A model for spatially embedded social networks

Johannes Illenberger, Technical University Berlin, Germany; Gunnar Flötteröd, École Polytechnique Fédéral de Lausanne, Switzerland; Matthias Kowald, Swiss Federal Institute of Technology Zurich, Switzerland; Kai Nagel, Technical University Berlin, Germany

Abstract This paper presents a stochastic model for spatially embedded social networks based on the ideas of spatial interaction models. Analysing empirical data, we find that the probability to accept a social contact at a certain distance follows a power law with exponent -1.6. With a simulation where the spatial distribution of vertices is defined by a synthetic population of Switzerland, we can reproduce the edge length distribution observed in the empirical data as well as some other typical properties of social networks.

1. Introduction

Most work in transport planning research concentrates on daily commuting and peak hour traffic, such as trips related to work or educational purposes. However, in recent years one observers an increasing share of leisure related travel, which in the case of Switzerland has even become the dominating travel segment. Looking at the 2005 Swiss Microcensus on Travel Behaviour (ARE/BFS, 2007), one finds that more than 40 % of activities are related to leisure. A share of 44 % for kilometers travelled and a share of 50 % in the category of time spent fortifies the importance of leisure traffic.

The limited work in leisure related travel research partly has pragmatic reasons. To analyse 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 generalised travel costs by mode and infrastructure. In most situations, data of this type is available from local administrations and can be used in a rather straightforward process of modelling commuter traffic. In contrast, 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). Data of such type is very rare since on the one hand the realisation of appropriate surveys is often very costly and extensive, and on the other hand respondents are required to disclose private data.

A common way to model the social relation between individuals is to use the representation of a graph. Research on social networks dates back to the 1960ies (Wasserman and Faust, 1994) and has gained increasing interest from sociologists and also physicists in the last years, where the latter community focuses on the characteristics of social networks as complex systems (Newman, 2003). Although there is a huge amount of studies about social networks, their use for transport planning is limited. On the one hand, studies of social networks are often embedded in an institutional setting, e.g., people working at the same company, pupils of the same school or same class, movie actors or collaborating scientists, whereas a transport planner requires a less localised sample which is representative in terms of travel behaviour. On the other hand, and this is the most important aspect, the majority of those studies provide no information about the spatial dimension of the social network.

The spatial embedding of a social network is crucial for the modelling of leisure travel. It provides the spatial distribution of an actor's social contacts, i.e., the spatial distribution of her leisure activity opportunities. 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 synthetically generate data that describes the social relations required for the leisure travel modelling process. We present a statistical model to generate a spatial embedded social network and essentially address the question to what extend the spatial distribution of actors is an explanatory variable for the formation of social ties.

The remainder of this article is organised as follows: In Sec. 2., we review the literature related to our problem. Section 3. introduces some terms common in social network analysis as well as the basic formulation of the occurrence probability of a social relation. We analyse some empirical data and formulate the model in Sec. 4.. Results of the simulation are presented in Sec. 5., and the paper concludes with a discussion in Sec. 6..

2. Related Work

Previous research on the detailed spatial distribution of social networks is sparse and especially so for non-local, sometimes transcontinental contacts. Quantitative research on the social content of long distance travel is missing, as the typical travel survey has little interest in this issue. Equally, empirical analysis on the structure of complete social networks outside of institutional settings, such as workplaces, schools or clubs, is rare.

Latané et al. (1995) propose that the frequency of social interaction is described by an inverse power law with slope -1. They analyse three data sets of studies on the social interaction of students in the United States and China as well as a study on social psychologists attending the same conference. All three studies show that the interaction frequency fits well in their proposed model.

Frei and Axhausen (2007) sample ego-centric social networks in the greater area of Zurich, Switzerland. Contrary to the study of Latané et al., they randomly select respondents, which means that their study is not embedded in an institutional setting. Frei and Axhausen also analyse the distance distribution of social contacts. However, they find that the distribution does not follow a simple parametric distribution, which seems to be reasonable since the underlying population is not equally distributed over space.

Since it is difficult to compile comprehensive data sets about social contacts, the approach proposed by Brockmann and Theis (2008) may be also helpful for the spatial analysis of social networks. Brockmann and Theis use the circulation of money as a proxy for human traffic. They analyse data from the online bill tracker www.wheresgeorge.com and find that the probability $p(d)$ of a bill traversing a distance d follows the power law $p(d) \propto d^{-1.6}$. However, there is no guarantee that the traversal of a bill matches the travel of a single person, and the data gives no information about the motivation of the travel.

Research on models of spatially embedded networks is even more sparse than literature on the analysis of the spatial dimension of social contacts. Wong et al. (2005) propose an exponential random graph model to study the spatial structure of random graphs. They model the probability of an edge formation as a simple step function that only differentiates between edges within and beyond a so-called neighbourhood radius. As a simulation scenario they use randomly scattered vertices. Wong et al. study the effect of the neighbourhood radius on small-world properties and community structures, but they make no statement about the edge length distribution.

The model of Hackney and Marchal (2009) 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.

3. Definitions

3.1. Social Networks

Social network analysis uses a graph to represent social relations between individuals. A vertex in a graph represents an individual, where an edge between two vertices denotes any kind of social relation, e.g., friendship. For simplicity, only undirected and unweighted graphs are considered, i.e, edges can be traversed in both directions (if i as a friend of j then j is also a friend if i), and all edges are equal in their strength. For the analysis of a social network, the following indicators are of interest in this article:

- The edge length distribution, where the length of an edge is defined by the geographical beeline distance between both vertices.
- The degree distribution, where the degree of a vertex denotes the number of adjacent vertices, i.e., the number of acquaintances.

The starting point for the simulation of the proposed model is a synthetic population of Switzerland. The modelling process of the synthetic population is not discussed in this paper. The interested reader is referred to Meister et al. (2008) for the basic concepts. The available synthetic population includes information about the residential locations and socio-demographic attributes of all persons. Thus, a set of vertices where each vertex is defined by a person in the synthetic population can be created. This also means that the size of the network and the spatial distribution of the vertices is fixed.

3.2. Spatial Interaction Models

The two-dimensional geographical coordinate of individual i is denoted by x_i, and the Euclidean distance between individual i and j is written as $d(x_i, x_j)$ or, when applicable without ambiguity, simply as d_{ij}.

The model we will propose in Sec. 4.3. relies on the idea that the intensity of an interaction between two individuals is dependent on the distance between the individuals. Models of such type are known as spatial interaction models (Wilson, 1971). In the case of a social network, individuals translate to vertices. Denoting by $n_i(d)$ the number of vertices that are (a) adjacent to a given vertex i in the network and (b) geometrically d distant units from i away, and denoting by $N_i(d)$ the number of all vertices that are d distant from i, this modeling assumption translates into the following equation:

$$\langle n_i(d) \rangle = p_{acc}(d) N_i(d) \tag{1}$$

where $\langle \cdot \rangle$ denotes the expectation and $p_{acc}(d)$ is the distance-dependent probability that a connection is formed ("accepted") between two vertices that are distance d apart.

The next section identifies $p_{acc}(d)$ from real data.

4. Analysis of Empirical Data and Model Discussion

4.1. Data Collecting

The data used for this work is obtained using a so-called ego-centric network approach where persons are asked to report leisure contacts (Kowald et al., 2009a,b). The newly reported contacts are then asked to report their leisure contacts. This step is repeated multiple times, thus one obtains an ascending sampling strategy, also known as snowball sampling. It reveals the spatial location of the respondents and their acquaintances.

The survey is still ongoing and the data used for this work is only an early preset. The first and second iteration of the snowball sample comprises in total 140 sampled Swiss respondents, which reported in total 1768 acquaintances. Figure 1 shows the spatial distribution of respondents and acquaintances. The survey is not restricted to Switzerland, so respondents are allowed to name leisure contacts outside of Switzerland, and these contacts are also requested to participate the survey. However, the synthetic population that is used for the network generation process is only available for Switzerland. So, in the following analysis all social contacts that are not located within Switzerland are ignored.

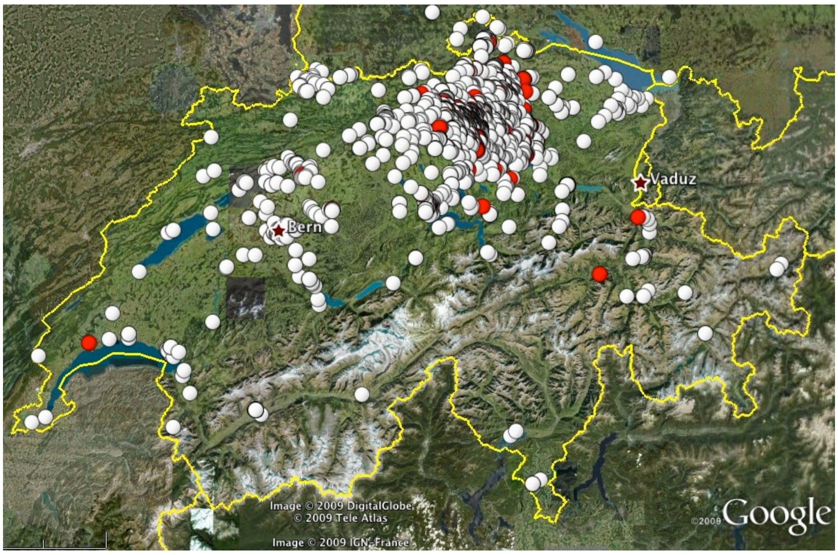

Fig. 1. Residential locations from empirical data obtained by snowball sampling. Red dots represent the respondents, white dots represent the reported acquaintances.

4.2. Data Analysis

Figure 2a) shows the edge length distribution of social contacts reported in the survey. The share of contacts decreases with increasing distance, but it does not exhibit a clear scaling law, such as a power-law or an exponential scaling. One can assume that the strong drop of the number of contacts at long distances (512 km) is due to the finite size system, which is introduced by ignoring all edges with target outside of Switzerland. Due to the small sample size at long distances we choose logarithmically scaled bin sizes (where values are re-weighted by the bin width) in the histograms as well as for the according regression analysis.

According to (1), $n(d)$ needs to be divided by $N(d)$ in order to obtain the behavioural $p_{acc}(d)$. This is done in Fig. 2b), where every instance of an edge from vertex i to vertex j of distance d_{ij} is re-weighted by $N_i(d_{ij})$, which is the number of vertices in distance d_{ij} from vertex i. To make this operational, distances are classified into distance classes of 1 km width.

A possible power law $p_{acc}(d) = cd^{\gamma}$ leads in combination with (1) to a linear model on a double-logarithmic scale:

$$\ln \frac{n_i(d)}{N_i(d)} = \ln c + \gamma \ln d + \varepsilon_i(d), \qquad (2)$$

where $\varepsilon_i(d)$ represents the unexplained random term in the regression model and $N_i(d)$ is obtained directly from the synthetic population. Figure 2b) also plots the according linear regression line into the respectively transformed histogram. Again, the data is cut off at the boundaries of the analysis zone. The power law fits well the

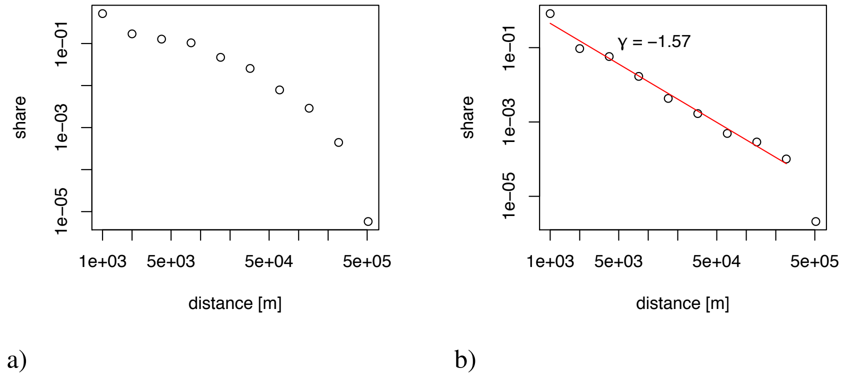

Fig. 2. a) Edge length distribution with values aggregated into logarithmic bins from empirical data. b) Acceptance probability $p_{acc}(d)$ from empirical data.

data and results in an exponent of $\gamma \approx -1.57$.

Little more information can be inferred from the currently available survey data because of its sparsity. The network built from empirical data exhibits a mean degree of 17, i.e., on average each respondent named 17 acquaintances. The sample size is still too small to determine a meaningful degree distribution. Also, clustering cannot be determined at the current stage of the survey.

4.3. Model Discussion

The above approach suggests an acceptance probability of $p_{acc}(d_{ij}) \sim d_{ij}^{\gamma}$. In comparison with other approaches, one would, however, arguably prefer an acceptance probability of $p_{acc}(d_{ij}) \sim e^{c_{ij}}$ (Wilson, 1971), where c_{ij} is the generalised cost of getting from i to j. This assumes that the attractiveness of the opportunities is homogeneously distributed; note that the usual re-weighting by the number of opportunities in a given zone is no longer necessary since every opportunity is individually considered.

Equating the two equations leads to $d_{ij}^{\gamma} \sim e^{c_{ij}}$ or

$$c_{ij} = \gamma \ln(d_{ij}) + \text{const}, \tag{3}$$

that is, it seems that the perceived generalised cost scales logarithmically in the geographical distance. The model, at this point, makes no statement if this transformation is caused by human perception, or possibly by the properties of the transportation system itself (which becomes faster for longer distances).

5. Simulation

5.1. Introduction

For a variety of reasons, it may be interesting to have a complete model of all social contacts of a region or a country. Possible applications are the model-based generation of an important segment of leisure traffic (visits of friends or relatives) or the spreading of information/rumours. Behavioural models such as the ones by Hackney (Hackney and Marchal, 2009) or by Jackson (Currarini et al., 2007) are, in principle, able to generate such data. But in practice they are too slow for large scale scenarios, and statistical properties such as clustering coefficients are difficult to control since they emerge from the behaviour. Statistical modelling, i.e., to generate social networks from simple statistical principles, is meant to fill this gap (also see Park and Newman, 2004). The following is a first step in this direction, attempting to recreate the distance distribution from the survey.

5.2. Simulation Model

The model always takes an existing synthetic population as input. Two synthetic populations will be explicitly tested in the following:

- A population of 10,000 randomly distributed vertices over a square of 200 km^2.
- A 0.5% unbiased random sample of the population of Switzerland, resulting in 36,000 persons.

In both cases, the simulation connects each possible pair of vertices (i, j) with probability $p_{acc}(d_{ij})$. That is, every pair of vertices is touched exactly once.

The exponent of the power law in $p_{acc}(d_{ij})$ is set to $\gamma = -1.6$ as observed from empirical data. The acceptance probability is renormalised in a way that the mean degree of the network can be adjusted to $\langle k \rangle = 17$.

5.3. Random Vertex Distribution

As stated above, we first test the model with 10,000 randomly distributed vertices over a square of 200 km^2.

The network obtained from the simulation exhibits as expected a mean degree of $\langle k \rangle = 16.9 \approx 17$. Since edges are inserted independently from each other, the degree distribution shown in Fig. 3a) follows the usual Poisson distribution of random graphs (Newman, 2003). Figure 3b) shows the edge length distribution over uniform distance bins of 1 km size. The distribution follows a power law with an exponential cut-off towards the system boundaries. The slope parameter of the power law in the medium range up to 50 km is approximately -0.6: With randomly distributed vertices, the number of vertices $N_i(d)$ increases proportionally to distance d since the circumference of a circle grows linearly with its radius. Thus, one obtains $\gamma - 1$ for the slope of the observed edge length distribution. However, in a finite system $N_i(d) \propto d$ does not hold towards the system boundaries.

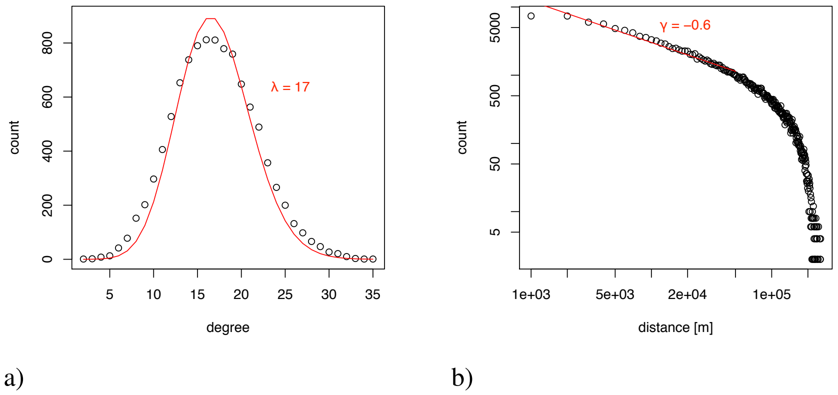

Fig. 3. Simulation with randomly distributed vertices. a) Degree distribution. The red curve indicates a Poisson distribution with $\lambda = 17$. b) Edge length distribution.

5.4. Synthetic Population of Switzerland

The starting point for this simulation is the synthetic population of Switzerland.

Figure 4a) shows the degree distribution of the simulated social network. The distribution is highly right-skewed and does not fit a Poisson distribution. Many real-world networks exhibit a power law degree distribution (Newman, 2003), and one can observe that the tail of the simulated distribution exhibits a power law in the range of 15 to 50 with exponent -2.1.

The simulated edge length distribution, again averaged over logarithmic distance bins, is shown in Fig. 4b). It fits quite well a power law with exponent -1.2 (red triangles) but does not exactly reproduce the shape of the empirical distribution (black circles). Figure 4c) shows the acceptance probability $p_{acc}(d)$ over logarithmic distance bins, which is determined for the simulation with the same method as for the empirical data in Sec. 4.2.. As expected, the simulation values scale with $d^{-1.6}$. The difference of the exponents of $n_i(d)$ and $p_{acc}(d)$ mean that the number of opportunities at distance d scales with approximately $N_i(d) \propto d^{0.4}$ in case of the population of Switzerland. One can validate this assumption by generating the fully connected graph for the synthetic population, i.e., by connecting all vertices. Figure 4d) shows the resulting edge length distribution of such a simulation over uniform distance bins of 1 km size. One observes that at least for the medium range up to 50 km the number of opportunities in fact scales with $d^{0.35}$. Since a high share of the population lives in Zurich and since Zurich is only 25 km distant from the southern border of Germany, one very quickly observes the boundary effects.

We can identify the distribution of the population to be the cause for two further effects. The first effect is visualised in Fig. 5a) and 5b). There is a correlation between the vertex degree and the population density: the higher the population density, the higher the degree. Generally speaking, simulated people living in urban areas tend to have a lot of social contacts, whereas people living in the country side have fewer contacts. Accordingly, the simulated network exhibits a high degree-degree correlation of 0.6 (calculated with the method of Newman (2002)), i.e., vertices of high degree tend to connect to other vertices of high degree. Furthermore, one also observes a correlation between edge length and population density (Fig. 5c). All short edges are concentrated in the highly populated areas, whereas long edges are concentrated in the countryside.

All of these results are plausible if one considers the methodology: Every possi-

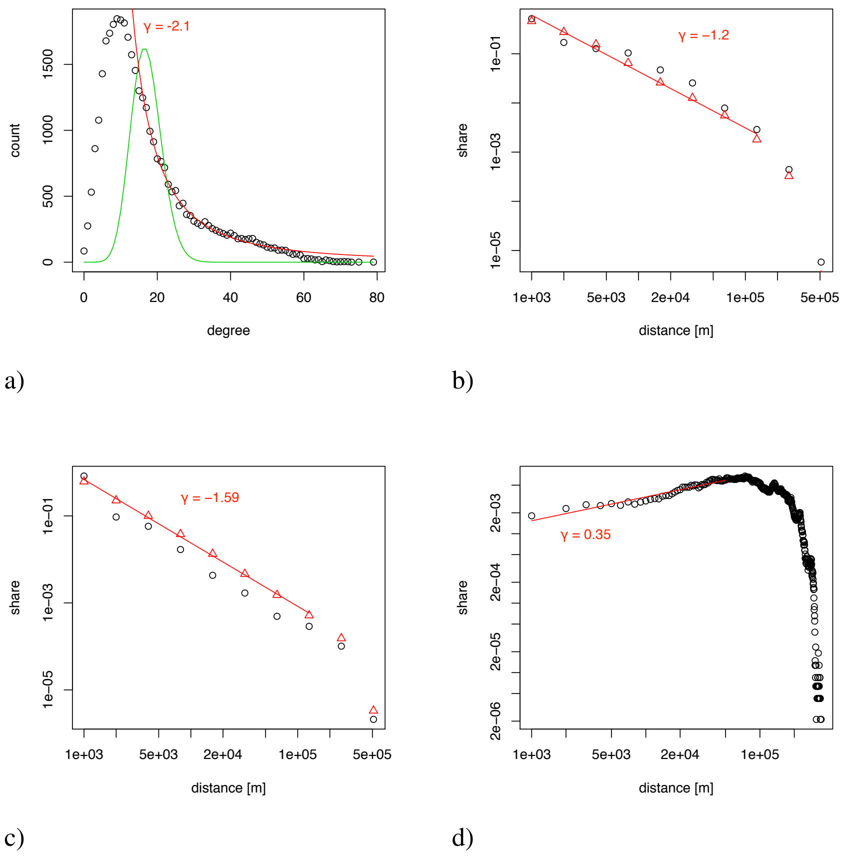

Fig. 4. Simulation with a synthetic population of Switzerland. a) Degree distribution. The green curve indicates a Poisson distribution with λ=17, the red curve indicates a power law distribution with exponent $\gamma = -2.1$. b) Edge length distribution. c) Acceptance probability $p_{acc}(d)$. Black circles are values from empirical data, red triangles are simulation values. d) Edge length distribution for the fully connected graph.

ble edge (i, j) is considered and accepted with probability $p_{acc}(d_{ij})$, which is larger for short distances. Since vertices in densely populated areas have many more opportunities in close distance, many more opportunities are accepted, resulting in larger degrees.

At the current stage of the survey, the observed data is to sparse to confirm or contradict our findings from the simulation. Figure 6a) shows that the degree over the population density randomly scatters, and similarly, the mean edge length over the population density (Fig. 6b) exhibits no clear pattern. Recall that this is an ongoing survey and that more data will be available in the future.

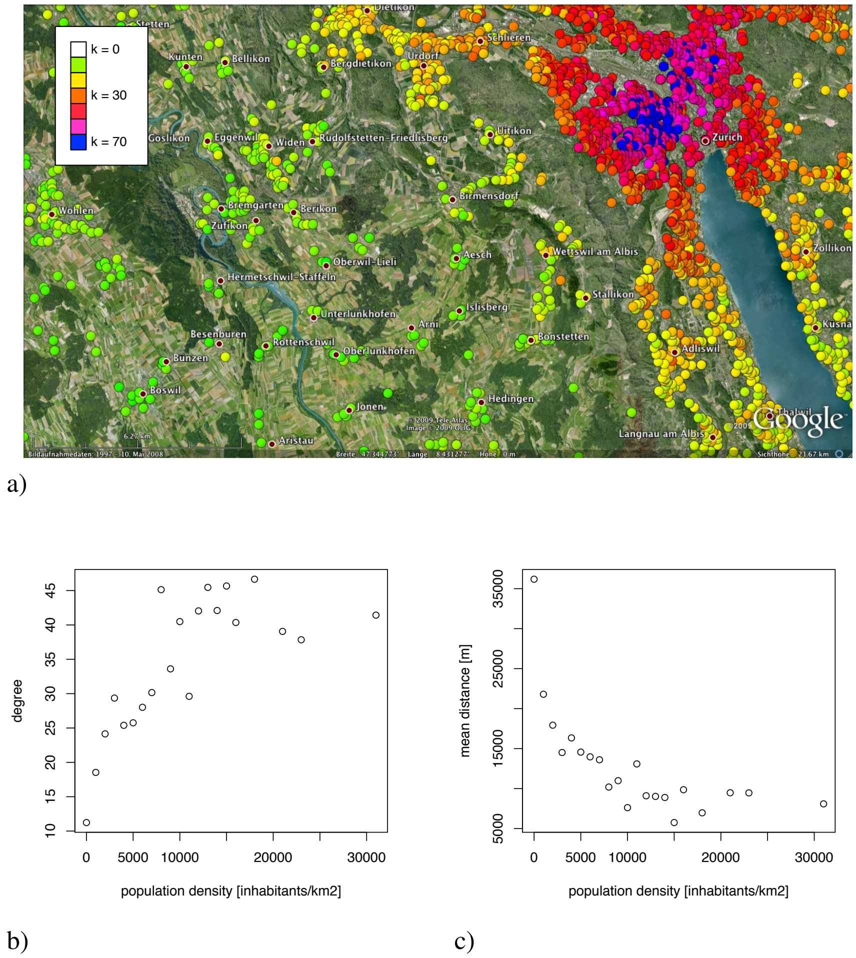

Fig. 5. Simulation with a synthetic population of Switzerland. a) Visualisation of the degree distribution. b) Degree over population density. c) Edge length over population density.

6. Discussion and Conclusion

This paper presents a stochastic model for spatially embedded social networks based on a spatial interaction model. From empirical data, we find that the probability to accept a leisure contact at a certain distance follows a power law with exponent -1.6. However, the resulting distance distribution of social connections is distorted by the underlying population distribution.

The paper shows that the observed scaling law can be explained with the traditional gravity model in that the probability of an edge can be modelled as a function of the distance between both actors. With a simulation on a synthetic population, we demonstrate that we can reproduce the observed edge length distribution with the proposed model.

The simulated network shows a right-skewed degree distribution and high degree-degree correlation, which are both typical properties for social networks. The model also produces a spatial distribution of degrees that is highly correlated with the population density. A the current stage of the survey, we cannot identify such a pattern in the empirical data because of its sparsity.

Turning to transportation research, the question arises if the beeline distance is the appropriate measure for the spatial distance between two individuals. Especially regarding the topology of Switzerland, one observes that the costs of visiting an acquaintance do not scale linearly with the beeline distance. For future studies, we intend to replace the beeline distance by travel time. Considering for example that a distance of 1 km is done in10 min by walking, 10 km in 30 min by public transport, and 100 km in 1 hour by car, this may indeed explain the logarithmic dependency of cost on distance observed in Sec. 4.3., and hence the currently observed power law as a function of distance would transforms into a exponential function of travel

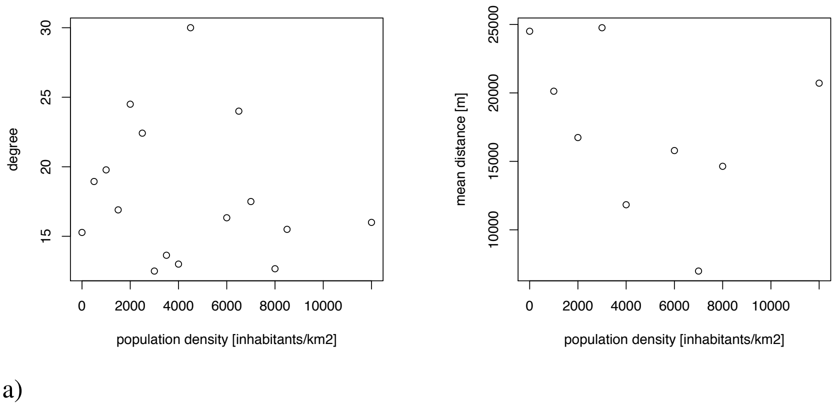

Fig. 6. a) Degree over population density from empirical data. b) Mean edge length over population density from empirical data.

time.

Regarding the findings of this study, one can conclude that distance does matter and explains certain characteristics of social networks, but it appears to be only one explanatory variable beside others.

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

References

ARE/BFS, 2007. Mobilität in der Schweiz, Ergebnissse des Mikrozensus 2005 zum Verkehrsverhalten. Federal Office for Spatial Development and Swiss Federal Statistical Office.

Brockmann, D., Theis, F., 2008. Money circulation, trackable items, and the emergence of universal human mobility patterns. IEEE Pervasive Computing 7 (4), 28–35.

Currarini, S., Jackson, M., Pin, P., 2007. An economic model of friendship: Homophily, minorities and segregation.

Frei, A., Axhausen, K. W., 2007. Size and structure of social network geographies. Tech. Rep. 439, ETH Zürich, Institute for Transport Planning and Systems.

Hackney, J., Marchal, F., 2009. A model for coupling multi-agent social interactions and traffic simulation. Paper presented at the 88th Annual Meeting of the Transport Research Board.

Kowald, M., Frei, A., Hackney, J., Illenberger, J., Axhausen, K. W., 2009a. Collecting data on leisure travel: The link between leisure acquaintances and social interactions. In: Applications of Social Network Analysis (ASNA).

Kowald, M., Frei, A., Hackney, J., Illenberger, J., Axhausen, K. W., 2009b. Using an ascending sampling strategy to survey connected egocentric networks: A field work report on phase one of the survey. Tech. Rep. 582, Institute for Transport Planning and Systems, Swiss Federal Institute of Technology, Zurich.

Larsen, J., Urry, J., Axhausen, K., 2006. Mobilities, Networks, Geographies. Ashgate, Aldershot.

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

Meister, K., Rieser, M., Ciari, F., Horni, A., Balmer, M., Axhausen, K., March 2008. Anwendung eines agentenbasierten Modells der Verkehrsnachfrage auf die Schweiz. In: Proceedings of Heureka '08. Stuttgart, Germany.

Newman, M. E. J., 2002. Assortative mixing in networks. Physical Review Letters 89 (20).

Newman, M. E. J., 2003. The structure and function of complex networks. SIAM Review 45 (2), 167–256.

Park, J., Newman, M. E. J., 2004. Statistical mechanics of networks. Physical Review E 70 (066117).

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

Wilson, A. G., 1971. A family of spatial interaction models, and associated developments. Environment and Planning 3 (1), 1–32.

Wong, L. H., Pattison, P., Robins, G., 2005. A spatial model for social networks. arXiv:physics/0505128v2.