Noname manuscript No.
(will be inserted by the editor)

The role of spatial interaction in social networks Working Paper 11-11

Johannes Illenberger · Kai Nagel

the date of receipt and acceptance should be inserted later

Abstract This article addresses the role of spatial interaction in social networks. We analyse empirical data describing a network of leisure contacts and show that the probability to accept a person as a contact scales in distance with $\sim d^{-1.4}$. Moreover, the analysis reveals that the number of contacts an individual possesses is independent from its spatial location and the spatial distribution of opportunities. This means that individuals living in areas with a low accessibility to other persons (rural areas) exhibit at average the same number of contacts compared to individuals living in areas with high accessibility (urban areas). Low accessibility is thus compensated with a higher a priori probability to accept other candidates as social contacts.

In addition, we propose a model for large-scale social networks involving a spatial and social interaction between individuals. Simulation studies are conducted using a real-world synthetic population as input. The results show that the model is capable of reproducing the spatial structure, but, however, fails to reproduce other topological characteristics.

Both, the analysis of empirical data and the simulation results provide a further evidence that spatial interaction is a crucial aspect of social networks. Yet, it appears that spatial proximity does only explain the spatial structure of a network but has no significant impact on its topology.

Keywords social networks · spatial interaction · accessibility

Johannes Illenberger
Group of Transport System Planning and Transport Telematics
Berlin Institute of Technology
E-mail: illenberger@vsp.tu-berlin.de

Kai Nagel
Group of Transport System Planning and Transport Telematics
Berlin Institute of Technology
E-mail: nagel@vsp.tu-berlin.de

1 Introduction

Research on social networks has made great advances in understanding the structure and dynamics of networks in the last decade. An increasing availability of proxy-data sets from which social networks can be inferred (for instance movie actors [1] and co-authors [2]) allow for the insight into large-scale networks. While extensive research is conducted on the organisational and social structures of networks the focus has just recently shifted to another dimension: the spatial structure. Among sociologists the relationship between physical space and social structure has already been identified in the middle of the 20th century [3]. However, detailed analysis on the spatial structure of social networks has just begun with the spatial analysis of networks in general such as transport and communication networks [4]. The lack of research in this area might have just pragmatic reasons: The above mentioned proxy-data sets usually do not involve any spatial information. And even if they provide spatial information, it often comes only with a coarse resolution at the level of municipality or cities.

The literature agrees that distance plays an important role in social networks [5,6,7,8]. This is rather unsurprising because face-to-face meetings, which are required to maintain a social contact, involve travel for at least one actor. The costs of travel usually scale in distance making the maintenance of long distance contacts more expensive. The literature also agrees that electronic information and communication technologies do not fully replace the need for physical contacts . Rather, they act as a complement [9,7,10]. Thus, understanding the role of the spatial dimension in social networks is also of importance for the forecasting of travel and communication demands.

Realising that spatial proximity influences the occurrence of social contacts rises the question how much does it contribute to the explanation of general network structures? Moreover, does the spatial distribution of individuals, which can be quite inhomogeneous in real-world (rural versus urban areas), have any impact? Considering the first question, results of theoretical studies appear to diverge from the observations of empirical studies. Models involving a spatial interaction between individuals can, under certain configurations, explain the emergence of network structures [11,12,13]. However, results of empirical studies suggest that a geographical process is not the only process that governs the organisational layout of a network [8,14]. This may indicate that the configurations considered in theory are rarely observed in reality. Considering the second question, the theoretical studies are usually based on random distributions or lattice-like layouts. Real-world distributions have gained just little attention (see for instance [15] and [16]).

This paper contributes to both questions. The results of studies presented in this article provide a further evidence that distance is a crucial factor concerning tie formation but is not the dominating variable explaining the topology of a network. We analyse empirical data describing a network of leisure contacts and show that local network structures, specifically the degree distribution and transitivity, are not determined by the spatial layout of the network. A model for large-scale social networks involving a spatial and social interaction between individuals is proposed. Simulation studies are conducted using a real-world synthetic population as input. The results show that the model is capable of reproducing the spatial structure of the observed network, but, however, fails to reproduce the topological characteristics.

The remaining article is organised as follows: Section 2 provides an overview of empirical studies on social networks and models for spatially embedded networks. In Sec. 3 a detailed analyses of the empirical data is presented. We turn to the description of the simulation model in Sec. 4 and present the simulation studies in Sec. 5. The paper is closed with a discussion of the results of the empirical analysis as well as the simulation studies in Sec. 6 and a conclusion in Sec. 7.

2 Related work

The influence of geographical distance on the presence and strength of social ties has been already identified in the middle of the 20th century (for instance [3]). From a personal perspective it is intuitive that the probability of a social contact decreases with increasing distance. Several empirical studies, however, have shown that this rather common assumption is well described by the so-called *gravity model* [17]. The gravity model can be used to explain to distinct observations: (i) the strength of a social contact (which may be quantified by the frequency of physical meetings) in dependency of distance or (ii) the occurrence of a social contact in a specific distance. The later is naturally dependent on the underlying spatial distribution of individuals and can be normalised accordingly.

Latané et al. [5] study three data sets describing the social interaction of college students, citizens of Florida and social psychologists. In all three data sets they observe that the interaction frequency is well approximated by a power law d^{-1}, where d denotes the distance between two individuals.

The analysis of a social network of bloggers by Liben-Nowell et al. [15] revealed that the relationship between friendship probability and distance can be described by $p \sim d^{-1}$. They determine the friendship probability by dividing the observed number of friendships with distance d by the number of possible pairs of individuals with same distance.

The same approach to calculate the probability that two individuals are connected is used by Lambiotte et al. [14], however, using a communication network constructed from mobile phone data. Lambiotte et al. find that the probability follows the same fundamental scaling law as in [15] but with a considerably smaller exponent: $p \sim d^{-2}$. Moreover, Lambiotte et al. observe that transitive connections (person i is connected to j, j is connected to k and k is connected to i) are not only composed of spatially adjacent persons, as one may expect considering that the density of social contacts is greater in the direct proximity, but that they can stretch out over large distances.

Frei and Axhausen [18] surveyed personal networks of respondents located in the metropolitan area of Zurich, Switzerland. The respondent were requested to report emotional important social contacts together with their residential location. To account for the inhomogeneous population distribution they divide the share of observed contacts by the population share, where both quantities are aggregated into concentric rings centred at the centre of mass of all respondents' residential locations. The calculated ratio of contact and population share exhibits a strong decay in distance, however, Frei and Axhausen do not make statements about the parametric form of the distribution.

In a recent study Daraganova et al. [8] consider various types of exponential and power-laws to describe the distance distribution of social contacts. The analysis of a data set of 551 individuals shows that the family of power-laws result in better fits of the distribution compared to exponential decay functions. For all considered power-laws the presence of edges decays in distance with an exponent of ≈ -1.

While there are a couple of models dealing with the generation of networks embedded in space, only few of them specifically address the generation of social networks. A quite intuitive approach to generate a spatially embedded network is to extend the model of preferential attachment by Barabási and Albert [19]: In the so-called *modulated BA* [20] the preference to attach to high degrees competes with the preference for short edges. The probability of a vertex introduced at time t to connect to an existing vertex i is adapted according to $p_i(t) \sim k_i(t)\, d^{\alpha}$, where $k_i(t)$ denotes the degree of i at time t, d denotes the euclidian distance between both vertices and α is a (usually negative) parameter. A further extended version of this model is used by Barrat et al. [11] to create weighted networks. In their model the preferential attachment mechanism relies not only on the vertex's degree but on the sum of the weights of the edges connect to a vertex.

In the *geographical threshold graph* model of Masuda et al. [21] two vertices i and j are connected according to the threshold mechanism $(w_i + w_j)\, h(d_{ij}) \geq \theta$, where w_i and w_j are a priori defined weights, $h(d_{ij})$ represents a decreasing function of distance d_{ij} and θ is a constant threshold.

Other models consider vertices placed on a lattice that connect to their nearest neighbours [22] or graphs where regions are iteratively partitioned (e.g. by triangulation [23]) into subregions by introducing new vertices and edges. Details on the above four models can also be found in a review of geographical scale-free networks by Hayashi [24].

Models that explicitly address the construction of social networks are, for instance, the models of Boguñá et al. [25], Wong et al. [26], Liben-Nowell et al. [15], Lambiotte et al. [14] and Daraganova et al. [8]. Boguñá et al. introduces the concept of *social distance* attachment in which vertices are connected with probability $p_{ij} = 1/\left(1 + (d_{ij}/b)^{\alpha}\right)$, where d_{ij} denotes the distance between both vertices in the *social space*, b and $\alpha > 1$ are parameters. If one considers the social space as a two-dimensional euclidean space, then this model connects vertices with probability just depending on their distance; probably the most simple model for a spatially embedded network.

Wong et al. [26] propose an exponential random graph model in which they describe the probability of an edge as a simple step function differentiating between edges with and beyond a so-called *neighbourhood radius* H. More precisely, the probability to connect vertices i and j is given by $p_{ij} = p + p_b$ if $d_{ij} \leq H$ or $p_{ij} = p - \Delta$ if $d_{ij} > H$, where p denotes the average edge density, p_b represents the *proximity bias* controlling the users sensitivity towards distance and Δ is a correction term to maintain the average edge density.

The model of Liben-Nowell et al. [15] explicitly accounts for the spatial population distribution. More precisely, the population distribution is the only variable in their model: The probability that persons i and j are connected is described by the reciprocal of the number of persons living closer to i than j does.

Lambiotte et al. [14] developed a model with moving agents which allows to explain why triangles are approximately equally distributed over space. Agents which are placed on a periodic one-dimensional lattice are either allowed to move and thus to deform and stretch triangles or to adapt to their neighbourhood by replacing long-distance connections with short-distance connections forming new local triangles.

An exponential random graph model involving geographical proximity is proposed by Daraganova et al. [8]. The model combines spatial processes described by a power-law equivalent to the function used by Boguñá et al. and network processes describing the emergence of star-like and triangular configurations. While some of the above theoretical models can be configured so that spatial processes can explain other network structures such as transitivity and degree correlation [25, 26, 11], Daraganova et al. conclude with a contrary statement. The discrepancy may be explained with the fact that the model of Daraganova et al. is fitted against real-world data. In fact, a common characteristic of most of the above presented studies is that they either use (i) random or poisson distributed vertices in a one or two-dimensional euclidean space, (ii) vertices positioned on a lattice, or (iii) vertex locations that are a result of the generating algorithm itself. Literature dealing with inhomogeneous vertex distributions is sparse (see for instance [15] and [16]). The present study addresses this knowledge gap. The proposed network model uses a real-world synthetic population as input and thus allows to gain insights into the effects of the population distribution on network structures.

3 Analysing empirical data

3.1 Data collection

Empirical data is obtained from a survey that collects data on a social network of leisure contacts in Switzerland [27]. The sampling design involves a so-called *snowball sampling* technique. In a snowball sample, respondents are asked to report their social contacts, which are then invited to participate in the survey as well. The new respondents are asked to report their social contacts which in turn also are invited. This iterative process is continued until a predefined number of iterations is conducted or the desired number of samples is collected. The name of the approach stems from the image of a snowball accumulating more and more material when it is rolled through the snow.

The drawback of snowball sampling is its inherent bias. Each additional contact of a person means an additional edge through which the sampling mechanism can find that person. This means that persons with many contacts are more likely to be included in the sample, and hence the resulting sample is biased towards respondents with many social contacts. With appropriate methods that account for the *degree bias* it is possible to obtain corrected statistics [28].

Each respondent and each reported person represents a vertex in the network. An edge between to vertices i and j denotes that either i named j as a contact or vice versa. The resulting network contains more than 7000 vertices and 7600 edges sampled within three snowball iterations. 406 vertices represent the respondents, i.e. persons that filled out a questionnaire. The remaining vertices represent the contacts that have been named by the respondents. This distinction is crucial because the degree (number of leisure contacts) is only known for respondents. For the majority of vertices socio-demographic attributes such as age and gender are known. Roughly 75 % of all vertices disclosed their residential location.

3.2 Network topology

The observed network exhibits a corrected mean degree of $\langle k \rangle = 13.2$. The corresponding degree distribution (Fig. 1) is heavily right-skewed, with a maximum degree of 41. Without detailed statistical test it is not possible to decide whether the distribution is exponential or log-normal. Other studies on social networks observe a decrease of the probability towards the very low end of the degree scale [18, 29]. It is thus plausible to fit a log-normal distribution into the data points. This is also supported by the presumption that missing samples at the very low end of the degree scale are a result of the sampling design which despite the correction has difficulties to capture vertices with low degree.

A further commonly observed property of social networks are transitive connections (the probability that the friend of my friend is also my friend). A method to quantify transitivity is the local clustering coefficient [30]:

$$C = \frac{1}{N} \sum_i \frac{2m_i}{k_i \left(k_i - 1\right)} \ , \tag{1}$$

where m_i denotes the number of edges that connect neighbours of i and N denotes the total number of vertices. A draw-back of snowball sampling is that it misses edges between neighbours (i.e. it underestimates m_i) if not at least one neighbour connected to such an edge participates in the survey. To compensate for this, the survey additionally collects data on neighbour-neighbour relations using a *sociogram*. In a sociogram respondents are asked to define activity-groups (for instance “hiking group” or “soccer club”) and assign their contacts to those groups. Connecting all persons within an activity-group with each other reveals the missing edges between the respondent’s neighbours. Of course, two persons being in the same activity-group do not necessarily consider each other as a leisure contact. A discussion of this aspect is, however, out of the scope of this paper. Including the edges obtained from the sociogram the clustering coefficient is $C = 0.21$[1]. A dependence of the clustering coefficient on the vertex’s geographical location can not be identified.

3.3 Spatial network properties

3.3.1 Distance distribution

Given the residential locations of vertices the (orthodromic) length of edges can be calculated (Fig. 2(a)). The resulting edge length distribution $p_{edge}(d)$ breaks up into a short range and a long range domain with the transition at about 20 km distance. Both domains follow a power law distribution

$$p_{edge}\left(d\right) \sim d^{\beta_{1/2}} \tag{2}$$

with $\beta_1 \approx -0.6$ for the short range and $\beta_2 \approx -1.8$ for the long range.

[1] Edges obtained from the sociogram are only used for the analysis of transitivity but ignored for all other analyses.

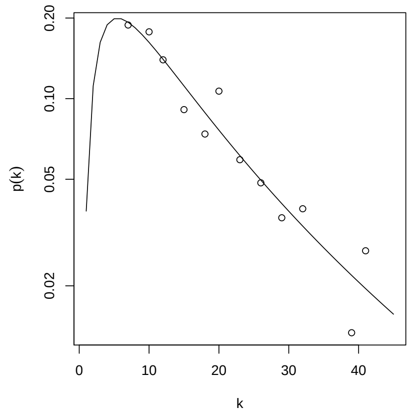
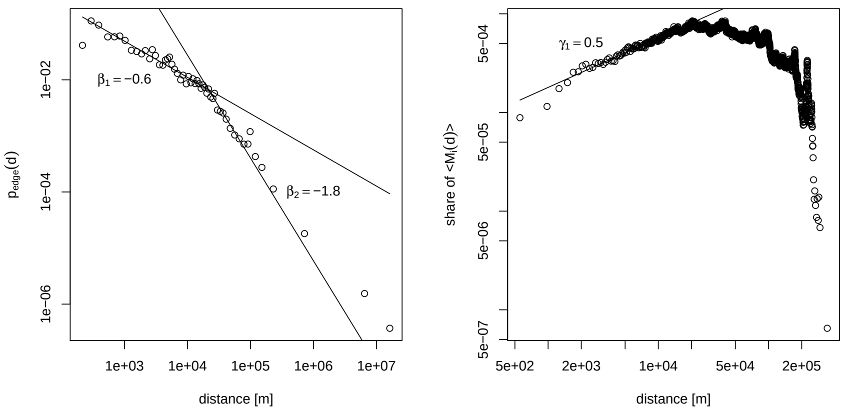


Fig. 1 Corrected degree distribution. Samples are averaged such that each data point represents the same number of samples. The solid line shows a fitted log-normal distribution with $\sigma = 0.9$ and $\mu = 2.6$.



(a) (b)

Fig. 2 (a) Edge length distribution of social contacts $p_{edge}(d)$. (b) Distribution of opportunities (inhabitants) $\langle M_i(d) \rangle$, averaged over the survey population. Samples are aggregated so that each data point represents (approximately) the same number of samples ((a): $n = 100$, (b): $n \approx 30'000$).

3.3.2 Acceptance probability

The distribution can be decomposed into the product of a behavioural model $p_{accept,i}(d)$, i.e. the probability of i to accept a person at distance d as social contact, and the underlying population distribution $M_i(d)$, i.e. the number of persons at distance d from individual i:

$$E(m_i(d)) = p_{accept,i}(d) \cdot M_i(d) \ , \tag{3}$$

where $E(m_i(d))$ is the expectation value of the number of contacts that person i possesses in distance d. Rearranging Eq. 3 yields

$$p_{accept,i}(d) = \frac{E(m_i(d))}{M_i(d)} \ . \tag{4}$$

Applying the population average operator results in

$$\left\langle p_{accept,i}(d) \right\rangle = \left\langle \frac{m_i(d)}{M_i(d)} \right\rangle \ , \tag{5}$$

where the person-specific expectation value $E(.)$ is subsumed in the population average $\langle . \rangle$. The population distribution, $M_i(d)$, can be obtained by counting the number of inhabitants at distance d for each respondent i. For this, a 1 % sample of the Swiss population obtained from the Swiss micro-census [31] is used, and distance d is discretised into 300 bins (rings) with the width adjusted so that all bins contain the same number of inhabitants. Areas outside Switzerland contribute zero opportunities.

Figure 3(a) shows $\left\langle p_{accept,i}(d) \right\rangle$, where the population average $\langle . \rangle$ is approximated by an average over the survey population. The distribution is well approximated by the power law $\left\langle p_{accept,i}(d) \right\rangle \sim d^{-1.4}$. The fact that $\left\langle p_{accept,i}(d) \right\rangle$ does not exhibit the breakup in a short and long range domain as the edge length distribution indicates that this effect is caused by the inhomogeneous distribution of the underlying population. Figure 2(b) shows $\langle M_i(d) \rangle$, where $\langle . \rangle$ denotes again the average over the survey population. The population distribution exhibits the same short and long range breakup as the edge length distribution. The sign of the slope changes from positive to negative at the transition, i.e. the number of opportunities first increases and then decreases in d. As a matter of the survey design, respondents are neither equally distributed over space, nor distributed proportionally to the population density, but are instead concentrated in Canton Zurich. The change of the slope in $\langle M_i(d) \rangle$, and consequently in $p_{edge}(d)$, can thus be explained with the border of Switzerland to Germany which is approximately 20 km north of Zurich. A behavioural interpretation of this would be that individuals do not consider persons beyond the national boundary as opportunities for social contacts.

3.3.3 Accessibility of the home location

A question at this point is if there is an acceptance probability that can be seen, at least in leading order, as independent from the person, i. For this, define

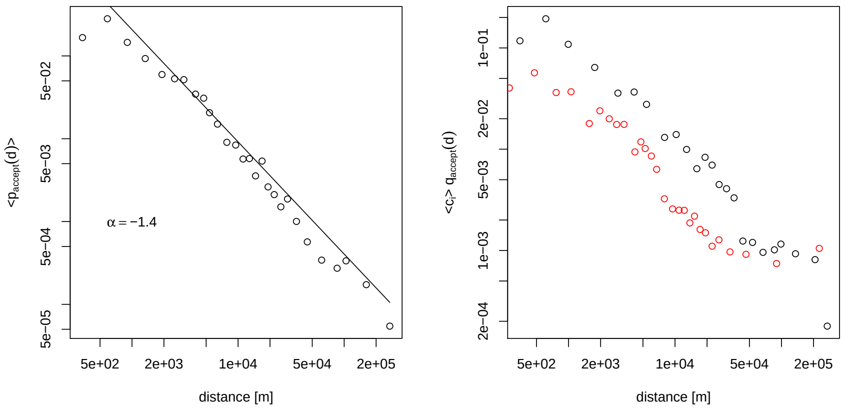
$$p_{accept,i}(d) =: c_i \, q_{accept}(d) \tag{6}$$

In order to test this, first re-express Eq. 3 by

$$E(m_i(d)) \overset{?}{=} c_i \, q_{accept}(d) \cdot M_i(d) \ , \tag{7}$$

where $\overset{?}{=}$ denotes that this is, at this point, a hypothesis. Then, rearranging and applying the population average operator yields

$$\langle c_i \rangle \, q_{accept}(d) \overset{?}{=} \left\langle \frac{m_i(d)}{M_i(d)} \right\rangle \ . \tag{8}$$



(a) (b)

Fig. 3 Probability to accept an opportunity as social contact. (a) $\langle p_{accept,i}(d)\rangle$, where the average goes over the survey population. (b) $\langle c_i\rangle\, q_{accept}(d)$ for two categories with different accessibility A_i, where the average goes over the population of the category. Black points indicate the category with low accessibility (rather rural areas), red points indicate the category with high accessibility (rather urban areas). Samples are aggregated so that each data point represents approximately the same number of samples ((a): $n = 200$, (b): $n = 100$).

To validate the hypothesis with the empirical data a measure of accessibility needs to be defined:

$$A_i := \sum_d q_{accept}(d) \cdot M_i(d) \equiv \sum_j q_{accept}(d_{ij}) \ . \tag{9}$$

Clearly, it is a sum over all opportunities, where the opportunities are weighted by the probability to actually accept them. The indicator does not depend on person i directly, but only on her home location.

For each vertex, its accessibility is calculated, and vertices are split into two categories at the median accessibility value. Thus, one obtains a category of vertices located in rather rural areas (low accessibility) and a category of vertices located in rather urban areas (high accessibility). Calculating $\langle c_i\rangle\, q_{accept}(d)$ conforming to Eq. 8 for both categories (where $\langle . \rangle$ is the average over the population of the category) reveals that both exhibit the same slope (except at distances above 50 km where the observations are due to the small sample size less meaningful) but a different offset in the log-log plot (Fig. 3(b)). The data points with the low offset, i.e. small c_i, correspond to the category with high accessibility.

3.3.4 Behavioral interpretation

It is useful to consider the meaning of c_i. If we assume the degree k_i of person i as given, one has

$$k_i = \sum_d E(m_i(d)) = c_i \sum_d q_{accept}(d) \cdot M_i(d) =: c_i\, A_i \ . \tag{10}$$

Eq. (10) can be re-written as

$$\frac{k_i}{c_i} = A_i \ .$$

That is, good accessibility either goes along with a large degree k_i or a small constant c_i. According to the first hypothesis, persons living in metropolitan areas, i.e. with high accessibility, should have high degrees k_i, and persons living in rural areas, i.e. with low accessibility, should have low degrees. However, an analysis of the empirical data regarding a correlation of degree and accessibility does not support this hypothesis. The alternative is that the constant c_i is not equal for all locations, but related to the degree over accessibility:

$$c_i = \frac{k_i}{A_i} \ . \tag{11}$$

An interpretation is that the number of social contacts is a strong personal preference, and it is so strong that it is neither modified by the accessibility of the home location, nor is there self-selection into locations that correspond to the preference for social contacts.

In addition, the scaling law of the acceptance probability, $\sim d^\alpha$ with $\alpha \approx -1.4$, seems to hold across the population. This is both independent from the degree and from the accessibility of the home location.

In consequence, as the only remaining degree of freedom, the pre-factor c_i varies by accessibility. That is, persons compensate for reduced accessibility neither by modifying the number of their social contacts not by modifying the functional form of their distance distribution. Rather, they have a larger a-priori probability to accept an opportunity as a contact.

3.4 Age and gender homophily

Analogous to spatial distance, decreasing "social distance" between two individuals increases the probability of being connected, where "social distance" denotes a measure of how much two individuals differ in their socio-demographic attributes. In social network analysis this phenomenon is known as homophily [32].

The attribute which induces the strongest degree of homophily is age, followed by gender. Both types of homophily can be quantified with the Pearson correlation coefficient of the variables' values at either ends of all edges. For a given network the Pearson correlation coefficient can be expressed by (adopted from [33])

$$r_x = \frac{\sum\limits_{(ij)} x_i x_j - M^{-1}\left(\sum\limits_{(ij)} \frac{1}{2}\left(x_i + x_j\right)\right)^2}{\sum\limits_{(ij)} \frac{1}{2}\left(x_i^2 + x_j^2\right) - M^{-1}\left(\sum\limits_{(ij)} \frac{1}{2}\left(x_i + x_j\right)\right)^2} \ , \tag{12}$$

where x_i and x_j denote the variable of vertices i and j, and M the number of edges. A correlation coefficient of $r_{age} = 0.54$ indicates a strong correlation with respect to age. Although, this correlation exists throughout all age groups the absolute age difference between ego and alter increases with the ego's age. The correlation with respect to gender, where the gender of individual i is encoded with $x_i = 0$ if i is male or $x_i = 1$ otherwise, is less pronounced ($r_{gender} = 0.34$).

4 Simulation model

Following the empirical observations we propose a simulation model which involves a spatial and social interaction. Both interaction forces are modelled with a utility function that quantifies the utility a persons gains from its social contacts. Generally, the dynamics of this model are driven by the ambition of individuals to connect to persons that are geographically close and similar in their socio-demographic characteristics. Referring to Sec. 3.3.4, the model does not attempt to describe the process that governs the degree distribution of a network. Therefore, the model takes an a priori defined degree distribution as input which remains unmodified during the simulation. The simulation comprises two steps:

1. Creating an initial random graph with an arbitrary but given degree distribution.
2. Re-organising edges so that the utility distribution of individuals converges to a stochastic user equilibrium.

A further input of the model is a synthetic population of Switzerland, i.e. a random realisation of census data such that a census on the synthetic population would approximately reproduce the original census. The information obtained from a synthetic person are its residential location, its gender, and its age. Persons are connected with undirected and unweighted edges.

4.1 Utility function

Let U_i be the utility individual i gains from its social contacts:

$$U_i = \sum_j y_{ij} U_{ij} \ , \tag{13}$$

where y_{ij} denotes the edge indicator variable which is 1 if persons i and j are connected, or 0 otherwise. U_{ij} is a composition of utility terms:

$$U_{ij} = U_{ij}^{(dist)} + U_{ij}^{(age)} + U_{ij}^{(gender)} \ . \tag{14}$$

Each utility term represent a specific interaction force:

- $U_{ij}^{(dist)}$ denotes the perceived (usually negative) utility of the geographical distance between i and j. It describes the travel costs involved to physically meet person j (at its residential location) in order to maintain the social contact.
- $U_{ij}^{(age)}$ captures homophily with respect to age and usually increases if i and j differ less in age.
- $U_{ij}^{(gender)}$ captures homophily with respect to gender and is either zero if i and j share the same gender or non-zero (and usually negative) otherwise.

4.1.1 Spatial interaction

The spatial interaction model is based on the assumption that the maintenance of a social contact requires regular physical contact between individual i and j. This means that either i visits j at its residential location or vice versa. Of course, the physical contact does not necessarily have to be located at the actors' residential locations. However, as the survey data does not include information about joint activities we stick to this simplified assumption. Further assume that the probability that i is willing to make a trip to j is proportional to the acceptance probability (Eq. 6), whereas we neglect constant c_i:

$$p_{trip,ij} \sim q_{accept}(d_{ij}) = d_{ij}^{\alpha} \ . \tag{15}$$

Since the network model considers undirected edges the missing constant of proportionality in the above equation needs to ensure the symmetry $p_{trip,ij} = p_{trip,ji}$. To solve this two-sided constrained problem we turn to the *production-attraction constrained* gravity model of Wilson [17]. The gravity model describes the probability of making a trip from i to j by

$$p_{trip,ij} = C_i \, D_j \, k_i \, k_j \, d_{ij}^{\alpha} \ , \tag{16}$$

where the *mass terms* are replaced by the degrees of i and j, and d_{ij}^{α} represents the impedance function. C_i and D_j are the *balancing factors* that ensure the symmetry of $p_{trip,ij}$. The balancing factors are specified in accordance to the constraints

$$\sum_j p_{trip,ij} = k_i \text{ and } \sum_i p_{trip,ij} = k_j \ , \tag{17}$$

so that

$$C_i = \frac{1}{\sum_j D_j \, k_j \, d_{ij}^{\alpha}} \tag{18}$$

and

$$D_j = \frac{1}{\sum_i C_i \, k_i \, d_{ij}^{\alpha}} \ . \tag{19}$$

To point the relation to the analysis of Sec. 3.3 more out Eq. 16 is re-written as

$$p_{trip,ij} = \tilde{C}_i \, \tilde{D}_j \, d_{ij}^{\alpha} \ , \tag{20}$$

where

$$\tilde{C}_i = \frac{k_i}{\sum_j \tilde{D}_j d_{ij}^{\alpha}} \tag{21}$$

and

$$\tilde{D}_j = \frac{k_j}{\sum_i \tilde{C}_i d_{ij}^{\alpha}} \ . \tag{22}$$

Both above equations equal the definition of the person specific constant (Eq. 11)

$$c_i = \frac{k_i}{A_i} = \frac{k_i}{\sum_j d_{ij}^{\alpha}} \tag{23}$$

if $\tilde{D}_j$ and $\tilde{C}_i$, respectively, are set to one. This corresponds to the one-sided constrained problem where the symmetry of $p_{trip,ij}$ is dropped. Regarding the analysis of Sec. 3.3 this means that the surveyed graph is considered to be *directed*.

One will see quickly that the equations for the balancing factors (Eq. 21 and Eq. 22) can only be solved numerically. However, as it will turn out later the determination of the balancing factors is not necessary as they cancel out during the process of re-ordering edges.

The logit random utility model [34] commonly used in transport planning describes the functional relation between the impedance function and the (dis)utility of distance $U_{ij}^{(dist)}$:

$$p_{trip,ij} = \tilde{C}_i \, \tilde{D}_j \, d_{ij}^{\alpha} \ \sim \exp\left(U_{ij}^{(dist)}\right) \ . \tag{24}$$

Isolating $U_{ij}^{(dist)}$ in the above equation yields

$$U_{ij}^{(dist)} = \alpha \ln d_{ij} + \ln \tilde{C}_i + \ln \tilde{D}_j + \text{const.} \tag{25}$$

The parameter α, formerly the exponent in the power-law, represents now the pre-factor of the spatial interaction force. Constants $\ln \tilde{C}_i$ and $\ln \tilde{D}_j$ both stem from the balancing constraints and are thus "extrinsic" utilities, better to be interpreted in terms of the choice probability: $\ln \tilde{C}_i$ expresses the fact that someone with a high degree or a low accessibility needs to have a larger a priori probability to accept in order to reach her desired number of contacts. In a logit choice model, this contribution is equal for all alternatives and thus will cancel out. $\ln \tilde{D}_j$ expresses the fact that, if the system is to remain balanced, one should have a larger probability to accept someone with a large degree k_j or with a low accessibility A_j. It can be seen as an expression quantifying that "competition for slots" will be reduced when alters either accept many contacts or when they are difficult to reach.

For better readability we change the notation of α for the remaining article to $-\alpha^{(dist)}$. This emphasizes that the pre-factor is usually negative and distinguishes from parameters $\alpha^{(age)}$ and $\alpha^{(gender)}$ which are introduced in the next section. Thus, Eq. 25 finally reads:

$$U_{ij}^{(dist)} = -\alpha^{(dist)} \ln d_{ij} + \ln \tilde{C}_i + \ln \tilde{D}_j + \text{const} \ . \tag{26}$$

4.1.2 Social interaction

The utility describing homophily in age is quantified by the product of the relative error from i to j and vice versa:

$$U_{ij}^{(age)} = -\alpha^{(age)} \frac{(a_i - a_j)^2}{a_i a_j} \ , \tag{27}$$

where $a_i > 0$ and $a_j > 0$ denote the age of i and j, and $\alpha^{(age)}$ is a parameter controlling the strength of homophily. This specification ensures the symmetry $U_{ij}^{(age)} = U_{ji}^{(age)}$. Furthermore, it accounts for the observation that the absolute age difference becomes less important with increasing age (Sec. 3.4).

The quantification of $U_{ij}^{(gender)}$ is straight forward:

$$U_{ij}^{(gender)} = -\alpha^{(gender)} g_{ij} \ , \tag{28}$$

where g_{ij} is a binary variable which is 0 if i and j are of same gender or 1 otherwise and $\alpha^{(gender)}$ controls the strength of homophily with respect to gender. The symmetry of g_{ij} is inherently given.

4.2 Creating an initial random graph

Creating a graph with an arbitrary degree distribution can be easily done with the following algorithm: First, generate a sequence of degrees $\{k_i\}$ corresponding to a given degree distribution and then randomly assign each person a *target degree* (or *edge stubs*) out of the degree sequence. Second, pairs of vertices are randomly chosen and connected if the current degree of both vertices is less than their target degree. This process is continued until all vertices reached their target degree.

Since the graph is undirected, the sum of all degrees $\sum_i k_i$ needs to be even in order to properly connect all edge stubs. Cases may occur where all edge stubs can only be connected by inserting multiple edges between the same vertex pair which, however, is not allowed. If the degree sequence does not satisfy both conditions it is discarded and a new sequence is generated. The obtained graph exhibits the given degree distribution but is random with respect to all other network properties.

4.3 Re-organising edges

Given an initial graph with the desired degree distribution edges are re-organised using a Markov Chain Monte Carlo simulation. Let $\boldsymbol{Y}_t$ be the graph at time t and $\boldsymbol{Y}_{t+1}$ the graph after a step which is defined as:

1. Randomly draw two connected pairs (ij) and (uv) of vertices satisfying the conditions $i \neq j$, $u \neq v$, $i \neq u$, $i \neq v$, $j \neq u$, and $j \neq v$.
2. With probability π_{t+1}, move edge (ij) to (iu) and edge (uv) to (jv).

The above method does not change the vertices' degrees. The transition probability from $\boldsymbol{Y}_t$ to $\boldsymbol{Y}_{t+1}$ is defined as

$$\pi_{t+1} = \frac{e^{U(\boldsymbol{Y}_{t+1})}}{e^{U(\boldsymbol{Y}_{t+1})} + e^{U(\boldsymbol{Y}_t)}} \ , \tag{29}$$

where $U(\boldsymbol{Y}) = \sum_{i<j} U_{ij}$ denotes the total utility in the graph $\boldsymbol{Y}$. Re-arranging Eq. 29 shows that one only needs to evaluate the utility difference ΔU:

$$\pi_{t+1} = \frac{1}{1 + e^{\Delta U}} \ , \tag{30}$$

where

$$\begin{aligned} \Delta U &= U(\boldsymbol{Y}_t) - U(\boldsymbol{Y}_{t+1}) && (31)\\ &= U_{ij,t} + U_{uv,t} + U_{iu,t} + U_{vj,t} && (32)\\ &\quad - U_{ij,t+1} - U_{uv,t+1} - U_{iu,t+1} - U_{vj,t+1} \ . \end{aligned}$$

Given the dyad states

- $y_{ij} = 1$, $y_{uv} = 1$, $y_{iu} = 0$, and $y_{vj} = 0$ for configuration $\boldsymbol{Y}_t$,
- $y_{ij} = 0$, $y_{uv} = 0$, $y_{iu} = 1$, and $y_{vj} = 1$ for configuration $\boldsymbol{Y}_{t+1}$

Eq. 31 collapses to

$$\Delta U = U_{ij} + U_{uv} - U_{iu} - U_{jv} \ . \tag{33}$$

Inserting Eq. 14 reveals that all constants, specifically the balancing factors in $U^{(dist)}$, cancel out. This means that with an a priori given degree distribution it is not required to solve the balancing factors. This is rather unsurprising because $\tilde{C}_i$ and $\tilde{D}_j$ are only dependent on the vertex's degree and its geographical location. Removing the degree distribution as a degree of freedom removes also $\tilde{C}_i$ and $\tilde{D}_j$ as a variable in the model.

5 Simulation studies

Simulation runs are conducted with varying strength of spatial and social interaction, and varying size of the network. Considered populations are random draws from the entire synthetic Swiss population containing between 500 and 40'000 persons. Depending on the size of the network the Monte Carlo Markov Chain is run for $5 \cdot 10^8$ to $2 \cdot 10^{10}$ steps.

5.1 Degree distribution

For each considered population and parameter configuration an initial random graph is created according to the method described in Sec. 4.2. The degree sequence is drawn from a log-normal distribution configured with the parameters roughly set to the observed values: $\sigma = 1$, $\mu = 2.5$ and a maximum degree of 41. It is noteworthy that this process is independent from the spatial distribution of vertices and thus there is no correlation between the vertices' degree and their location.

For large networks ($> 10'000$ vertices) it is no problem to generate a valid graph. The probability that a graph meets the given degree distribution is in the order of 10^{-2}. With decreasing network size it becomes more difficult to generate a graph that does not violate the constraints. The probability to generate a valid graph with 500 vertices lies in the order of 10^{-4}.

5.2 Spatial properties

To validate the model's capabilities to reproduce the observed spatial network structure, a simulation configuration with 40'000 persons, spatial interaction set to the observed value $\alpha^{(dist)} = 1.4$ and no social interaction ($\alpha^{(age)} = 0$ and $\alpha^{(gender)} = 0$) is analysed in detail.

Figure 4 shows the edge length distribution of the simulated network. The distribution follows the power law $p_{edge}(d) \sim d^{\beta}$ with $\beta \approx -0.8$ with an exponential cutoff towards the borders of Switzerland. The distribution does not exhibit the split into a short and long range domain as observed in the empirical data (Fig. 2(a)). This indicates that the effect introduced by the northern border of Switzerland vanishes if one considers a representative sample of the entire Swiss population.

In analogy to Sec. 3.3, we extract $\langle p_{accept,i}(d) \rangle$ and validate if the simulated network exhibits the correct behavioural model (Fig 5(a)). Except for very short distances ($\lesssim 3$ km), $\langle p_{accept,i} \rangle$ follows the power law d^{α} with $\alpha \approx -1.4$. For the very short distances the slope is slightly flatter. The effect becomes even more pronounced if the network size deceases or if one considers the population located

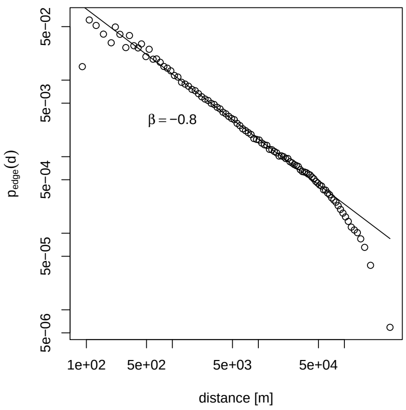


Fig. 4 Edge length distribution of a simulated network with 40'000 vertices and $\alpha^{(dist)} = 1.4$.

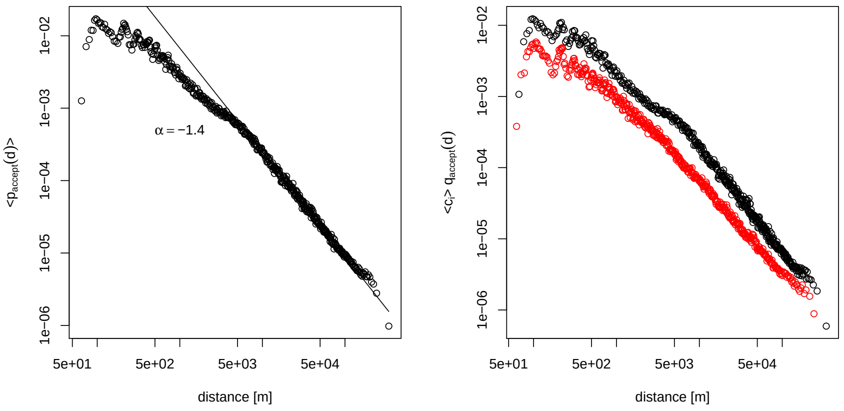
in rural areas (Fig. 5(b)). The observation indicates that this inaccuracy is caused by an insufficient sample size: The behavioural model cannot be upheld if the density of opportunities in the direct proximity is too sparse. Choosing a sample size that exhibits a sufficient density of opportunities also in rural areas would, however, lead to infeasible computation times.

In analogy to the analysis of the empirical data, we test if the acceptance probability is in leading order independent of individual i. Therefore, $\langle c_i \rangle \, q_{accept}$ is calculated, again, for two sub-populations with different accessibility (Fig 5(b)). Both distributions exhibit the same slope but a different offset: The distribution with the smaller offset represent the sub-population with higher accessibility. Although, $\tilde{C}_i$ and $\tilde{D}_j$, which represented the two-sided constrained counterparts of c_i and c_j in the simulation model, are not explicitly given, they are implicitly given by the a priori given degree distribution and the geographical location of i and j.

Investigating the spatial distribution of edge lengths reveals that there is a clear spatial separation of edges regarding their lengths. Short edges are concentrated in areas with high accessibility. This result conforms to observations also made with the empirical data (Fig. 6).

5.3 Homophily

Using the parameter setup for size and $\alpha^{(dist)}$ as above the effects of social interaction are investigated by varying $\alpha^{(age)}$ and $\alpha^{(gender)}$. With increasing values of $\alpha^{(age)}$ and $\alpha^{(gender)}$, respectively, the correlation coefficients asymptotically approximate 1, i.e. the perfect assortative network where edges exclusively connect vertices of same age and gender, respectively. To obtain the empirically observed values $r_{age} = 0.54$ and $r_{gender} = 0.34$, the pre-factors need to be set to $\alpha^{(age)} = 0.9$ and $\alpha^{(gender)} = 0.8$. The edge length distribution shows no measurable change if the pre-factors of $U^{(age)}$ and $U^{(gender)}$ are varied. Moreover, $\alpha^{(age)}$ and $\alpha^{(gender)}$ show no effect on the other property, r_{gender} and r_{age}, respectively.



(a)

(b)

Fig. 5 Probability accept an opportunity as social contact. (a): $\langle p_{accept,i}(d)\rangle$ where the average goes over the synthetic population. (b): $\langle c_i\rangle\, q_{accept}(d)$ for two categories with different accessibility A_i, where the average goes over the population of the category. Black points indicate the category with low accessibility (rather rural areas), red points indicate the category with high accessibility (rather urban areas).

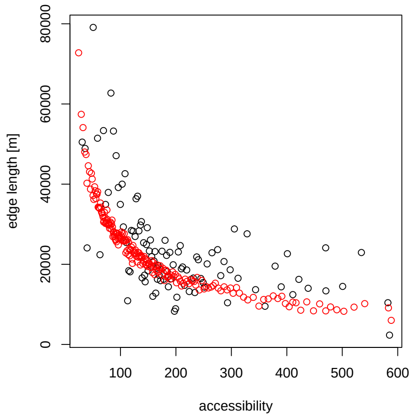


Fig. 6 Mean edge length over accessibility. Black = empirical data, red = simulated data.

5.4 Transitivity

A simulation configuration with the interaction forces set to the values corresponding to the empirical observation ($\alpha^{(dist)} = 1.4$, $\alpha^{(age)} = 0.9$ and $\alpha^{(gender)} = 0.8$) and a network size of 40'000 individuals produces networks that exhibit no transitivity.

Varying the network size and spatial interaction shows that one either needs to decrease the population size or increase $\alpha^{(dist)}$ to obtain transitivity (Fig. 7). Because the mean degree is constant for all configurations, decreasing the population size naturally increases the probability that two vertices have a common

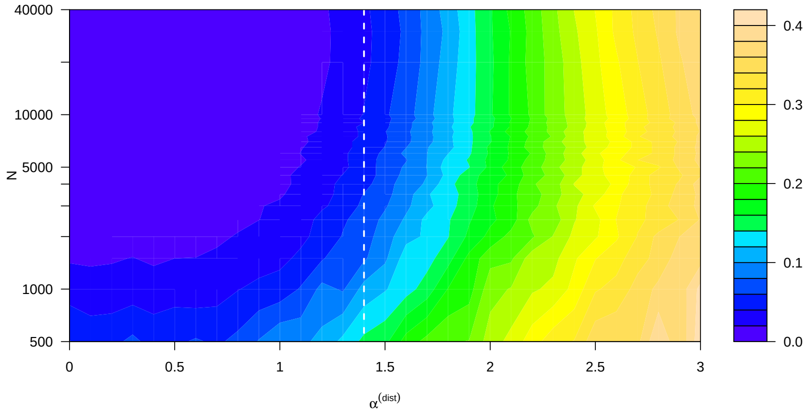


Fig. 7 Transitivity in dependency of network size and $\alpha^{(dist)}$ ($\alpha^{(age)} = 0$ and $\alpha^{(gender)} = 0$). Colours indicate the value of the local clustering coefficient. The white vertical line indicates the empirical observed value of $\alpha^{(dist)}$.

neighbour. A similar effect is obtained if one increases $\alpha^{(dist)}$ (while holding the size constant): It forces the individuals to connect to vertices in their direct spatial proximity which increases the probability of two vertices to have a common neighbour as well. Using the empirically observed value of $\alpha^{(dist)} = 1.4$, significant transitivity ($C = 0.14$, still less than the observed $C = 0.21$ in the survey data) is only observable for network sizes of 500 vertices. Generally, the sensitivity of transitivity towards the network size diminishes for population sizes of more than approximately 5000 vertices. The sensitivity towards the spatial interaction force is present throughout all considered networks sizes.

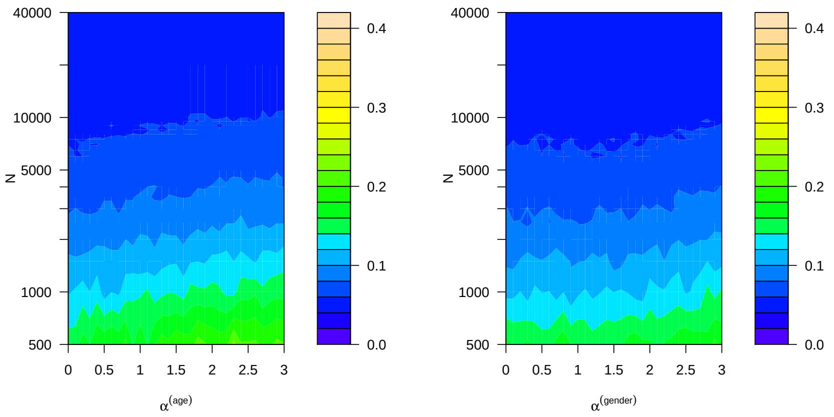
Referring to the social interaction force based on age, the sensitivity is of considerably smaller magnitude (Fig. 8(a)). Decreasing size and increasing $\alpha^{(age)}$ while holding the spatial interaction constant at $\alpha^{(dist)} = 1.4$ shows only a minor increase of transitivity. The increase of transitivity for small networks is predominantly governed by the spatial interaction.

Variations of $\alpha^{(gender)}$ result in only minor changes of transitivity, regardless of network size (Fig. 8(b)). Even in a perfect assortative network (with respect to gender) the remaining set of candidates is still so large that the probability of having a common neighbour remains small.

Networks with high transitivity exhibit a negative correlation between accessibility and the local clustering coefficient. This is obvious because if the density of candidates in the direct proximity increases it is less likely that two vertices share the same vertex as a contact. Yet, this observation contradicts the observation from the empirical analysis (Sec. 3.2): The survey data shows no correlation between the local clustering coefficient and the geographical location.

6 Discussion

The analysis of empirical data shows that individuals living in rural areas appear to compensate for the lower accessibility with a higher a priori probability to



(a) (b)

Fig. 8 (a) Transitivity in dependency of network size and $\alpha^{(age)}$ ($\alpha^{(dist)} = 1.4$ and $\alpha^{(gender)} = 0$). (b) Transitivity in dependency of network size and $\alpha^{(gender)}$ ($\alpha^{(dist)} = 1.4$ and $\alpha^{(age)} = 0$). Colours indicate the value of the local clustering coefficient.

accept other persons as social contacts. Consequently, the number of contacts a person posses is independent from its geographical location. The translation of this finding into a utility function shows that the a priori probability is expressed by a constant in the utility model. In terms of travel costs, the constant represents the costs involved in reaching the next node of the transportation network, for instance the next highway or transit stop. The usually lower density of the transportation network in rural areas yields higher travel costs in these areas. The willingness of individuals to accept higher travel costs in order to maintain the same number of social contacts explains also the occurrence of longer edges in rural areas. The above observation suggests that the process that governs the degree distribution of a network is not, or only to a minor extent, related to the spatial distribution of individuals.

Considering social network transitivity, the empirical observations and the simulation results diverge in two aspects. First, the survey data shows that a vertex's clustering coefficient, as well as its degree, does not depend on its geographical location. Second, after adjusting the spatial interaction force, age and gender homophily force so that the simulation results match the observed distributions, the simulated network does not exhibit significant transitivity. Spatial interaction, age and gender homophily interaction become only relevant for transitivity if one considers small networks. In such regimes, distance appears to be the dominating force, followed by age, which, however, has comparable little impact. Gender appears to have no influence. Naturally, given a degree distribution, the probability that two vertices have a neighbour in common increases with decreasing network size. According to this, the assumption that an individual considers the entire population as candidates for social contacts is debatable. One may expect to observe sub-structures within the population such that individuals consider only a limited *choice set* of persons as candidates. This means that spatial and social proximity alone do not explain the emergence of triangles. The findings, however, give rise to the hypothesis that considering choice sets in a social network model could be a step towards more realistic social networks.

The findings of this work are in line with other work. Daraganova et al. [8] use exponential random graph models involving various types of spatial interaction functions to reproduce an observed network. Although the considered networks are comparably small (551 and 306 vertices), the authors conclude that spatial proximity alone does not explain the structure of the networks. From the analysis of the Live-Journal community, Liben-Nowell et al. [15] draw a similar conclusion: They find that two thirds of observed friendships are derived from geographical processes, whereas the remaining friendships are governed by some non-geographical process. The fact that the presence of triangles is almost independent from space is also observed by Lambiotte et al. [14]. Results of their simulation model suggest that this effect is caused by the mobility of individuals. If people move to new residential locations they keep existing ties and thus triangles are stretched over longer distances.

7 Conclusion

In this article we investigate the spatial structure of a social network and the impacts of spatial interaction on its topology. From the analysis of a leisure-contacts network we draw the following conclusions:

- The probability of individuals to accept other persons as social contacts scales in distance with $d^{-1.4}$. This scaling law holds across the population and is both independent from the degree and the geographical location.
- The number of contacts an individual establishes is independent from its geographical location and the spatial distribution of opportunities. In consequence, as the remaining degree of freedom persons with low accessibility, i.e. living in rural areas, exhibit a higher background probability to accept opportunities as social contacts.
- As a consequence from the above item, people living in rural areas show to have contacts that are at average more distant compared to people living in urban areas.
- The occurrence of triangles is almost independent from the spatial structure of the social network.

The above conclusions gives rise to the hypothesis that spatial interaction forces are not the solely explanatory variable for the emergence of more complex social network structures. The hypothesis is supported by simulation results of a social network model. The model involves spatial and social interaction and uses a synthetic real-world population as input. While the model is capable to reproduce the spatial structure and the characteristics of homophily with respect to age and gender it contradicts the observations of the empirical social network in the following aspects:

- The simulated networks show no significant transitivity.
- The occurrence of triangles depends on space in that the clustering coefficient decreases with increasing accessibility.
- The model requires the degree distribution to be a priori given in order to avoid a correlation between degree and accessibility.

In summary, the analysis provides a further evidence that spatial interaction is a crucial aspect of social networks. Yet, it appears that distance only explains the spatial structure of the social network. An impact on the emergence of other social network properties is not observed.

Acknowledgements This work was funded by the VolkswagenStiftung within the project "Travel impacts of social networks and networking tools".

References

1. L.A.N. Amaral, A. Scala, M. Barthélémy, H.E. Stanley, Proceedings of the National Academy of Sciences of the United States of America **97**(21), 11149 (2000)
2. M.E.J. Newman, Physical Review E **64**(016131) (2001)
3. L. Festinger, S. Schachter, K. Back, *Social Pressures in Informal Groups* (Stanford University Press, Stanford, California, 1963)
4. M.T. Gastner, M.E.J. Newman, The European Physical Journal B **49**(2), 247 (2006)
5. B. Latané, J.H. Liu, A. Nowak, M. Bonevento, L. Zheng, Personality and Social Psychology Bulletin **21**(8), 795 (1995)
6. J. Larsen, J. Urry, K.W. Axhausen, *Mobilities, Networks, Geographies* (Ashgate, Aldershot, 2006)
7. D. Mok, B. Wellman, R. Basu, Social Networks **29**, 430 (2007)
8. G. Daraganova, P. Pattisona, J. Koskinen, B. Mitchell, A. Bill, M. Watts, S. Baum, Social Networks (2011)
9. P.L. Mokhtarian, Journal of Industrial Ecology **6**(2), 43 (2002)
10. J. Larsen, J. Urry, K.W. Axhausen, Information, Communication and Society **11**(5), 640 (2008)
11. A. Barrat, M. Barthélemy, A. Vespignani, Journal of Statistical Mechanics: Theory and Experiment **5**, 1 (2005)
12. P.D. Hoff, A.E. Raftery, M.S. Handcock, Journal of the American Statistical Association **97**(460), 1090 (2002)
13. C.T. Butts, Predictability of large-scale spatially embedded networks. Tech. rep., Institute for Mathematical Behavioral Sciences, UC Irvine (2002)
14. R. Lambiotte, V.D. Bondel, C. de Kerchove, E. Huens, C. Prieur, Z. Smoreda, O.V. Dooren, Physica A **387**, 5317 (2008)
15. D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, A. Tomkins, Proceedings of the National Academy of Sciences of the United States of America **102**(33), 11623 (2005)
16. C. Butts, K. Carley, Spatial models of large-scale interpersonal networks. Tech. rep., Center for Computational Analysis of Social and Organizational Systems (CASOS) (2001)
17. A. Wilson, Environment and Planning **3**, 1 (1971)
18. A. Frei, K.W. Axhausen, Size and structure of social network geographies. Working Paper 439, ETH Zürich, Institute for Transport Planning and Systems (2007)
19. A.L. Barabási, R. Albert, Science **286**, 509 (1999)
20. S.S. Manna, P. Sen, Physical Review E **66**(066114), 1 (2002)
21. N. Masuda, H. Miwa, N. Konno, Physical Review E **71**(036108), 1 (2005)
22. D. ben-Avraham, A.F. Rozenfeld, R. Cohen, S. Havlin, Physica A **330**, 107 (2003)
23. J.S. Andrade, H.J. Herrmann, R.F.S. Andrade, L.R. da Silva, Physical Review Letters **94**(018702), 1 (2005)
24. Y. Hayashi, IPSJ Digital Courier **2**, 155 (2006)
25. M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, A. Arenas, Physical Review E **70**(056122), 1 (2004)
26. L.H. Wong, P. Pattison, G. Robins, Physica A **360**(1), 99 (2006)
27. M. Kowald, A. Frei, J. Hackney, J. Illenberger, K.W. Axhausen, Using an ascending sampling strategy to survey connected egocentric networks: A field work report on phase one of the survey. Working Paper 582, ETH Zürich, Institute for Transport Planning and Systems (2009)
28. J. Illenberger, G. Flötteröd, Estimating properties from snowball sampled networks. VSP Working Paper 11-01, TU Berlin, Transport Systems Planning and Transport Telematics (2011). See www.vsp.tu-berlin.de/publications

29. J.A. Carrasco, Social activity-travel behaviour: A personal network approach. Ph.D. thesis, University of Toronto (2006)
30. D.J. Watts, S.H. Strogatz, Nature **393**(440-442) (1998)
31. ARE/BFS. Mobilität in der Schweiz, Ergebnissse des Mikrozensus 2005 zum Verkehrsverhalten (2007)
32. M. McPherson, L. Smith-Lovin, J.M. Cook, Annual Review of Sociology **27**, 415 (2001)
33. M.E.J. Newman, Physical Review Letters **89**(20), 1 (2002)
34. M. Ben-Akiva, S.R. Lerman, *Discrete choice analysis* (The MIT Press, Cambridge, MA, 1985)