# Insights into a spatially embedded social network from a large-scale snowball sample

Johannes Illenberger Matthias Kowald Kay W. Axhausen Kai Nagel

March 28, 2011

Working Paper 10-10

## Abstract

Much research has been conducted to obtain insights into the basic laws governing human travel behaviour. While the traditional travel survey has been for a long time the main source of travel data, recent approaches to use GPS data, mobile phone data or the circulation of bank notes as a proxy for human travel behaviour are promising. The present study proposes a further source of such proxy-data: the social network. We collect data using an innovative snowball sampling technique to obtain details on the structure of a leisure-contacts network. We analyse the network with respect to its topology, the individuals' characteristics, and its spatial structure. We further show that a multiplication of the functions describing the spatial distribution of leisure contacts and the frequency of physical contacts results in a trip distribution that is consistent with data from the Swiss travel survey.

## 1 Introduction

Understanding human travel behaviour is a major research field in the community of transport science. The understanding of human mobility is not only required for urban planning and traffic forecasting, but also to gain insights into the dynamics of the spreading of diseases, information, or social values and norms. Tools to monitor human travel behaviour range from the traditional travel survey, such as the German "MiD" [1] or the Swiss micro-census [2], via long-term travel diaries [3], to automated processing of GPS data [4], mobile phone data [5] or tracking of dollar bills [6]. The latter two approaches provide much greater sample sizes compared to the traditional travel survey, with relatively less effort. However, they do not directly monitor the travel behaviour, but use other data as a proxy. The circulation of bank notes describes a diffusion process rather than repeated movement patterns. Such data may by quite useful for the modelling of, say, virus spreading, yet the inference of individual travel behaviour is limited since the link between the circulating bill and the travelling individual is missing. For instance, it is unclear when a bank note was passed from one person to another, or even if the displacement is caused by the transfer of bank notes between institutions. Mobile phone data, in contrast, allows to obtain individual trajectories. Still, characteristics of travellers and their motives remain unrevealed. All of these monitoring tools, including the traditional ones, have in common that they lack any information about the social connectivity of individuals. Table 1 provides an overview of the strengths and weaknesses of selected monitoring tools.

With the increasing availability of large data sets, such as co-authorship networks [8], movie actor networks [9], or email networks [10], social network analysis has made great advances. Yet, the link from social networks to travel behaviour is missing as those networks provide no information about their spatial structure. In sociology, it has been pointed out by Latané [11] that distance is a significant explanatory

Table 1: Overview of selected monitoring tools (adopted from [7]).

| aspect | paper & pencil [1, 2] | mobile phone data [5] | person-based GPS [4] | dollar bills [6] | present study |
|---|---|---|---|---|---|
| delimitation of journeys | easy | impossible | difficult | no | only leisure |
| coverage of journeys | depending on respondent | incomplete | complete | no | depending on respondent |
| duration of journeys | rounded (usually 5 mins) | no | exact | no | possible (with data from travel diary) |
| identification of locations | precise | zones | exact | no | precise (depending on respondent) |
| trip purpose | yes | no | imputed | no | yes (leisure) |
| social content | in part | no | no | no | rich |

variable for the structure of social networks, but only recently research in sociology has focused on this aspect [12, 13, 14, 15].

In fact, both communication and face-to-face encounters are crucial for the maintenance of social networks [16]. In consequence, there is a reciprocal interaction between travel and communication on the one hand, and social networks on the other hand. Social networks cause travel and communication, yet travel and communication enable the spatial spread of the social network.

To shed more light on the link between network structures and travel behaviour, research in transport planning makes increasing use of the methods of social network analysis. By focusing on actors and their relations simultaneously, these methods prove productive, producing new empirical insights and results (for examples see [16, 17, 18, 19, 20, 21]). A common characteristic of those studies is their limitation to first degree relations. This means that the obtained network data is limited to isolated star-like network structures.

The present study widens the scope of network studies in transport planning by collecting data on iteratively connected personal leisure networks in Switzerland with a so-called *snowball sampling* approach. This approach allows to make statements about the network structure beyond first degree relations, such as transitivity and average path length between individuals. At the same time, travel-related attributes are collected, most importantly the geographical locations of the persons' home locations, but also the frequencies of face-to-face meetings or demographic attributes. We choose the context of leisure contacts, as leisure traffic is an increasing share of the travel patterns. For example in Germany, "leisure" is now the main trip purpose (31 % of all trips) [1], followed by "shopping" (19 % of all trips). "Work" follows third, with 15 % of all trips. Similar behaviour is observed in Switzerland [2], in the U.K. [22], and in the U.S. [23]. Yet, compared to commuting traffic which is routinely surveyed in most countries, there is comparatively little data for leisure traffic. The present study is not only meant to generate data for that travel segment, but also to enhance our understanding.

The remainder of this article is organised as follows. Section 2 presents a brief overview of related work, describes the survey instrument, the construction of the leisure network from the raw survey data and presents a technique to correct the inherent "degree bias" of the snowball approach. Section 3 presents an analysis of the network with respect to its topology, spatial structure and socio-demographic attributes of the individuals. Furthermore, travel patterns are inferred from the spatial structure of the network. The article is closed with a discussion in Sec. 4 and a summary in Sec. 5.

# 2 Data collection and parameter estimation

## 2.1 Related work

In transport research a number of studies using the methods of social network analysis have been recently conducted. To provide a brief overview, the studies that are closely related to the present study are presented.

Carrasco [18] collected data of 350 participants in the East York area of Toronto with a so-called *egocentric* sampling design. This means that respondents, denoted as *egos*, are asked about their social contacts, denoted as *alters*. The research question of his project was to investigate the explanatory power of data describing personal interaction on the generation of joint activities.

The egocentric network approach has also been used by Frei and Axhausen [20] to sample personal networks of more than 300 respondents. The study focused on the individual mobility biographies and the geographical distribution of emotionally important contacts.

Silvis et al. [19] used a different approach to collect data on trips and social interactions. In a three-day interaction diary respondents were asked to report trips and social interactions together with information about their purpose, mode, and participants. Furthermore, respondents were requested to pass postcards to people with whom they met face-to-face inviting them to participate in the survey as well. This sampling design is related to the present design in that respondents invite other people to participate the survey. Within a two-month timeframe 24 participants were recruited reporting 505 trips and 972 social interactions.

The interaction diary approach has also been recently used by van den Berg et al. [21] but without the recruitment mechanism of Silvis et al. Respondents were requested to report up to three participants of joint activities conducted on two consecutive days. The sample size counts nearly 750 diaries.

The present study differs from the above studies in three major aspects:

- The sampling technique collects data on *connected* personal networks. This allows us to make statements about network characteristics that go beyond vertex-local properties.
- The study explicitly focuses on leisure contacts. Emotionally important contacts, the main interest of the above studies, are only subordinated.
- The targeted sample size (800 egos reporting 12'000 alters) considerably exceeds the sample size of the above studies.

## 2.2 Survey instrument

The data used for this article describes a social network of leisure contacts in Switzerland. It is obtained with a snowball sampling design. In a snowball sample, respondents (egos) are asked to report their social contacts (alters), which are then invited to participate in the survey as well. The new respondents are asked to report their social contacts which in turn are invited as well. This iterative process is continued until a predefined number of iterations is conducted or the desired number of respondents are sampled. The name of the approach stems from the image of a snowball accumulating more and more material when it is rolled through the snow. Each respondent is requested to fill out a questionnaire which is divided into four sections. The following paragraphs summarise the survey instrument, details can be found in [24, 25].

An introductory questionnaire asks for the ego's socio-demographics and its mobility biography by collecting postal addresses of former residential locations as well as workplaces and schools.

The second part asks the ego to report its alters. It implements the so-called *name generator*, i.e. the definition of the criteria conforming to which alters should be reported or not. The name generator comprises two questions:

1. "Please list the people with whom you make plans to spend free time (Examples: sports, club or organised activities, cultural events, cooking together or going out to eat, taking holidays or excursions together)."

2. "If there are other people with whom you discuss important problems, please list them here."

Naturally, it is up to the respondent to decide if a social contact meets the above criteria. Both kinds of contacts can be considered as crucial in terms of leisure travel as the ego meets those persons frequently. The questionnaire has space for 40 contacts. Respondents are allowed to additionally name (note them on the back) further contacts if they feel so. Since this opportunity is only rarely used, the current analysis neglects the additional contacts.

The so-called *name interpreter* represents the third section of the questionnaire. While the name generator asks only for the alters' "names", the name interpreter requests the respondents to provide further information. Of interest are (i) the alters' socio-demographics, (ii) the type of relationship between ego and alter, (iii) contact modes used, and (iv) the contact frequency.

Finally, the last part of the questionnaire, a so-called *sociogram*, asks egos to report groups of alters that make plans to spend free time together. This part of the survey instrument is influenced by the work of Carrasco et al. [18] in Toronto (see also [26]). Activity groups can be reported by mentioning the context of the activity, for example "hiking group", and identifying all alters from the name generator that join in this activity. The sociogram does not allow to mention further contacts that have not been reported in the name generator. Egos are allowed to mention up to 20 groups.

The range of topics and the level of confidentiality implied in the questions result in a high amount of response burden. Respondents are offered a 20 CHF incentive.

A subset of respondents in addition participate in a 8-day travel diary. The diary records daily activities together with the information with whom these activities are conducted, how frequently they are conducted and who initiated the activity. The travel diary is still in an early state. For that reason, its analysis is not included in this article. However, it is expected that the diary enriches the current data set.

## 2.3 Application and issues of snowball sampling

A part of the snowball survey is still under way. The survey was started with 40 ego-seeds. Two seeds did not disclose their contacts and are excluded in further analysis For 20 seeds the snowball is expanded up to the second iteration. The remaining 18 seeds are expanded up to the fourth iteration. At the time of the writing of this article the survey is in the process of expanding the third iteration (date of data: September 2010).

Several issues have to be considered because snowball sampling is well known for various sources of bias [27, 28]. First, snowball sampling does not produce unbiased random samples because the probability of becoming part of the sample is influenced by the egos, as egos report their alters selectively, whether unintentionally or intentionally (for a study with such problems see [19]). This selectivity limits the number of possible paths the chain can take to be continued.

The second source of bias results from the snowball chain itself as it provides a higher chance for persons with many social contacts to be named by someone else and thus be included in the sample. We address this source of bias, also called "degree bias", in Sec. 2.5.

Third, bias can result from similarities between egos and their alters. These characteristic similarities, also addressed as status homophily, are well

documented in social network studies [29, 30]. The present study aims to observe homophily in connected personal networks, as persons with similar characteristics have a higher probability of establishing a relationship than dissimilar pairs of persons [31].

The overall response rate is 27 % (calculated conforming to the guidelines of the AAPOR [32]), but differs between the snowball iterations: Starting with a low rate around 16 % in the 0-th iteration, it increased to 31 % on iteration 1, 29 % on iteration 2, and around 23 % at iteration 3 (which is not completed by the time of the writing of this paper). This is satisfactory considering the amount of response burden of this study: Filling out the questionnaire requires, depending on the egos' network size, between one and four hours. It is also satisfactory in view of the fact that the survey asks for very confidential information such as names and postal addresses of friends and family members (for arrangements to increase the response rate see [24]). When using the instrument's response burden as input, a tool from commercial survey research estimated a lower response rate [33].

To our knowledge, this is the first time a snowball sample approach is used to sample a survey population of this size (targeted are 800 egos reporting about 12'000 alters) with so few restrictions on the persons included. Of course, language or national frontiers affect the spread of the snowball (see below). However, the survey instrument makes no limitations regarding institutional settings (such as workplace, school or clubs), personal characteristics, or communication modes.

## 2.4 Constructing the snowball graph

For the following analysis, the raw survey data is transformed into a graph data structure. Vertices represent egos and alters; edges represent the reported leisure contacts. Even if, strictly speaking, the survey data represents directed edges (from ego to alter) the graph is assumed to be undirected.[1]

The raw survey data comes in the form of an edge list, i.e. a listing of all reported leisure contacts. All vertices are assigned an id and are checked for equality. This means that if multiple egos report the same surname, the identity of an alter is first verified based on name and address. However, the residential locations of approximately 25 % of alters are missing because the reporting egos did not disclose their addresses. In such situations, further attributes such as age, civil status, and citizenship are used to identify an alter. In critical cases, the respondents are contacted for clarification. Still, the determination of uniqueness remains a critical issue.

The data of the sociogram (Sec. 2.2) is ignored in the following analysis with the exception of Sec. 3.1.3; that section addresses the changes in network transitivity if edges from the sociogram are included.

In the remainder of this article, the following notations will be used. The index of an iteration is denoted by $i$, where the 0-th iteration represents the initial random draw of the seed vertices. Vertices that represent an ego are called *ego-vertices*. Vertices representing an alter are called *alter-vertices*. Ego-vertices are those who participated in the survey, i.e. vertices that have filled out a questionnaire. Regarding the statistical analysis, this distinction is crucial since, for example, the degree of an alter is unknown.

Some alter attributes are initially reported by the ego. The reliability of the information of such indirect alter information is tested, where possible, by comparing it with the details that the alter reports herself in the subsequent iteration. The comparison shows that the information matches in more than 90 % of all cases which is consistent with the findings of other studies [34].

Quantities that are calculated based on different iterations are denoted with the iteration index in parentheses in the superscript. For instance, the number of ego-vertices sampled in iteration $i$ is denoted with $n^{(i)}$, the number of vertices that have been sampled up to and including iteration $i$ is denoted by $n^{(\leq i)}$. Symbols without an iteration index refer to

[1] Consider the situation where the alter participates in the survey but does not report the back-link to the ego. It is now unknown if the back-link is intentionally missing in the sense of "this in not my friend", intentionally missing in the sense of "I know that our friendship has already been reported, so I do not report it again", or forgotten.

Table 2: Graph size per iteration. Note that "response rate" and "recruitment rate" are not the same (see text). *By the time of the writing of this article iteration 3 has not been fully expanded.

| iteration ($i$) | egos ($n^{(i)}$) | alters ($a^{(i)}$) | edges ($m^{(i)}$) | recruitment rate ($\alpha^{(i)}$) |
|---|---|---|---|---|
| 0 | 38 | 568 | 573 | – |
| 1 | 103 | 1644 | 1794 | 0.18 |
| 2 | 238 | 4464 | 4812 | 0.14 |
| 3* | 27 | 451 | 496 | 0.006 |
| total | 406 | 7127 | 7675 | 0.06 |

the entire sample obtained up to (and including) iteration 3. For instance, $n$ corresponds to $n^{(\leq 3)}$, where this includes both, vertices of the components that emerge from the seeds that are expanded only up to iteration 2 and vertices of the components that are expanded up to iteration 4.

Based on the vertices' ids, the edge list can be merged into one graph. The resulting graph consists of 7165 vertices and 7675 edges, with 406 vertices representing egos. An overview of the graph size per iteration is given in Tab. 2. The recruitment rate per iteration $\alpha^{(i)}$ is defined as the number of egos $n^{(i)}$ over the number of alters $a^{(i-1)}$ in the previous iteration. Note that the definition of *recruitment rate* is different to what is commonly referred to *response rate* in sociology [32]. To calculate the response rate one would use the total number of all *enquired* vertices in the denominator. Due to missing contact information not all alters can be enquired.

## 2.5 Estimation

As mentioned in Sec. 2.3, there are several sources for bias in a snowball sample. To properly estimate topological characteristics of the network, an approach to correct the degree bias is used. The problem of estimating properties of snowball sampled networks is well known and has been addressed by other articles [35, 36, 37, 38, 39, 40, 41, 42, 43]. However, since snowball sampling can be implemented in different variants, each specification requires its own estimation method. A major aspect in which snowball sampling designs differ is the so-called *branching rule*. This rule defines the process of recruiting new alters. For instance, in Respondent-Driven Sampling [36], a common real-world application of snowball sampling, each ego recruits a constant number of alters. In contrast, in the present snowball sampling design it is attempted to recruit all reported alters. Both branching rules lead to quite different estimation techniques.

This section will summarise the ideas of our estimation method. For details see [44].

The progress of a snowball sampling is heavily determined by the topology of the underlying network. Well connected vertices are covered relatively fast by the sampling algorithm, whereas it takes more iterations until the less well connected vertices are reached. As a consequence, vertices with high degree are overrepresented in the early iterations. Even though this effect is undesired for network parameter estimation, it interestingly is of advantage for, say, immunisation strategies: Randomly select a person to immunise, but also immunise her friends since it is likely that one reaches persons with higher connectivity than average [45].

In terms of estimation theory, snowball sampling can be regarded as sampling with unequal inclusion probabilities. The inclusion probability $\pi_v$ of a vertex $v$ cannot be calculated directly. Yet, it can be estimated by the following considerations.

First, expand the notation of $\pi_v$ to account for the iteration index. Thus, denote by $\pi_v^{(\leq i)}$ the probability that vertex $v$ is included in a snowball sample that has been run up to and including the $i$-th iteration. This probability can be expressed as the probability that one of the vertex's neighbours has been sampled in or before the previous iteration $i-1$. More formally, the probability that vertex $v$ is *not* sampled in or before iteration $i$ is the joint probability that none of its neighbours $w$ has been sampled in or before iteration $i-1$:

$$\pi_v^{(\leq i)} = 1 - \prod_{w=1}^{k} \left(1 - \pi_w^{(<i)}\right) \ , \qquad (1)$$

where $k$ denotes the degree of vertex $v$. The probability $\pi_w^{(<i)}$ is, however, just as unknown as $\pi_v^{(\leq i)}$.

An approximation that turns out to be useful is to ignore the details of the snowball sampling process and assume that all neighbours are equally and independently sampled. Then, the inclusion probability of a neighbour can be approximated by

$$\pi_w^{(<i)} \approx \frac{n^{(<i)}}{N} \tag{2}$$

where $N$ is the total number of vertices. Replacing $\pi_v^{(\leq i)}$ with the estimator $\hat{\pi}_v^{(\leq i)}$ Eq. 1 becomes

$$\hat{\pi}_v^{(\leq i)} = 1 - \prod_{w=1}^{k} \left(1 - \frac{n^{(<i)}}{N}\right) \ . \tag{3}$$

Since now the product does not depend on the running index any more, it can be rewritten as

$$\hat{\pi}_k^{(\leq i)} = 1 - \left(1 - \frac{n^{(<i)}}{N}\right)^k \ . \tag{4}$$

Obviously, this estimator is only applicable for $i > 0$. In the 0-th iteration samples are obtained with an unbiased random draw, i.e. $\pi_v^{(0)} = n^{(0)}/N$. Eq. 4 implicitly accounts for the recruitment rate as $n^{(<i)}$ decreases with decreasing recruitment rate.

The estimator requires knowledge of the size $N$ of the network, which is unknown. However, the growth of the snowballs shows that both the national border (e.g. between Switzerland and Germany) as well as the language boundary (between the German-speaking part of Switzerland and those with other languages) restrict the expansion. In consequence, it is plausible to set $N$ to the size of the German-speaking Swiss population, approximately 5.2 million inhabitants.

Given the estimator for the inclusion probability, $\hat{\pi}_k^{(\leq i)}$, one can obtain additional statistical quantities. An estimator for any population total is

$$\hat{t}_y = \sum_v \frac{y_v}{\hat{\pi}_v} \ , \tag{5}$$

where $y$ is the quantity of interest and $\sum_v$ denotes the sum over all ego-vertices. An estimator for any population mean is the weighted sample mean

$$\hat{\bar{y}} = \frac{1}{\sum_v 1/\hat{\pi}_v} \sum_v \frac{y_v}{\hat{\pi}_v} \ . \tag{6}$$

In the following section, Eq. 6 will be used to obtain estimators for the mean degree, degree distribution, and the mean clustering coefficient.

# 3 Analysis

## 3.1 Topological network properties

### 3.1.1 Degree

According to Eq. 6, an estimator for the mean degree that corrects the bias of the snowball is

$$\hat{\bar{k}} = \frac{1}{\sum_v 1/\hat{\pi}_v} \sum_v \frac{k_v}{\hat{\pi}_v} \ , \tag{7}$$

where $k_v$ is the degree of vertex $v$. The sum goes over all ego-vertices because no information about the degree of the alters is available. The estimated degree distribution is obtained by

$$\hat{p}\,(k) = \frac{1}{\sum_v 1/\hat{\pi}_v} \sum_{v_k} \frac{1}{\hat{\pi}_{v_k}} \ , \tag{8}$$

where $\sum_{v_k}$ denotes the sum over all ego-vertices with degree $k$.

Calculating the mean degree, without correction, for each iteration reveals the snowball bias. The uncorrected mean degrees for iterations 0 to 3 are $\bar{k}^{(0)} = 15$, $\bar{k}^{(\leq 1)} = 17.6$, $\bar{k}^{(\leq 2)} = 20.1$, and $\bar{k}^{(\leq 3)} = 20.0$. Using the estimator from Eq. 7, one obtains a corrected mean degree of $\hat{\bar{k}}^{(\leq 3)} = 13.2$. The fact that this is only slightly less than the mean degree $\bar{k}^{(0)}$ of the initial draw, which is by definition unbiased, is an indicator for the validity of the correction method.

Figure 1(a) shows the uncorrected (circles) and the corrected (triangles) degree distribution. Both distributions are right-skewed, and it is clearly visible that the estimator shifts probability mass from the high degrees to the low degrees. Different from other studies [9, 10, 46], the tail of the (corrected) distribution seems to follow an exponential (Fig. 1(b)) rather than a power law decay. Recall that because of the survey design reported contacts beyond the first 40 are currently ignored. Egos can still have a higher degree than 40 if they named 40 contacts and then are additionally named by other egos. If some respondents

with $k = 40$ had in truth a higher degree $k > 40$, this would shift probability mass to degrees above 40.

### 3.1.2 Degree correlation

Another interesting property is the degree correlation which can be expressed as the Pearson correlation coefficient of the degrees of the vertices on either ends of all edges [47]:

$$r_k = \frac{\sum\limits_{(vw)} k_v k_w - M^{-1}\left(\sum\limits_{(vw)} \frac{1}{2}\left(k_v + k_w\right)\right)^2}{\sum\limits_{(vw)} \frac{1}{2}\left(k_v^2 + k_w^2\right) - M^{-1}\left(\sum\limits_{(vw)} \frac{1}{2}\left(k_v + k_w\right)\right)^2} \ , \tag{9}$$

where $k_v$ and $k_w$ denote the degrees of the two adjacent vertices $v$ and $w$ of an edge $(vw)$ and $M$ denotes the total number of edges in the network. To properly determine the degree correlation for a sampled network we evaluate Eq. 9 only for a sub-network: $\sum_{(vw)}$ goes only over all edges connecting ego-vertices and $M$ is set to their number.

Social networks are known to exhibit a positive degree correlation indicating that vertices of similar degrees tend to be connected [48]. The nature of the snowball sampling technique also biases the degree correlation. The uncorrected degree correlation is close to zero ($r_k = 0.07$). In contrast to vertex-oriented properties, such as the degree, the item of interest is now an edge. The inclusion probability $\pi_{(vw)}$ of an edge follows from the observation that the probability that an edge is sampled before or in iteration $i$ equals the probability that at least one of its adjacent vertices $v$ or $w$ is sampled before or in iteration $i - 1$. Hence,

$$\hat{\pi}_{(vw)}^{(\leq i)} = \left(\hat{\pi}_v^{(<i)} + \hat{\pi}_w^{(<i)}\right) - \left(\hat{\pi}_v^{(<i)} \hat{\pi}_w^{(<i)}\right) \ , \tag{10}$$

where again independence of the sampling events is assumed. An estimator of the degree correlation is obtained by weighting each summand (in analogy to Eq. 6) in Eq. 9 with the inverse inclusion probability

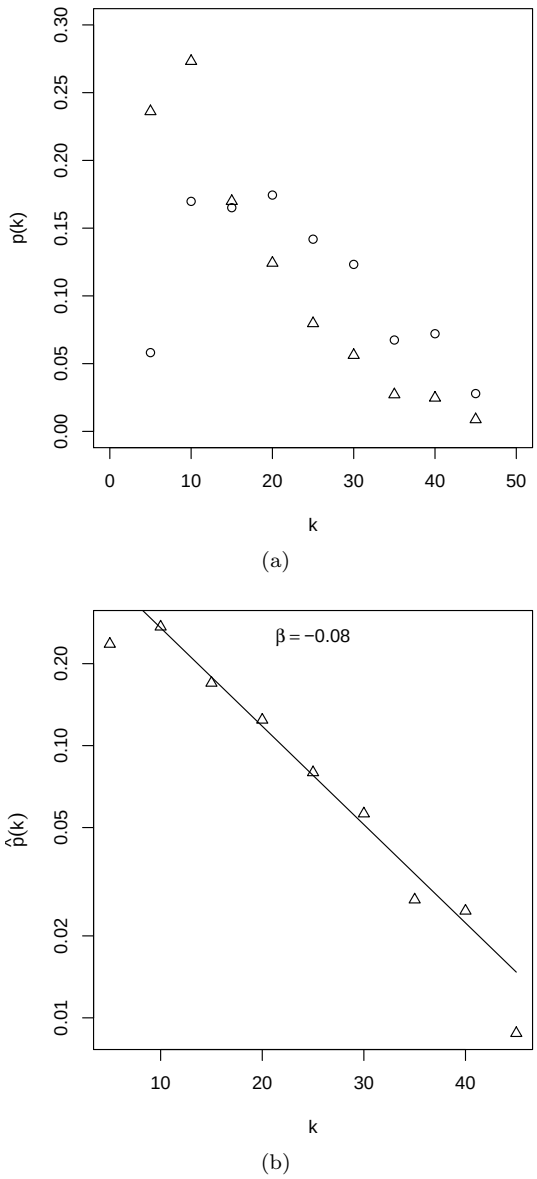



Figure 1: (a) Uncorrected ($p\left(k\right)$, circles) and corrected ($\hat{p}\left(k\right)$, triangles) degree distribution. Results are aggregated in bins of width 5. (b) Log-linear plot of the corrected degree distribution. The solid line indicates a fit of the function $p\left(k\right) \sim \exp\left(\beta k\right)$, resulting in $\beta = -0.08$.

$\hat{\pi}_{(vw)}$ of an edge:

$$\hat{r}_k = \frac{\sum\limits_{(vw)} \frac{k_v k_w}{\hat{\pi}_{(vw)}} - \hat{M}^{-1} \left( \sum\limits_{(vw)} \frac{1}{2} \frac{(k_v + k_w)}{\hat{\pi}_{(vw)}} \right)^2}{\sum\limits_{(vw)} \frac{1}{2} \frac{(k_v^2 + k_w^2,)}{\hat{\pi}_{(vw)}} - \hat{M}^{-1} \left( \sum\limits_{(vw)} \frac{1}{2} \frac{(k_v + k_w)}{\hat{\pi}_{(vw)}} \right)^2} , \quad (11)$$

where $\hat{M}$ is an estimator for the unknown total number of edges:

$$\hat{M} = \sum_{(vw)} \frac{1}{\hat{\pi}_{(vw)}} . \quad (12)$$

The sum $\sum_{(vw)}$ goes, again, only over edges connecting ego-vertices. Applied to the survey data, Eq. 11 results in $\hat{r}_k = 0.16$. This means that the present sampled network is slightly assortative with respect to degree, however, not as pronounced as in, for instance, networks of movie actors ($r_k = 0.21$ [9]) or company directors ($r_k = 0.28$ [49]).

### 3.1.3 Transitivity

Social networks are often an example for complex networks with high transitivity, i.e. a lot of triangular configurations [8, 9, 49]. There are two common methods to measure transitivity in a network. Transitivity can be quantified with the network clustering coefficient

$$C_{(1)} = \frac{3 \cdot n\,(\text{triangles})}{n\,(\text{connected triples})} , \quad (13)$$

or by the mean clustering coefficient over all (ego-)vertices

$$C_{(2)} = \frac{1}{n} \sum_v \frac{2m_v}{k\,(k-1)} , \quad (14)$$

where $m_v$ denotes the number edges connecting neighbours of vertex $v$, and $n$ the number of ego-vertices. The first equation (Eq. 13) represents a global definition: It puts the total number of triangles in relation to the total number of triples. The second equation (Eq. 14) is an average over a local definition: It puts the number of triangles connected to a vertex in relation to the triples centred at the same vertex and then averages over all vertices. The sampled network exhibits no significant transitivity, neither quantified by the global definition $C_{(1)} = 0.015$ nor by the local definition $C_{(2)} = 0.02$.

Considering the correction technique, it proves to be better applicable to $C_{(2)}$ than to $C_{(1)}$: Since $C_{(2)}$ is a vertex-local property, the value $\hat{C}_{(2)}$ can be estimated according to the weighted sample mean:

$$\hat{C}_{(2)} = \frac{1}{\sum_v 1/\hat{\pi}_v} \sum_v \frac{2m_v}{k\,(k-1)} \cdot \frac{1}{\hat{\pi}_v} . \quad (15)$$

When estimating this quantity, only ego-vertices that have been sampled strictly before the last iteration are considered in the sum of Eq. 15 because only then all information about the connections between their alters is available from the survey. Yet, even after the correction by the estimation technique, the mean clustering coefficient $\hat{C}_{(2)}$ remains at 0.06.

Non-recruited alters reduce the value of the clustering coefficient, since without recruitment of at least one alter, an alter-alter link cannot be detected. Generally, the missing alter-alter relations are a drawback of snowball sampling. An attempt to shed more light on these relations is made in capturing information about cliques. A clique is defined as a fully connected set of vertices and is obtained from the sociogram data (Sec. 2.2). Respondents are asked to define activity-groups (for instance "hiking group" or "soccer club") and assign their alters to those groups. Connecting all alters within an activity-group with each other results in a clique. One may argue that alters within an activity-group are not necessarily connected to each other: Especially for large groups, the probability of being connected is likely to decrease. However, half of all reported cliques contain less than 4 persons. This is consistent with the findings of Dunbar [50] that groups of core contacts are rarely larger then four persons. Given the clique information, the number $m_v$ of observed edges between alter-vertices increases significantly, and the mean clustering coefficient changes to $C_{(2)} = 0.21$. The network clustering increases considerably to $C_{(1)} = 0.54$.

This discrepancy confirms, once more, that both definitions of transitivity can lead to quite different results [48]. Figure 2 shows the size of cliques as well

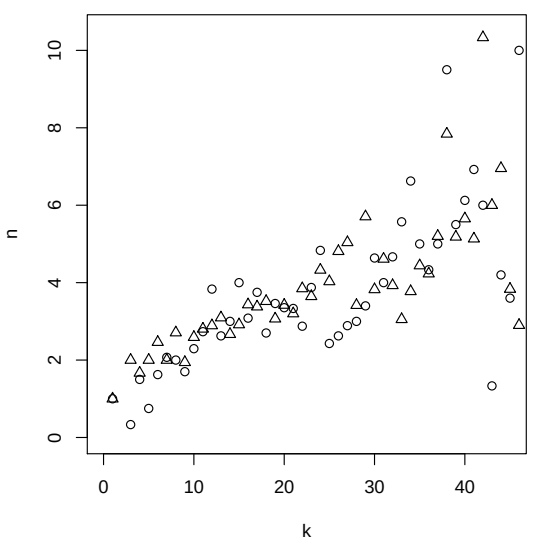



Figure 2: Average number of cliques (circles) and size (triangles) of cliques over the degree of the reporting respondent.

as the average number of cliques in relation to the degree of the reporting ego. Both quantities increase approximately linearly with the degree. This means that the few high degree respondents contribute a lot of cliques containing many persons and consequently contribute many triangles, thus increasing the network clustering $C_{(1)}$. In contrast, the mean vertex clustering coefficient $C_{(2)}$ assigns the same weight to the high degree vertices (contributing many triangles) as well as to the low degree vertices (contributing few triangles) and thus results in a considerably smaller value.

#### 3.1.4 Components

Starting with 38 seed-vertices, the snowball graph starts with 38 isolated components. A component denotes a graph where every vertex can be reached by every other vertex. Already within the first iteration, the survey detects bridge-persons that connect components. Four seed-to-seed connections, i.e. pairs of seed vertices connected by a path, can be identified (see bold edges in Fig. 3). Three of these paths have a length of four edges. The fourth path is an indirect path composed of two of the above paths and thus connecting two seeds through the original component of a third seed.

Within the second iteration, the survey graph merges to 27 isolated components. A total of 40 seed-to-seed connections can be identified, where 16 connections are direct and 24 connections are indirect, going through other components. The longest seed-to-seed connecting path consist of 19 edges, where the connecting path is defined as the path with the least number of edges which connects the two seed-vertices. Within the third iteration no additional bridge-vertices are found. The average length of paths connecting seed-vertices is 9.9 edges. It may be too early for statements about small world properties. Assuming that there should be a path form each seed-vertex to each other seed-vertex, only 40 of 703 ($n \cdot (n-1)/2 = 38 \cdot 37/2 = 703$) possible seed-to-seed paths are present in the sample and detected (approximately 7 %).

Figure 3 visualises the sampled network at the time of the writing of this article. A giant component is identified with 4096 vertices. It is a composition of components that originally emerged from nine of the seed-vertices. A further component of 693 vertices containing three seed-vertices and a component of 389 vertices containing two seed-vertices are identified. All remaining components are still the isolated egocentric networks emerging from the seed-vertex.

### 3.2 Spatial properties

Given the residential locations of more than 75 % of all reported egos and alters, the edge length distribution is calculated (Fig. 4(a)). The distribution appears to break up into a short range domain up to approximately 20 km and a long range domain including transcontinental contacts up to 16.000 km distance. Both domains follow a power law distribution $p_{edge}(d) \sim d^{\beta_{1/2}}$ with $\beta_1 \approx -0.5$ for the short range domain and $\beta_2 \approx -2.1$ for the long range domain. Half of all connected individuals are located within a distance of 11 km to each other. Note that the spatial analysis does not require a correction for

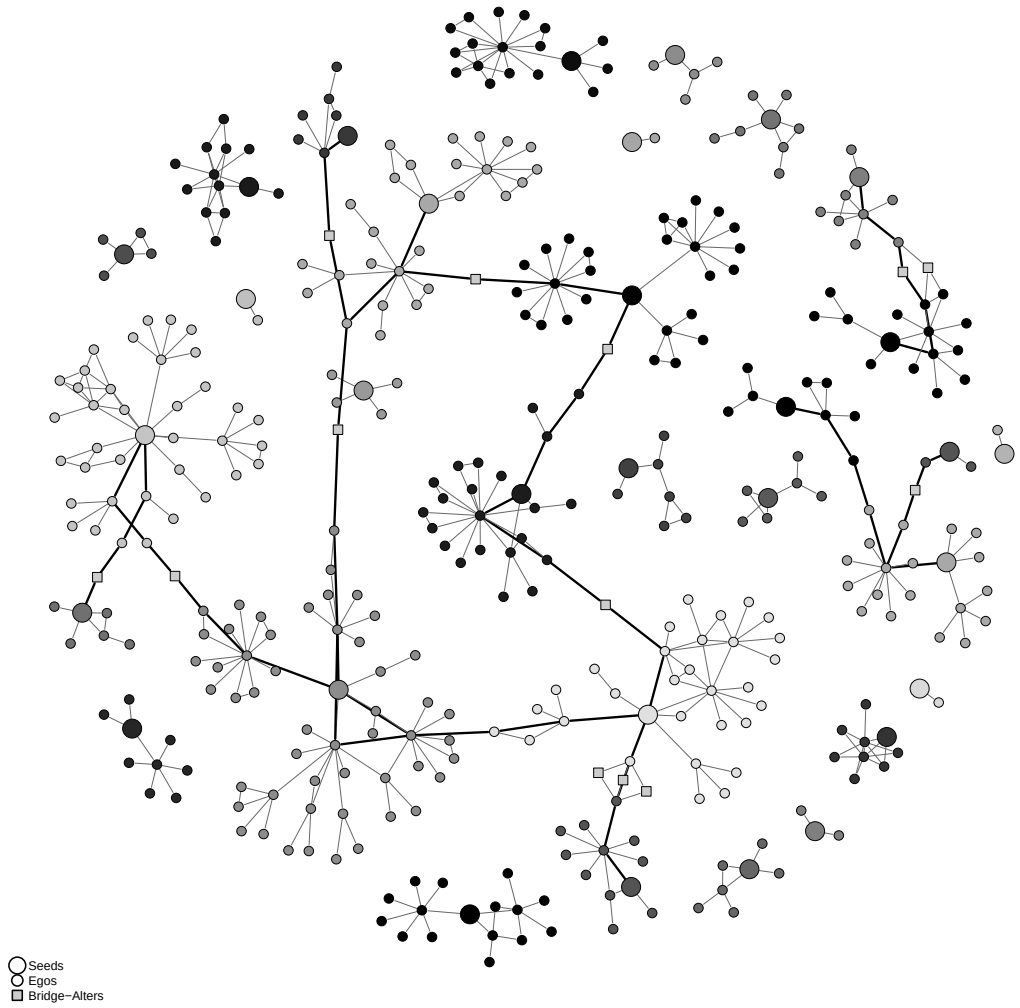



Figure 3: Sampled leisure network. Drawn are only ego-vertices, and alter-vertices if they act as a bridge-vertex (gray squares) connecting two components. Paths connecting seed-vertices are highlighted with bold edges.

Table 3: seed-vertex connection matrix. Row/column names are the ids of the seed-vertices. The entries represent the path length between both seed-vertices. Listed are only those seed-vertices that are reachable by at least one other seed-vertex.

| id | 709 | 207 | 63 | 31 | 845 | 89 | 769 | 329 | 799 | 754 | 559 | 790 | 241 | 297 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 709 | 0 | – | – | 4 | – | 10 | 11 | 5 | 4 | 10 | – | – | 14 | 9 |
| 207 | – | 0 | – | – | 5 | – | – | – | – | – | – | 8 | – | – |
| 63 | – | – | 0 | – | – | – | – | – | – | – | 6 | – | – | – |
| 31 | 4 | – | – | 0 | – | 14 | 15 | 9 | 8 | 14 | – | – | 18 | 13 |
| 845 | – | 5 | – | – | 0 | – | – | – | – | – | – | 5 | – | – |
| 89 | 10 | – | – | 14 | – | 0 | 13 | 15 | 6 | 12 | – | – | 4 | 14 |
| 769 | 11 | – | – | 15 | – | 13 | 0 | 11 | 7 | 5 | – | – | 17 | 7 |
| 329 | 5 | – | – | 9 | – | 15 | 11 | 0 | 9 | 8 | – | – | 19 | 4 |
| 799 | 4 | – | – | 8 | – | 6 | 7 | 9 | 0 | 6 | – | – | 10 | 8 |
| 754 | 10 | – | – | 14 | – | 12 | 5 | 8 | 6 | 0 | – | – | 16 | 4 |
| 559 | – | – | 6 | – | – | – | – | – | – | – | 0 | – | – | – |
| 790 | – | 8 | – | – | 5 | – | – | – | – | – | – | 0 | – | – |
| 241 | 14 | – | – | 18 | – | 4 | 17 | 19 | 10 | 16 | – | – | 0 | 18 |
| 297 | 9 | – | – | 13 | – | 14 | 7 | 4 | 8 | 4 | – | – | 18 | 0 |

the snowball bias: The bias affects only properties that correlate with the degree, yet the current observations do not reveal any correlation between the spatial and topological structure of the network.

Let us assume that the observed edge length distribution is a multiplication of an individual's probability $p_{accept}(d)$ to accept a contact at distance $d$ and the number of opportunities $M(d)$ at distance $d$, so that

$$p_{edge}(d) = p_{accept}(d) \cdot M(d) \ . \tag{16}$$

Using land use data to obtain $M(d)$ it is possible to extract $p_{accept}(d)$ from the survey data (Fig. 4(b)). For this, it is necessary to re-weight every occurrence of an edge connected to ego $v$ by $1/M_v(d)$. Here, every $M_v(d)$ is individually computed for every ego as the sum of opportunities at distance $d$. Areas outside Switzerland contribute zero opportunities.

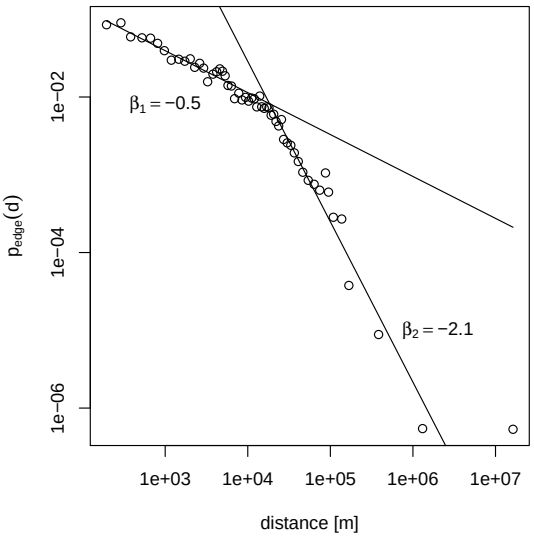
The function $p_{accept}(d) \sim d^{\alpha}$ with $\alpha \approx -1.6$ fits well to the resulting distribution. This may be an indicator that the change of the exponent in the edge length distribution $p_{edge}(d)$ is induced by boundary effects. In fact, the initial seeds of the snowball are drawn within Canton Zurich, i.e. samples are concentrated within the metropolitan area of Zurich. The southern border of Germany is approximately 20 km north of Zurich city.

The spatial distribution of social contacts defines possible origin-destination relations of leisure related travel but makes no statement about the actual number of trips. It is assumed that reported physical contacts, i.e. face-to-face meetings, are located at either one actor's residential location.[2] Then, given the frequency distribution $f(d)$ of physical contacts, the distribution of trips $p_{trip}(d)$ is obtained by

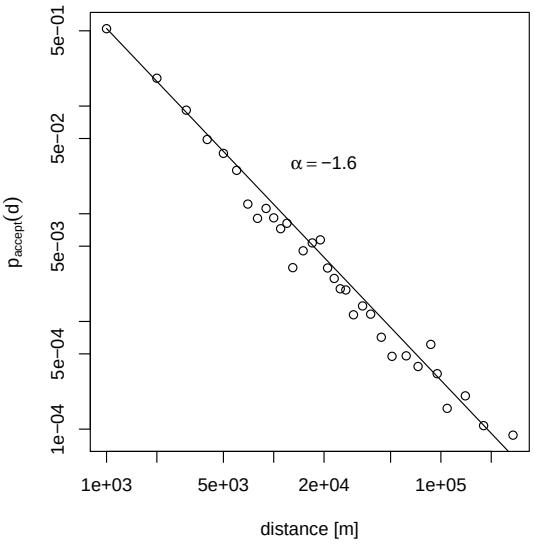
$$p_{trip}(d) = f(d)\, p_{edge}(d) \ , \tag{17}$$

i.e. by a multiplication of the functions describing the frequency distribution and the spatial distribution of leisure contacts. Figure 5(a) shows the trip distance distribution $p_{trip}(d)$. Similar to the edge length distribution (Fig. 4(a)), $p_{trip}(d)$ also exhibits a short range and long range domain. Both domains follow again a power law $p_{trip}(d) \sim d^{\gamma_{1/2}}$, however, with smaller exponents $\gamma_1 \approx -1.1$ and $\gamma_2 \approx -3.5$. The qualitative similarities between $p_{edge}(d)$ and $p_{trip}(d)$

[2] Once the data from the 8-days travel diary (Sec. 2.2) is available the precise locations of face-to-face meetings are known.




(a)




(b)

Figure 4: (a) Edge length distribution $p_{edge}(d)$, (b) acceptance probability distribution $p_{accept}(d)$. Samples are aggregated into distance bins each containing 100 samples.

indicate that the frequencies of visits given a contact follow the same basic scaling law, i.e. $f(d) \sim d^{\eta}$, as the probability of a contact given an opportunity. Figure 5(b) shows the frequency distribution $f(d)$ with respect to the face-to-face contact mode. The distribution does not (and should not) exhibit the two distance domains as $p_{trip}(d)$ and $p_{edge}(d)$, but, apart from some outliers at very long distances, follows roughly $f(d) \sim d^{\eta}$ with $\eta \approx -0.4$.

We further conclude that leisure contacts do not only exist more frequently with short distances, they are also activated more frequently at short distances. Contacts that are met at least once a week have an average length of less than 10 km, whereas contacts that are met just once per year are more than 100 km distant.

## 3.3 Homophily

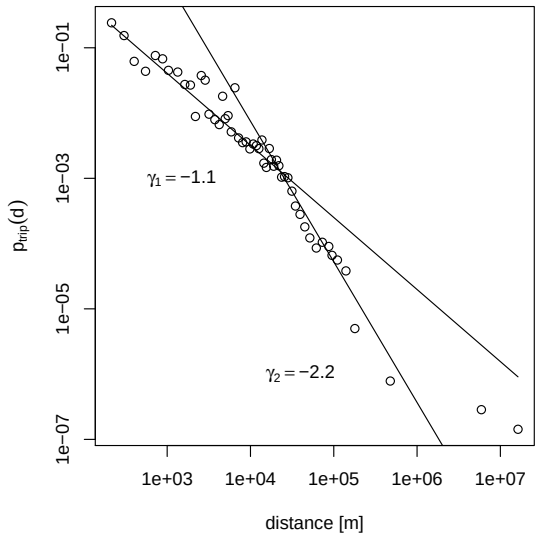
Analogous to spatial distance, decreasing "social distance" between two actors increases the probability of being connected, where "social distance" denotes a measure of how much two individuals differ in their socio-demographic attributes. In social network analysis this phenomenon is known as homophily [29].

The attribute which induces the strongest degree of homophily is age. It can be quantified with the Pearson correlation coefficient of the age values at either ends of all edges:

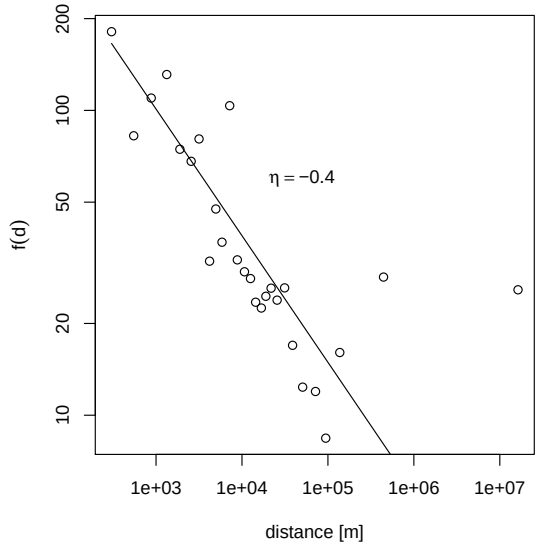
$$r_a = \frac{\sum\limits_{(vw)} a_v a_w - M^{-1} \left( \sum\limits_{(vw)} \frac{1}{2} (a_v + a_w) \right)^2}{\sum\limits_{(vw)} \frac{1}{2} (a_v^2 + a_w^2) - M^{-1} \left( \sum\limits_{(vw)} \frac{1}{2} (a_v + a_w) \right)^2}, \tag{18}$$

where $a_v$ and $a_w$ denote the age [years] of vertices $v$ and $w$, and $\sum_{(vw)}$ goes over all edges (contrary to Eq. 9 because both age values are known).

A correlation coefficient of $r_a = 0.55$ indicates a strong correlation. Interesting details on how homophily with respect to age changes during the course of life can be revealed if one looks at the alters' age distribution. For respondents of age below 30 years the distribution is rather narrow (Fig. 6).




(a)




(b)

Figure 5: (a) Trip distribution $p_{trip}(d)$, (b) frequency distribution of physical contacts $f(d)$. Samples are aggregated into distance bins each containing 100 samples.

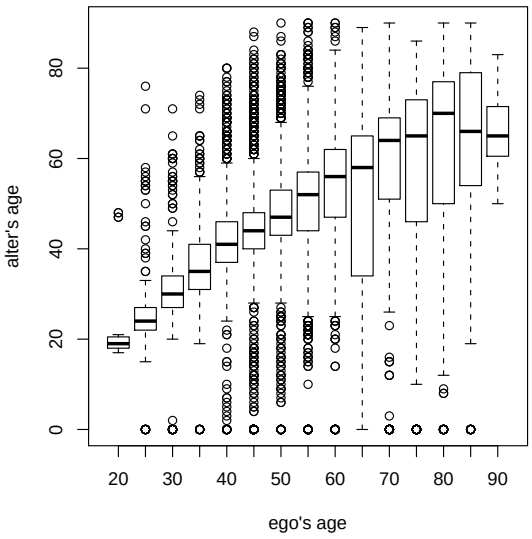



Figure 6: Box plot (conforming to [51]) of the distribution of alters' age. Lower and upper box bounds are first and third quartile respectively.

This means that alters are of nearly the same age as the ego. With increasing age the distribution becomes broader which is rather intuitive since the absolute age difference becomes less important as one becomes older. Also remarkable is that at an age of approximately 40 years more contacts at the lower and upper end of the scale occur. The contacts at the lower end include the respondent's children, the contacts at the upper bound indicate that the parental generation becomes more relevant considering leisure contacts.

Apart from age, homophily with respect to gender plays a significant role. Interestingly, the degree of homophily of this attribute differs between female and male. The probability that a contact of a female ego has the same gender is 0.72. For male respondents, the probability that the alter has the same gender decreases to 0.64. As a consequence of the resulting snowball recruitment bias, the sample itself is biased such that 58 % of the vertices are female, whereas the share in the Swiss micro-census is only 51 %.

Considering the level of education, a similar effect is observed. Categorising respondents into "academics" (university or university of applied science) and "non-academics" reveals that the survey data contains surprisingly many academics. However, homophily within academics is less pronounced compared to non-academics. Regarding academics, alters' level of education is almost equally distributed over both categories. In contrast, for a non-academic respondent the probability that an alter belongs to the same category is 0.75. Consequently, the above average share of academics (45 % in survey data, 15 % in Swiss micro-census) can only be explained by the greater uncorrected average degree ($\bar{k}_{\text{academic}} \approx 23$ and $\bar{k}_{\text{non-academic}} \approx 18$).

# 4 Discussion

The present study confirms a couple of findings from other studies but also reveals some new aspects of social networks and the link to travel behaviour. The average number of leisure contacts per individual is estimated to $\hat{\bar{k}} = 13.3$ and is comparable to, for instance, the study of Frei and Axhausen [20] ($\bar{k} = 12.4$). Carrasco [18] observes a mean degree of $\bar{k} = 12.1$ but does not explicitly ask for leisure contacts. Carrasco's and the present name generator partly overlap in that they both also ask for emotionally important contacts. During a comparison one should, however, bear in mind that the name generator is a sensitive aspect in such surveys. The tail of the degree distribution exhibits an exponential decay which is different from the often observed power law decay.

Compared to other studies, triangular configurations in the graph appear to be less frequent. Even if the edges from the sociogram data are included, the mean clustering coefficient of, for instance, networks of company directors ($C_{(2)} = 0.59$) [49] or physics co-authorship ($C_{(2)} = 0.43$) [8] is two or three times as large as ours. However, one should recall that the present study is not embedded into any institutional setting, such as company directors or co-authors. As we observe in the sociogram data, alters are organised into several different communities: each ego reports at average 4.25 cliques. It is thus not surprising that leisure contacts show less transitivity.

The question if the present study supports the notation of "six degrees of separation" [52] is still open. The current average path length between two seed-vertices is 9.9. However, it is still possible that the snowball detects shorter paths in the next iterations. Non-responding vertices can, at first, disrupt the expansion of the snowball, but can still be reached from other components and thus act as bridge-vertices.

With the knowledge about the spatial structure of the leisure network, we can show how to infer the leisure related trip distribution as a multiplication of the functions describing the spatial distribution of contacts and their frequency of physical meetings. A comparison with data from the Swiss micro-census reveals that the present approach is in fact able to infer reasonable leisure travel patterns. The micro-census includes a representative travel survey that also captures trip purposes. Figure 7 shows both, the trip distribution of the present study as well as the distribution of 3833 trips with purpose "visit" from the micro-census. The comparison needs to be treated with care because the sample of this study is spatially biased towards Canton Zurich, whereas the micro-census is representative for all of Switzerland. Yet, the distribution inferred form the social network follows quite well the distribution obtained from the micro-census.

Furthermore, the scaling laws of human travel identified in this study can be put in relation to the observations of Brockmann et al. [6] and González et al. [5]. Brockmann et al. propose that the circulation of bank notes can be used as a proxy for human travel. They show that the probability of a bill traversing a distance $d$ (in a short time period) is well described with $p(d) \sim d^{\gamma}$ and estimate the exponent to $\gamma = -1.6$. The approach of González et al. in which they use trajectories obtained from mobile phone data shows that the probability of an individual to make a displacement of distance $d$ is described by $p(d) = (d + d_0)^{\gamma} \exp(-d/\kappa)$ with $\gamma = -1.75$. Both studies show significant negatively greater exponents compared to the present study ($\gamma \approx -1.1$).

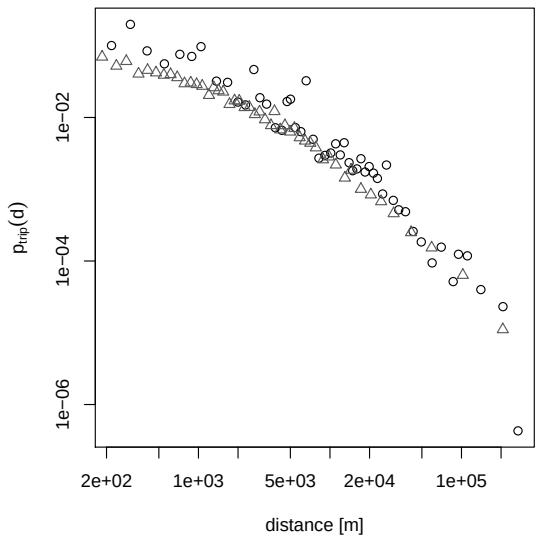



Figure 7: Comparison of $p_{trip}(d)$ between the present study (circles) with data from the Swiss micro-census (triangles). Samples are aggregated into 50 bins each containing (approximately) the same number of samples.

# 5 Summary

This article presents insights into the structure of a large-scale spatially embedded social network. The survey instrument accounts for both, revealing the topology of the network as well as its spatial structure. While it seems practically impossible to obtain complete networks of regular leisure contacts, it is useful to go beyond the egocentric network by employing the snowball sampling approach. With the large sample size (406 respondents naming more than 7000 contacts) the density of personal networks is so high that even paths connecting the initial seed vertices are found. Clearly, just having connecting components does not allow to estimate network-global parameters directly. The data, however, will provide evidence about the order of magnitude of the "degree-of-separation" distribution. This is not only useful for a much better estimate of statistical models for transport and communication modelling, but it particularly provides a sound basis for the spreading of diseases or rumours (see for instance chapter 5 of [53]).

Regarding the reciprocal interaction of social networks and travel, we focused in this article on one direction: from the social network to travel. We show that the trip distribution with respect to leisure travel is a multiplication of the functions describing the spatial distribution of leisure contacts and the frequency distribution of physical meetings. Our results are consistent with the Swiss micro-census.

The other direction, from travel to the social network, represents an aspect that is open for further research. Some initial work in this direction, by generating social networks with agent-based transport micro-simulations, has already been conducted [54, 55].

Once the snowball survey is completed it represents a large data set covering both, the topology of the social network and its spatial structure. The data from the travel diary will enrich the network data with details on the individuals' mobility patterns.

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

## References


[1] R. Follmer, U. Kunert, J. Kloas, H. Kuhfeld, Tech. rep., infas/DIW (2004), www.kontiv2002.de

[2] ARE/BFS, *Mobilität in der Schweiz, Ergebnisse des Mikrozensus 2005 zum Verkehrsverhalten* (2007)

[3] K.W. Axhausen, A. Zimmermann, S. Schönfelder, G. Rindsfüser, T. Haupt, Transportation **29**(2), 95 (2002)

[4] N. Schüssler, K.W. Axhausen, Transportation Research Record **2105**, 28 (2009)

[5] M. González, C. Hidalgo, A.L. Barabási, Nature **453**, 779 (2008)

[6] D. Brockmann, L. Hufnagel, T. Geisel, Nature **439**(26), 462 (2006)

[7] K.W. Axhausen, N. Schüssler, *Improving and replacing travel diaries using mobile tracing?*, presentation at Mobile Tartu (2010)

[8] M.E.J. Newman, Physical Review E **64**(016131) (2001)

[9] 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)

[10] H. Ebel, L.I. Mielsch, S. Bornholdt, Physical Review E **66**(3), 1 (2002)

[11] B. Latané, J.H. Liu, A. Nowak, M. Bonevento, L. Zheng, Personality and Social Psychology Bulletin **21**(8) (1995)

[12] C. Johnson, R.P. Gilles, Review of Economic Desgin **5**(3), 273 (2000)

[13] L.H. Wong, P. Pattison, G. Robins, Physica A **360**(1), 99 (2006)

[14] D. Mok, B. Wellman, R. Basu, Social Networks **29**, 430 (2007)

[15] G. Daraganova, Ph.D. thesis, University of Melbourne, Department of Psychology (2008)

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

[17] J. Larsen, J. Urry, K.W. Axhausen, Information, Communication and Society **11**(5), 640 (2008)

[18] J.A. Carrasco, Ph.D. thesis, University of Toronto (2006)

[19] J. Silvis, D. Niemeier, R. D'Souza, *Social networks and travel behaviour: Report from an integrated travel diary*, in *Proceedings of the meeting of the International Association for Travel Behavior Research (IATBR)* (Kyoto, Japan, 2006)

[20] A. Frei, K.W. Axhausen, Working Paper 439, ETH Zürich, Institute for Transport Planning and Systems (2007)

[21] P. van den Berg, T. Arentze, H. Timmermans, *Social networks, ICT use and activity travel patterns: Data collection and first analyses*, in *Proceedings of the International Conference on Design & Decision Support Systems in Architecture and Urban Planning* (Eindhoven, 2008)

[22] Department for Transport, *Transport statistics bulletin: National travel survey: 2006*, London (2006)

[23] U.S. Department of Transportation, *Summary of travel trends: 2001 national household travel survey*, Federal Highway Administration, Washington (2001)

[24] M. Kowald, A. Frei, J. Hackney, J. Illenberger, K.W. Axhausen, Working Paper 582, ETH Zürich, Institute for Transport Planning and Systems (2009)

[25] M. Kowald, K.W. Axhausen, Environment and Planning A (forthcoming)

[26] B. Hogan, J.A. Carrasco, B. Wellman, Field Methods **19**(2), 116 (2007)

[27] B.H. Erickson, Sociological Methodology **10**(1), 276 (1979)

[28] S. Gabler, ZUMA-Nachrichten **16**(31), 47 (1992)

[29] M. McPherson, L. Smith-Lovin, J.M. Cook, Annual Review of Sociology **27**, 415 (2001)

[30] C. Steglich, T.A.B. Snijders, Sociological Methodology (2010)

[31] G. Kossinets, D.J. Watts, American Journal of Sociology **115**(2), 405 (2009)

[32] *The American Association for Public Opinion Research* (2009)

[33] K.W. Axhausen, C. Weiss, Survey Practice **3**(2) (2009)

[34] P.V. Mardsen, Annual Review of Sociology **16**, 435 (1990)

[35] L.A. Goodman, The Annals of Mathematical Statistics **32**(1), 148 (1961)

[36] D.D. Heckathorn, Social Problems **44**(2), 174 (1997)

[37] D.D. Heckathorn, Social Problems **49**(1), 11 (2002)

[38] O. Frank, *Estimation of population totals by use of snowball samples* (Academic Press, New York, 1979), pp. 319–346

[39] J. Johnson, J. Boster, D. Holbert, Social Networks **11**, 135 (1989)

[40] T.A.B. Snijders, Bulletin de Méthodologie Sociologique **36**, 59 (1992)

[41] O. Frank, T. Snijders, Journal of Official Statistics **10**(1), 53 (1994)

[42] E. Volz, D.D. Heckathorn, Journal of Official Statistics **24**(1), 79 (2008)

[43] K.J. Gile, M.S. Handcock, Sociological Methodology **40**(1), 285 (2010)

[44] J. Illenberger, G. Flötteröd, Working Paper 11-01, TU Berlin, Transport Systems Planning and Transport Telematics (2011)

[45] M. Andre, K. Ijaz, J.D. Tillinghast, V.E. Krebs, L.A. Diem, B. Metchock, T. Crisp, P.D. McElroy, American Journal of Public Health **96**(11), 1 (2006)

[46] W. Aiello, F. Chung, L. Lu, *A Random Graph Model for Massive Graphs*, in *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing* (Association of Computing Machinery, New York, 2000), pp. 171–180

[47] M.E.J. Newman, Physical Review Letters **89**(20), 1 (2002)

[48] M.E.J. Newman, SIAM Review **45**(2), 167 (2003)

[49] G.F. Davis, M. Yoo, W.E. Baker, Strategic Organization **1**(3), 301 (2003)

[50] R. Dunbar, Behavioral and brain sciences **16**, 681 (1993)

[51] J.W. Turkey, *Exploratory data analysis* (Addison Wesley, 1977)

[52] S. Milgram, Psycology Today **2**, 60 (1967)

[53] T. Smieszek, Ph.D. thesis, ETH Zurich, Switzerland (2010)

[54] F. Marchal, K. Nagel, Transportation Research Record **1935**, 141 (2005)

[55] J. Hackney, Ph.D. thesis, ETH Zurich, Switzerland (2009)