slightly larger than average for alter-alter links. Furthermore, the sociogram data allows us to calculate the cluster index as defined in 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.

The population of Switzerland was considered here as the target population given the fact that the sample (mainly) originates from this area. To generate a network after having estimated the friendship model requires a (synthetic) population of agents representing the true population. 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 for the illustration, 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, increasing the sample (e.g., doubling the size) should not have any effect in the later application on characteristics such as transitivity, degree distribution and so on, provided that the threshold for primarily ties is properly adjusted by re-calibration (the larger the population the higher the threshold needs to be). In the model application below (Section 3.3) we will vary the sample size to illustrate this property of the model.

3.2. Estimation results

Estimation of the model is based on observations of ego-alter ties. As explained in Section 2.4, an independent random sample (5 in Eq. (15)) of the target population is needed to provide sociodemographic data regarding negative decisions (no tie exists). For this purpose a random sample of 100 persons was drawn for each observed ego from the (synthetic) population of 3301 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. 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 Iltenberger et al., 2011). Based on gender and age-class, similarity measures were defined in a straight forward 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.

3.3. Generating a population-wide social network

The model was used to generate a social network for the sample population (3301 persons). The thresholds are calibrated such that the mean personal network size and transitivity index match the observed values (22.0 persons and 20.6% respectively). The following iterative procedure was implemented to calibrate the two parameters:

1. Initialize the thresholds as \( u_0^t = u_0^s \) and \( \theta_1 = \theta_0 \), and set \( t = 1 \).
2. Run the network generation model using threshold values \( u_0^t \) and \( \theta_1 \).
3. Compute \( \bar{x}_t \) and \( c_t \) of the resulting network.
4. If \( x_t < \bar{x} - d_1 \) then set \( u_{t+1}^0 = u_0^s - s_1 \), otherwise if \( x_t > \bar{x} + d_1 \) then set \( u_{t+1}^0 = u_0^s + s_1 \).
5. If \( c_t < c^* - d_2 \) then set \( c_{t+1} = c_t - s_2 \); otherwise if \( c_t > c^* + d_2 \) then set \( c_{t+1} = c_t + s_2 \).
6. If \( x_t \neq \bar{x}_t \) or \( c_t \neq c^* \), then set \( t := t + 1 \) and repeat from 2.

where \( x^* \) and \( c^* \) are the target values of average personal network size and transitivity index respectively, \( t \) is an index of iteration, \( \bar{x}_t \) and \( c_t \) are current values of the two network characteristics, \( u_0^t \) and \( \theta_t \) are current settings of the threshold and threshold reduction parameter, \( u_0^s \) and \( \theta_0 \) are initial values, \( d_1 \) and \( d_2 \) are error-tolerance parameters, and \( s_1 \) and \( s_2 \) are step-size parameters for changing current parameter settings. This procedure makes small adjustments to the two threshold values depending on a comparison of the computed and target value of each of the two associated network characteristics. The process stops when the computed values fall within error margins of the target values (\( x^* - d_1 < x_t < \bar{x} + d_1 \) and \( c^* - d_2 < c_t < c^* + d_2 \)). The smaller the error margins are set the more accurate the calibrated values and the larger the expected number of iterations will be.

The error-margin parameters of the procedure were set as \( d_1 = 0.5 \) and \( d_2 = 0.02 \) and the step-size parameters as \( s_1 = 0.05 \) and \( s_2 = 0.005 \). The number of iterations depends on the accuracy of the settings of the initial values. Setting the initial value of the threshold to the estimated value from the likelihood estimation and the initial value of the reduction parameter to zero, the procedure typically converges in about 40 iterations. As result of this procedure, the calibrated values are \( u_0^* = 1.85 \) and \( u_0^* - \theta = -0.275 \). Implementing this setting the computed values for the generated network are \( \bar{x} = 22.0 \) and \( c^* = 0.192 \), which approximates the target values of \( x = 22.0 \) and \( c^* = 0.206 \) closely.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Estimation results of a basic model.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Null model</td>
</tr>
<tr>
<td>Log distance</td>
<td>0</td>
</tr>
<tr>
<td>Gender: same</td>
<td>0</td>
</tr>
<tr>
<td>Age: same class</td>
<td>0</td>
</tr>
<tr>
<td>Age: 2 classes difference</td>
<td>0</td>
</tr>
<tr>
<td>Age: 3 classes difference</td>
<td>0</td>
</tr>
<tr>
<td>Age: 4 classes difference</td>
<td>0</td>
</tr>
<tr>
<td>Threshold</td>
<td>4</td>
</tr>
<tr>
<td>LL</td>
<td>−37,540</td>
</tr>
<tr>
<td>Rho-square</td>
<td>0.378</td>
</tr>
</tbody>
</table>

T.A. Arenz et al. / Social Networks 35 (2013) 451–459
Tables 3 and 4 show all statistics of interest of the generated social network with these calibrated threshold parameters. The tables show statistics for primary and secondary ties separately as well as for all ties together. The last row shows the statistics for observed personal-networks of the sample for comparison. Since primary and secondary ties cannot be distinguished in the observed case, these latter figures refer to ego-alter ties in general. 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 at large due to sampling error. Therefore, we expect differences in this comparison even if the model is accurate.

Table 3 shows statistics related to homophily and distance relationships. The model reproduces observed age and gender similarity fairly accurately. 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 ($t$-value of difference in means is $-9.83$, $p < 0.001$). 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). Furthermore, contributing to this difference is the restriction inherent in the assumed synthetic population that predicted ties are confined to fall within the borders of Switzerland, whereas reported ties are not. Visual comparison of the graphs in Fig. 2 shows that the model fairly accurately reproduces the distribution of ties across age-class difference categories.

Fig. 3, which shows the distribution of ties across distance categories, confirms the earlier observation that the model underpredicts the largest distance category. Apart from that, visual comparison of the graphs indicates that the distance distribution reasonably closely matches the observed one.

Although predicted and observed tie distributions match each other fairly accurately on the dimensions considered, Chi-square tests indicate that the differences are statistically significant. This should, however, not surprise us. Due to the large sample size (41,456 predicted ties and 11,168 observed ties) small differences in the prediction become significant very quickly in a statistical sense.

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 (0.035), as expected. So, according to the model, the common-friend mechanism is predominantly responsible for the level of transitivity that networks display. Fig. 4 shows graphically the predicted and observed frequency distribution of persons across size categories of the personal network. The difference between the distributions is statistically significant (Chi-square value is $\chi^2 = 18.8$, $df = 7$, $p = 0.0088$). The standard deviation

---

1 The Chi-square values for difference between predicted and observed distributions are, for age-class difference, $\chi^2 = 99.8$ ($df = 5$, $p < 0.001$), for gender difference, $\chi^2 = 25.8$ ($df = 1$, $p < 0.001$) and, for distance category, $\chi^2 = 726$ ($df = 7$, $p < 0.001$).
of network sizes is somewhat overpredicted by the model (17.9 versus 13.4 persons). 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 only small.

The above analysis assumed a sample size of 10% of the population. As said, the statistical properties of a simulated network should not change when a different sample size is used provided that the thresholds are properly adjusted. To illustrate this property of the model, Table 5 shows the results of different model runs where the sample size was varied. Four sample sizes are considered, namely 7.5%, 10%, 12.5% and 15% of the population of 33,390 persons. The 10% sample corresponds to the case considered above. For comparison the first row of the table shows the statistics of the observed network. For each sample size the thresholds were recalculated using the iterative procedure. As expected, the calibrated threshold value for primary ties increases monotonically with sample size: the value increases from 1.25 to 2.40 when the sample size increases from 7.5 to 15%. The calibrated threshold for secondary ties (\(u^s - \theta\)) does not show a clear trend, but varies around a value of \(-0.23\), again, as expected. As a result of the calibration, the mean personal network size and transitivity index of the networks are accurately reproduced. The other network characteristics are also robust for sample size. That is to say, the homophily characteristics, as indicated by the same-age and same-gender ratios, are accurate and are not affected by the choice of sample size. On the other hand, the (geographic) distance distribution of ties does show an influence of sample size: with increasing sample size the average distance and standard deviation both show a tendency to decrease with sample size.

Table 5 Simulated network characteristics for different sample sizes.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Thresholds</th>
<th>Personal network size</th>
<th>Transitivity</th>
<th>Same age</th>
<th>Same gender</th>
<th>Distance (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(u^s)</td>
<td>(u^s - \theta)</td>
<td>Mean</td>
<td>SD</td>
<td>Ratio</td>
<td>Ratio</td>
</tr>
<tr>
<td>Observed</td>
<td></td>
<td></td>
<td>21.9</td>
<td>13.4</td>
<td>0.206</td>
<td>0.46</td>
</tr>
<tr>
<td>7.5%</td>
<td>1.25</td>
<td>-0.260</td>
<td>21.7</td>
<td>20.7</td>
<td>0.192</td>
<td>0.42</td>
</tr>
<tr>
<td>10%</td>
<td>1.85</td>
<td>-0.275</td>
<td>22.0</td>
<td>17.9</td>
<td>0.190</td>
<td>0.42</td>
</tr>
<tr>
<td>12.5%</td>
<td>2.25</td>
<td>-0.200</td>
<td>22.1</td>
<td>20.7</td>
<td>0.208</td>
<td>0.42</td>
</tr>
<tr>
<td>15%</td>
<td>2.40</td>
<td>-0.205</td>
<td>22.0</td>
<td>22.7</td>
<td>0.202</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Figure 5. Distance distribution of ties in simulated networks based on different sample sizes.

4. Conclusions and discussion

We introduced an agent-based model to generate social networks for populations explicitly in geographic space. The core of the model is a friendship-tie formation model which is formulated in the random-utility-maximization (RUM) framework. We showed how the model can be estimated by likelihood estimation based on observations of personal social networks. An application to a synthetic population of Switzerland indicated that the model is able to reproduce several characteristics of social networks that are generally important in spatial modeling of human behavior, including geographic distance, attribute similarity (homophily), size of personal networks (degree distribution) and clustering of networks (transitivity). The model is scalable to large populations. Not only are computational demands of the model modest. The model also supports a sampling procedure where only a fraction of a population instead of the full population is simulated. The chosen sampling fraction can be taken into account in the generation of a network by adjustment of a single threshold parameter. Mentioned properties (spatial proximity, degree distribution, homophily and clustering) of a generated network are maintained given that the threshold parameter is adapted to the size of the fraction of the population represented in the simulation.

The model we presented contributes to the emerging field of stochastic actor-based models. One such field is micro-simulation of individuals’ daily activity-travel behavior for transport and spatial planning. Individuals’ social-leisure activities, which account for an important segment of travel in modern societies, are to an important extent influenced by spatial characteristics of the social network. The friendship-tie formation model we proposed fits seamlessly in existing micro-simulation systems.
which assume RUM behavior of agents and, therefore, can be used to further enhance the usefulness of these micro-simulation systems as well as the integration of RUM and theories of social selection.

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 mechanisms of meetings instead such as emphasized in theories of social opportunities (Mollenhorst et al., 2008). An aspect of this is related to geographic space. Factors such as urban density, which vary spatially, may influence opportunities of meetings (Neuens et al., 2008; Butter et al., 2012). In terms of the model this means that geographic place factors should be incorporated in threshold functions. Another aspect of this concerns the mutual influence 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 and life-cycle dynamics such as proposed in life-cycle theories of social activities (Kalmijn, 2003) deserve attention. Although the present model can account for correlations between personal networks through the mechanism of common friends, it ignores the special role ties between household members may play in this regard.

References

Cambridge.