Modeling spatial embedded social networks

Working Paper

Author(s):
Frei, Andreas; Axhausen, Kay W.

Publication date:
2011-03

Permanent link:
https://doi.org/10.3929/ethz-a-006562051

Originally published in:
Arbeitsberichte Verkehrs- und Raumplanung 685

Modeling spatially embedded social networks

A. Frei and K.W. Axhausen

Working Paper Transport and Spatial Planning 685 March 2011

Institut für Verkehrsplanung und Transportsysteme
Institute for Transport Planning and Systems

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Working Paper

Modeling spatially embedded social networks

Andreas Frei
Institute for Transport Planning and Systems,
Swiss Federal Institute of Technology Zurich
Wolfgang-Pauli-Str. 15
8093 Zurich
Switzerland
Tel: +1 281 889 6821
Email: andi.frei@gmail.com

Abstract

In this paper different models for social networks are reviewed. Two models which allow for spatial embedding social networks in physical space are chosen and applied. Both of the models are based on spatial interaction, where the probability for a tie between people decreases with increasing distance. The models are applied based on empirical data of egocentric social networks collected in Zurich, Switzerland. The analysis of the empirical data shows, that there is a strong exponential decrease of probability to maintain a social tie. Based on the empirical distribution of the social ties, the models are used to simulate spatially embedded networks to reproduce the tie length distribution in the empirical data.

Keywords

Social network, distance, geography, spatial

Preferred citation style

A. Frei and K.W. Axhausen (2011) Modeling spatial embedded social networks, *Working paper Transport and Spatial Planning,* **685**, IVT, ETH Zurich, Zurich.

1 Introduction

Many problems in transportation arise from peak hour traffic on daily commutes. It is therefore not surprising that most work in transportation planning concentrated on these problems. However, there are also increasing problems which arise from leisure travel. For example the peaks in traffic during beginning and ends of holidays, or peaks on weekends for going to skiing trips. In fact, leisure travel dominates the travel market in terms of the miles traveled and journeys undertaken. In industrialized countries, leisure travel shares of 33% of daily travel and 40% of daily mileage are common, see (Bundesamt für Raumentwicklung und Bundesamt für Statistik, 2007; Bureau of Transportation and Statistics, 1995). The reason why leisure travel related research is rather limited compared to research of commuting related travel is partly also due to its complexity. The dominant paradigm for transport planning work is the individual satisfying his or her needs while maximizing the utility derived from the activities undertaken. To analyze and model trip making, transport researchers use measurable indicators such as socio-demographic attributes of the population, the spatial distribution of activity opportunities, and the generalized travel costs by mode and infrastructure de Dios Ortúzar and Willumsen (2001). In most situations, data of this type is available from local administrations and can be used for modeling commuter traffic. But for leisure traffic it has become clear that the location choice for executing these activities cannot be understood with just the attributes of the location, the available infrastructures to access the locations and the traveler's characteristics without understanding how social networks of relatives and friends are distributed. As leisure travel is influenced by individual lifestyles, values and essentially is driven by social motivations, i.e., to visit friends or relatives or to join them in activities (Larsen *et al.*, 2006). The 1995 American Travel Survey classified 33% (Bureau of Transportation and Statistics, 1995) of the reported leisure journeys as visiting friends and others, but a further 33% involved purposes which people rarely or never engage in alone, such as outdoor sports (e.g., golf, sailing) or family events (e.g., weddings, funerals), so that the true share of journeys which are fully or partly motivated by the wish to see families and friends is even higher. Data of such type is very rare since on the one hand the realization of appropriate surveys is often very costly and extensive, and on the other hand respondents are required to disclose private data (Axhausen, 2008). A common way to model the social relation between individuals is to use graph theory. Research on social networks started in the 1960ies (Wasserman and Faust, 1994) and has gained increasing interest from sociologists and also physicists in the last years. Although there is a huge number of studies about social networks, their use for transport planning is rather limited. This has pragmatic reasons, as empirical social networks are very hard to explore in unlocalized environments, they are often embedded in a closed environment, e.g., people working at the same company, pupils of the same school or same class, movie actors or collaborating scientists. A Social network which is representative in terms of travel behavior is not localized and therefore samples of such networks are very rare. In addition, the majority of social network studies provide no information about spatial dimensions of the social network. However, research in social network analysis started only recently to investigate the spatial dimension of networks. The motivation for the work presented in this paper is to develop a methodology to generate data that describes the social relations required for the leisure travel modeling process. Statistical models are presented to generate a spatial embedded social network with physical distance as the explanatory variable.

2 Background and empirical data

The driving force behind social network structure is the homophily, which has been shown in many empirical studies (see McPherson *et al.*, 2001). For example a big part of friendships' origin is work or family, the so called "inbreed homophilie". Born into a family, we share automatically many characteristics, such as race and religion. With our friends from work we often share characteristics such as social class, education. All those characteristics are not exclusive characteristics, which means that they do not limit our active choice. Many models of social networks are based on these characteristics (e.g., van Dujin *et al.*, 2004; Newman, 2003). But much more important than the "inbreed homophily" is the homophily which actually limits directly the pool of potential ties by geographic space. Many studies show a strong relationship between physical distance and social structure (e.g., Merton, 1948; Festinger *et al.*, 1950; Caplow and Forman, 1950; Whyte, 1956; Sommer, 1969; Snow *et al.*, 1981; Latané *et al.*, 1995; Wellman, 1996; Wellman *et al.*, 2001).

Previous research on the detailed spatial distribution of social networks is sparse. Empirical analysis on the structure of complete social networks outside of institutional settings, such as workplaces, schools or clubs, is rare. The information obtained about the locations of the members in localized "complete" or ego-centric networks is spatially rough, if available at all. Geographers and sociologists have generally ignored this issue. Transport planners have therefore recently begun to undertake new survey work satisfying their needs, while drawing on the extensive sociological experience in the capturing of ego-centric networks.

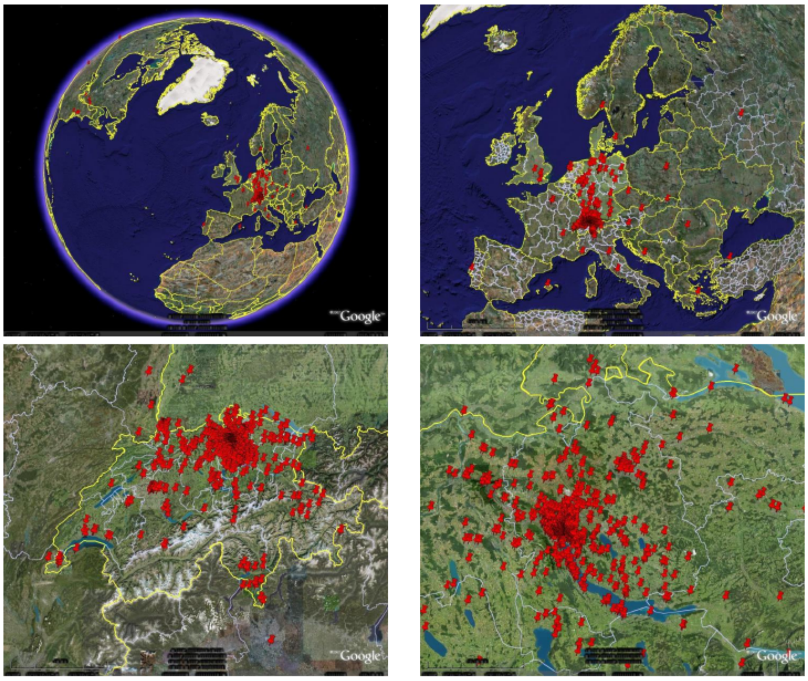
This paper uses the – to our knowledge - largest of these new quantitative surveys. Its closest, but substantially smaller, comparison was undertaken independently at about the same time in Toronto by a team involving sociologists and transport planners (Carrasco *et al.*, 2007). The scope of our survey was derived from initial qualitative work

(Ohnmacht and Axhausen, 2005; Larsen *et al.*, 2006) and related quantitative work on mobility biographies (Beige and Axhausen, 2006). The survey instrument addresses:

- The basic socio-demographics of the respondent today.
- The mobility biography of residential and employment moves over the lifetime, including information about income levels, mobility tool ownership, main mode of transport to work.
- The ego-centric network through a name-generator with four prompts: With whom the respondents "discuss important problems, with whom you stay in regular contact or which you can ask for help" and further persons, not yet given, with whom the respondents spend leisure time. A name interpreter establishes the exact home location of the contacts (alteri) of the respondents and the frequency of their interactions by four modes: face-to-face, phone, email and texting (short-message-service – SMS).

In total the data consists of 307 randomly sampled respondents from Zurich, which reported in total 3,791 acquaintances. Figure 1 shows the spatial distribution of respondents and acquaintances. The survey was not restricted to Switzerland, so respondents are allowed to name leisure contacts outside of Switzerland. Nevertheless, the physical space that is used for the network generation process in this paper is bounded. So, in the following analysis links with a distance greater than 120 km are ignored, as the physical space for the generation of links in the following experiments is chosen to be squared, with a diagonal of 120 km.

Figure 1: Residential locations of the respondents and acquaintances from the survey in Zurich



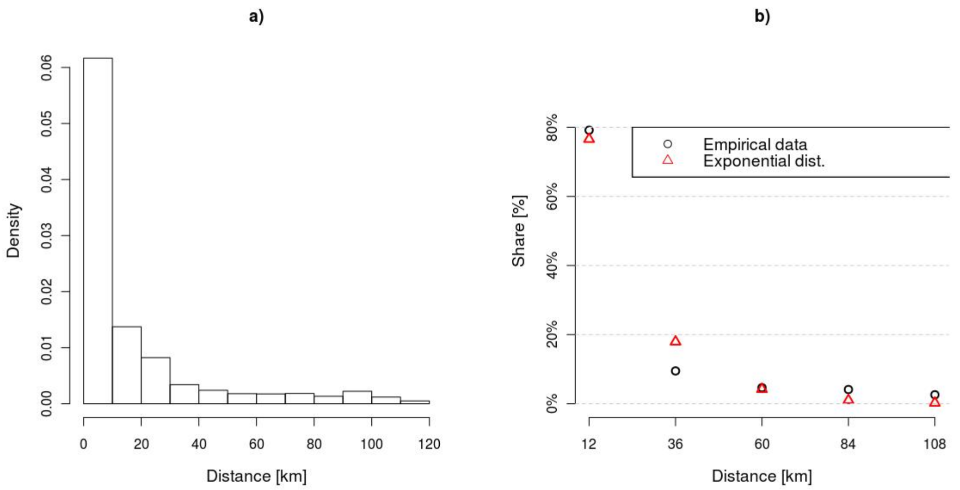
3 Analysis of empirical data and model discussion

3.1 Analysis of empirical data

The collected data does not include information about socio-demographics of the acquaintances. Therefore the social network model will rely on geographical distance only. However, the analysis and method used could be extended by adding further explanatory variables, such as age, gender etc. to include so called "inbreed homophily".

Figure 2 shows the edge length distribution of social contacts reported in the survey and a fitted exponential distribution. The share of contacts decreases with increasing distance. The distribution of the great circle distances between the respondents' residence and their relationships has two elements. Nearly two-thirds of the alteri live locally within 25 km. The bulk of the remaining distances are divided into regional and national relationships. The distance shows a strong exponential decay which is clearly visible in the histogram, but the distribution does not follow a simple parametric distribution, which seems reasonable because people are not equal distributed over space because of places which are not habitable e.g. mountains or the sea and because people tend to cluster in cities. To approximate the distance decay an exponential distribution seems to be appropriate, even tough the goodness of fit statistics rejects the null hypothesis (Chi-Square, Kolmogorov-Smirnov and Anderson-Darling Test) even at the 0.25-level. The Best-fit parameter for the exponential distribution is 0.060.

Figure 2: a) Histogram of the edge length distribution b) Empirical data of the edge length and fitted exponential distribution by quintiles



3.2 Model discussion

Research on models of spatially embedded networks is even more sparse than literature on the analysis of the spatial dimension of social contacts. **?** propose a model which is based on the idea that there is a certain probability that people become friends if they remain at the same place in an overlapping time interval. To extract the trajectories of peoples' mobility patterns, they use a microscopic traffic simulation, which simulates daily traffic in the greater area of Zurich, Switzerland. Hackney and Marchal find that distance distribution of social contacts is heavily right-skewed, but they provide no detailed analysis on the shape of the distribution.

Besides the model of Hackney and Marchal, there are several social network models, which are based on probability, but which are not especially used to model geographically embedded social networks. However, Toivonen *et al.* (2009) classifies these social network models in two main categories: those were the link updates depends on local network structures (network evolution models, NEMs) and those were the update of the links is dependent on the attributes of the vertices or nodes (nodal attribute models, NAMs). A third family of model, which can be interpreted as NEM or NAM, is the exponential random graph model (ERMG). The difference between NEMs and ERMGs are that the ERMGs do not include a link update step based on probabilities. The ERGMs link existence is defined by the probability distribution. Markov Chain Monte Carlo (MCMC) sampling methods, which are used in ERMGs, can also be used to model the evolution of social network, as proposed by Snijders (1996).

All of these models generate non bipartite, social networks, which are not directed.

Since the spatial dimension of a social network is defined as an attribute of each link and not as a network structure, a recent NAM (Boguna *et al.*, 2004) (BPDA) and an ERGM (Wong *et al.*, 2006) (ERGM) are studied following in more depth. Both approaches, the BPDA and ERGM will be tested to generate spatial social networks, to see, which approach fits the needs of this research better in terms of reproduction of spatial patterns obtained in a the Zurich survey and in general how they produce typical social network structures. These typical network structures, which are observed in social networks are:

- Large clustering $c_i = \dfrac{2e_i}{k_i(k_i-1)}$, where e_i represents the number of triangle connected to vertex c_i and k_i is c_i's degree. The denominator represents the possible triangles.
- Positive degree correlation, which means that the degree k_n is positively correlated with the average degree of the nearest neighbors of n: $cor\,[k_n, \overline{k_{nn}}(k)] > 0$.
- The existence of communities; In many cases it was shown that clustering does not occur evenly over the entire network: Subgroups of actors who are highly connected within themselves are loosely connected to other subgroups, which are themselves highly connected.

3.2.1 BPDA

? propose a class of models based on the concept of social distance, e.g. based on the idea, that friendship links get established whenever individuals feel close in some sense to each other. This leads to a notion of social distance, which rule the establishment of friendship in a way that individuals with short distance have a large probability of being linked, and vice versa for long distances. Their model has 3 free parameters N, α and β. The mechanism can be described as: Distribute N nodes with uniform probability in a (one-dimensional) social space (a segment of length h_{max}). Link nodes with probability$p = 1/(1+(d/\beta))$, where d represents the distance in the social space and h_{max} can be absorbed within β. If the model is treated many-dimensionally, similarity along one of the social dimensions is sufficient for the nodes to be seen as similar.

.

3.2.2 ERGM

Wong *et al.* (2006) propose a spatial random graph model. The model uses four free parameters, N, H, p and p_b. The model's mechanism can be described as: Distribute N nodes according to a homogeneous Poisson point process in a (two-dimensional) social space of unit area. Create a link between each node pair separated by distance d with probability $p+p_p$ if $d < H$, and with probability $p–p_\Delta$ if $d > H$ (where $p_\Delta(p,\, p_b,\, H)$ is such that the total fraction p of all possible links is generated). They model the probability of an edge formation as a simple step function that only differentiates between edges within and beyond a so-called neighborhood radius. As a simulation scenario they use randomly scattered vertices. Wong et al. study the effect of the neighborhood radius on small-world properties and community structures, but they make no statement about the edge length distribution.

4 Implementation and simulation

4.1 Basic assumptions

The simulated social networks represent always loop-less undirected social networks $G = (N; H)$ with a known vertex set N and a set of links H. G is spatially embedded, in the sense that a space S is defined and the set $N = \{n_1, ..., n_N\}$ such that $N \subset S$. The distance function, d, on S, is the great circle distance in kilometers. The models' specification include x and y coordinates for each Node n_i, which may cause boundary problems, which will not be analyzed in this paper. A less critical model could be generated by assigning the space in one dimension by just modeling the distance between each node, instead of assigning each node x and y coordinates. But for further usage of the model and its network structure implications on activity/location choice, the model calculates distances by assigning coordinates to each node in a physical space. For simplicity a population of size $N = 750$ is placed uniformly within a square region of size $l * l = 100km$. Such a model may be thought of as a first approximation to a very sparse population distribution in physical space over large areas. The models are configured to generate social networks in physical space, with characteristics to represent the Zurich data (Frei and Axhausen, 2007), with a mean degree of 12.36 and a edge length distribution flowing an exponential distribution $f(x;\lambda) = \lambda e^{-\lambda x}$, $x \geq 0$. *and* 0, $x < 0$ with $\lambda = 0.06$.

4.2 BPDA-Simulation

The BPDA model rules the establishment of friendship in a way that individuals with short distance have a large probability of being connected, and vice versa for long distances. Each Node i has a location h_i. Connection between each pair of individuals, $\vec{h}_i$ and $\vec{h}_j$ is given by: $r(\vec{h}_i, \vec{h}_j) = \sum_{n=1}^{N} \omega_n r_n(h_i^n, h_j^n)$, where ω_n is a normalized weight factor and r_n is the probability for a friendship given by: $r_n(h_i^n, h_j^n) = \frac{1}{1 + [\beta_n^{-1} d_n(h_i^n, h_j^n)]^{\alpha_n}}$, where β_n is a characteristic length scale (that controls eventually the average degree) and $\alpha_n > 1$ is a homophily measure, the tendency people connect to similar people. n is the number of possible links: $i * (i - 1)$.

To generate the social network two parameters alpha and beta have to be adjusted to reproduce the empirical Zurich data. For optimizing alpha and beta to reproduce the right-skewed distribution and the degree distribution, an incremental search is used to minimize the deviance of the generated and empirical link distance distribution within six bins, weighted by cases , as well as the mean degree of social contacts. The resulting, chosen social network model configuration is given by $\beta = 2.25$ and $\alpha = 1.99$. The resulting distance distribution are visualized in Figure 3. The target distribution can be very adequately reproduced.

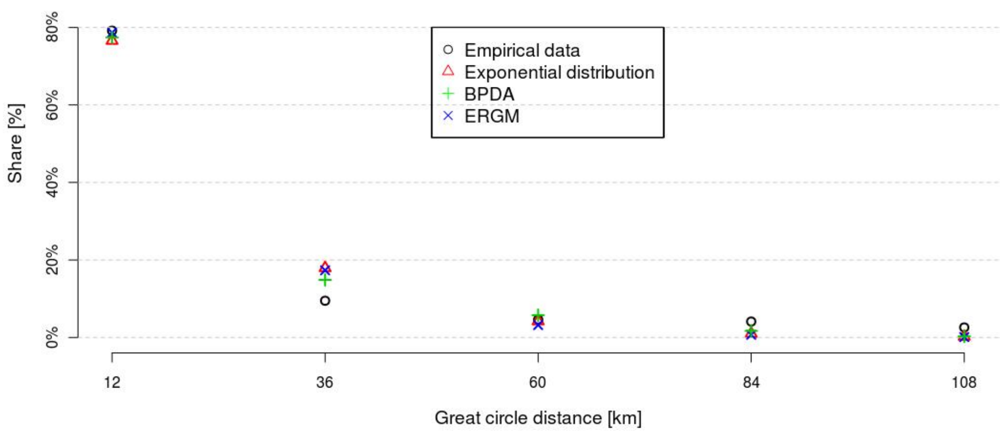
The simulated network meets also the the required criteria for a social network. The network has a large clustering coefficient of 0.42 with a low number of total ties (9450 of 561,750). The degree of correlation is positive and relative high (0.42).

4.3 ERGM-Simulation

The ERGM model describe the local selection forces that shape the global structure of a network. The probability of a tie is described as$P(Y = y \mid X) = exp[\theta^T g(y, X)]/\kappa(\theta)$, where$Y$ is random set of relations (edges and non-edges),y is a particular given set of relations, X is the matrix of attributed for the vertices, $g(y, X)$ the vector of network statistics,θ vector of coefficients and$\kappa(\theta)$ normalizing constant. The log-odds of any given edge given the current state of the rest of the network is described bye the logit function: $logit(Y_{ij} = 1) = \theta^T \delta[g(y, X)]_{ij}$, where Y_{ij} is an actor pair in Y, and $\delta[g(y, X)]_{ij}$ is the change in $g(y, X)$ when the value of $y_{ij}$is toggled from 0 to 1. This is in this simple case the fraction of possible edges which are realized. Since we use for this paper the edge distance as the local selection force, $g(y, X)$ contains the distance of the edges and the θ vector of coefficients is given by the parameters for the exponential distribution. To simulate a network, the Gibbs-sampling method is used, where from a constant initial set of edges G of all possible edges N are chosen, to generate the desired degree. Further is in each step one edge of H and one edge of $G - H$ is randomly drawn and if $\frac{p(G-H)_i}{p(G-H)_i + p(H)_j} \geq U(0, 1)$ the set of H is updated. As the probability to chose an edge with a certain distance is not uniform, the probability to chose a certain edge is therefore weighted bye the inverse kernel density estimation for each edge: $\frac{max(density(distance))}{density(distance)}$. Candidates of edges to include in N are 10^7 times sampled. Which is for the size of the Network enough long, as the distance distribution did no more change after about 10^6 samples. The resulting distance distribution is visualized in Figure 3. The target distribution can be very adequately reproduced, but deviates a little bit from the exponential distribution, which is targeted, due to the weighting procedure, which is not 100% accurate, as the weighting is done in bins of finite length of 0.5 km. Nevertheless, if the edge length distribution of the initial whole network were uniform, the simulated network would follow exactly the desired exponential distribution.

The simulated network meets also the the required criteria for a social network. The network has a large clustering coefficient of 0.38 with a low number of total ties (9446 of 561,750). The degree of correlation is positive and relative high (0.38).

Figure 3: Empirical data of the edge length distribution, exponential distribution and simulated distribution



5 Discussion and conclusion

This paper presents two stochastic models for spatially embedded social networks. Empirical data show that the edge probability does not follow a simple distribution, because of geographical reasons and the tendency of people to cluster in cities. However, an exponential distribution based on the edge length with a parameter of 0.060 can approximate the edge distance distribution. The paper shows that the observed exponential distribution can be explained with a relatively simple homophily model where the probability of an edge can be modeled as a function of the distance between two persons. With a simulation on generated vertices, we demonstrate that we can reproduce the observed edge length distribution with the two proposed models. The simulated network shows a right-skewed degree distribution and high degree- degree correlation, which are both typical properties for social networks. Where as both models perform good on reproducing the edge distance distribution, the ERGM approach has to be favored, as it allows to incorporate further explanatory variables.

In the context of transportation research, the great circle distance is probably not the appropriate measurement for the spatial effects between two persons. Especially regarding topology e.g. mountains etc. have a strong influence on the selection process. To include the real cost, which one has to be overcome, to maintain an edge, travel time should be used as a measurement instead of distance. However, distance could easily be replaced by travel time in both of the suggested models.

For a variety of reasons, it may be interesting to have a complete model of all social contacts of a region or a country, as it is suggested in this paper. Possible applications are the location choice of leisure traffic or visiting friends and relatives. Dynamic models such as the ones by Hackney (Hackney and Marchal, 2009) are able to generate similar results. But statistical measures such as distance, clustering coefficients etc. are very difficult to control. Statistical modeling to generate social networks can produce robust static results, which can then be used for further models, which represent short to mid time decisions, which do not affect the social network immediately.

References

Axhausen, K. W. (2008) Social networks, mobility biographies, and travel: Survey challenges, *Environment and Planning B: Planning and Design*, **35** (6), 981–996.

Beige, S. and K. W. Axhausen (2006) Residence locations and mobility tool ownership during the life course: Results from a retrospective survey in switzerland, in *paper presented at the European Transport Conference*, Strasbourg.

Boguna, M., R. Pastor-Satorras, A. Diaz-Guilera and A. Arenas (2004) Models of social networks based on social distance attachment, *Physical Review E*, **70** (5), 1–8.

Bundesamt für Raumentwicklung und Bundesamt für Statistik (2007) Mikrozensus zum verkehrsverhalten 2000, Technical report, Bundesamt für Statistik und Bundesamt für Raumentwicklung.

Bureau of Transportation and Statistics (1995) Nationwide personal transportation survey 1995, Technical report, Bureau of Transportation Statistics.

Caplow, T. and R. Forman (1950) Neighborhood interaction in a homogeneous community, *American Sociological Review*, **15**, 357–366.

Carrasco, J. A., B. Hogan, B. Wellman and E. J. Miller (2007) Collecting social network data to study social activity-travel behavior: An egocentered approach, *Environment and Planning B: Planning and Design*, **34** (6), 961–980.

de Dios Ortúzar, J. and L. G. Willumsen (2001) *Modelling Transport*, Wiley, Chichester, Chichester.

Festinger, L., S. Schachter and K. Back (1950) *Social Pressures in Informal Groups: A Study of Human Factors in Housing*, Harper, New York, NY, USA.

Frei, A. and K. W. Axhausen (2007) Size and structure of social network geographies, Arbeitsberichte Verkehrs- und Raumplanung 444, IVT, ETH Zurich, Zurich.

Larsen, J., J. Urry and K. W. Axhausen (2006) *Mobilities, Networks, Geographies*, Ashgate, Aldershot.

Latané, B., J. Liu, A. Nowak, M. Bonevento and L. Zheng (1995) Distance matters: Physical space and social impact, *Personality and Social Psychology Bulletin*, **21** (8), 795–805.

McPherson, M., L. Smith-Lovin and J. M. Cook (2001) Birds of a feather: Homophily in social networks, *Annual Review of Sociology*, **27** (1), 415–444.

Merton, R. K. (1948) The social psychology of housing, in *Current Trends in Social Psychology*, ed. W. Dennis, 163–217, University of Pittsburgh Press, Pittsburgh.

Newman, M. E. J. (2003) Mixing patterns in networks, *Physical Review E*, **67** (2), 26–126.

Ohnmacht, T. and K. W. Axhausen (2005) "wenn es billiger ist als die bahn - na ja, warum nicht?": Qualitative auswertung zu mobilitätsbiographien, mobilitätswerkzeugen und sozialen netzen, in *Arbeitsbereichte Verkehrs- und Raumplanung*, 327, IVT, ETHZ.

Snijders, T. A. (1996) Stochastic actor-oriented models for network change, *Journal of Mathematical Sociology*, **21** (1-2), 149–172.

Snow, D., P. Leahy and W. Schwab (1981) Social interaction in a heterogeneous apartment: An investigation of the effects of environment upon behavior, *Sociological Focus*, **14** (4), 303–319.

Sommer, R. (1969) *Personal Space: The Behavior Basis of Design*, Prentice-Hall, Englewood Cliffs.

Toivonen, R., L. Kovanen, M. Kivelä, J. Onnela, J. Saramäki and K. Kaski (2009) A comparative study of social network models: Network evolution models and nodal attribute models, *Social Networks*, **31** (4), 240–254.

van Dujin, M., T. Snijders and B. Zijlstra (2004) P: a random effects model with covariates for directed graphs, *Statistica Neerlandica*, **58** (2), 234–254.

Wasserman, S. and K. Faust (1994) *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge, MA.

Wellman, B. (1996) Are personal communities local? a dumptarian reconsideration, *Social networks*, **18** (4), 347–354.

Wellman, B., A. Haase, J. Witte and K. Hampton (2001) Does the internet increase, decrease, or supplement social capital, *American behavioral scientist*, **45** (3), 436–455.

Whyte, W. H. (1956) *The Organization Man*, Simon and Schuster, New York.

Wong, L. H., P. Pattison and G. Robins (2006) A spatial model for social networks, *Physica A: Statistical Mechanics and its Applications*, **360** (1), 99–120.