UNIVERSITÄT BERN

CRED
CENTER FOR REGIONAL ECONOMIC DEVELOPMENT

Cities and the Structure of Social Interactions: Evidence from Mobile Phone Data*

CRED Research Paper No. 13

Konstantin Büchel
University of Bern,
CRED

Maximilian von Ehrlich
University of Bern,
CRED

December, 2016

Abstract

Social interactions are considered pivotal to urban agglomeration forces. This study employs a unique dataset on mobile phone calls to examine how social interactions differ across cities and peripheral areas. We first show that geographical distance is highly detrimental to interpersonal exchange. We then reveal that individuals residing in high-density locations do not benefit from larger social networks, but from a more efficient structure in terms of higher matching quality and lower clustering. These results are derived from two complementary approaches: Based on a link formation model, we examine how geographical distance, network overlap, and sociodemographic (dis)similarities impact the likelihood that two agents interact. We further decompose the effects from individual, location, and time specific determinants on micro-level network measures by exploiting information on mobile phone users who change their place of residence.

Key words: Social Interactions; Agglomeration Externalities; Network Analysis; Sorting.

JEL classification: R1; R23; Z13; D85.

CRED
Universität Bern
Schanzeneckstrasse 1
Postfach
CH-3001 Bern

Tel. +41 031 631 37 11
info@cred.unibe.ch
www.cred.unibe.ch

*We benefited from numerous suggestions by Juan Becutti, Aymo Brunetti, Fabian Gunzinger, Diego Puga, Elisabet Viladecans-Marsal, and participants at the IEB research seminar in Barcelona, the Verein für Socialpolitik Meeting in Tellow, and the Meeting of the Urban Economics Association in Minneapolis. A very special thanks is due to *Swisscom AG* for providing the facilities and data to conduct this research project; we are particularly indebted to Dr. Imad Aad, who accompanied the project for two years. We also want to thank *search.ch* and *comparis.ch* for providing data on travel times and usage statistics of messenger apps.

1 Introduction

Social interactions lie at the nexus of two key themes in economics: sustained aggregate growth and the concentration of economic activity in cities. In a widely cited paper, Robert Lucas (1988) models human capital accumulation as main driver behind economic development. Interpersonal exchange is pivotal to the narrative of his framework, with Lucas (1988, p. 19) defining "human capital accumulation [a]s a *social activity*, involving groups of people". Through this social learning process, human capital not only provides an internal value to its owner but also exerts a positive externality on peers, which fosters the creation of new ideas and with it sustained development. In reference to urban theorist Jane Jacobs (1969), Lucas (1988) suggests that these externalities are especially prevalent in cities, which consequentially act as engines of growth. This notion reflects one of the classic agglomeration forces described by Alfred Marshall (1890), who argues that the dense concentration in cities facilitates the flow of information and knowledge, since social interactions diminish over space. Although social interactions are considered to play a decisive role for the aggregate dynamic and spatial organisation of the economy, empirical work uncovering the alleged micro-mechanisms has remained fragmentary at best.

This paper studies the relation between spatial structure and social interactions in order to test fundamental assumptions underlying the agglomeration forces discussed in the literature. The analysis builds on anonymised mobile phone calls between June 2015 and May 2016. This allows us to examine the interplay between local characteristics and social interactions as we not only observe comprehensive communication patterns but also location information derived from transmitting antennas and billing data. Based on this rich dataset and concepts from the network literature (e.g. Jackson, 2008), we investigate three main questions. *First*, how does geographical distance impact social interactions? *Second*, what is the relation between population density and the size of an individual's social network? *Third*, does population density affect the quality / efficiency of social interactions in terms of matching quality, clustering and network perimeter? To answer these questions we employ link formation models in the spirit of Graham (2014) and additionally estimate the impact of population density on various micro-level network measures. The sorting of individuals with specific characteristics can distort the results of both approaches. We therefore base our inference on individuals who change their place of residence (i.e. "movers") to back out time constant unobservables and correctly identify the role of distance as well as density-related externalities.

We show that distance is highly detrimental to social interactions, despite epoch-making progress in communication technologies. Contrary to the conventional view, this does not translate into larger networks in cities compared to the periphery. Density-related externalities rather arise in terms of network efficiency, namely better matching quality, lower clustering, and smaller distance costs. We are not aware of any study that has delivered comparable evidence on regional differences in both network size and network efficiency. Below, we discuss the main findings with reference to the related literature.

1.1 Related Literature

Social Interactions and Distance. Models that incorporate knowledge and learning spillovers as an agglomeration force typically assume that distance is costly to social interactions (e.g. Glaeser, 1999). The widespread adoption of information and telecommunication technologies popularised the "death-of-distance" argument (e.g. Cairncross, 2001), which raises the intriguing question of whether these technologies will fundamentally change the structure of cities (see Ioannides et al., 2008) or even make them obsolete (see Gaspar and Glaeser, 1998). We demonstrate that the social interactions recorded by mobile phones are surprisingly localised, with more than 60 percent of ties occurring between individuals that reside within less than 10 km distance of each other. Importantly, we aim for a causal interpretation and therefore estimate a network formation model based on movers. This novel analysis confirms that distance is highly detrimental to forming and maintaining social ties. A recent study by Levy and Goldenberg (2014) uncovers similar patterns for email traffic and online social media contacts. We interpret this as solid evidence against the death-of-distance hypothesis in the social exchange context.

Quantity of Social Interactions. Building on the assumption that distance is costly to social interactions, numerous micro-founded models of urban agglomeration economies have been developed (cf. Duranton and Puga, 2004). One body of literature focuses on the claim that the quantity of social interactions increases with local population density. Glaeser (1999) formalises the classic idea of Marshall (1890) that individuals acquire skills by interacting with each other. As cities are more densely populated than the hinterland, they facilitate more meetings in his framework and thus accelerate the social learning process. Another example is that of Sato and Zenou (2015), who model social interactions and their impact on employment outcomes. They propose that city residents maintain larger networks than rural dwellers, enabling them to acquire more information on the labour market, which reduces job search frictions and unemployment. Two empirical studies support the hypothesis that cities facilitate interpersonal exchange. Charlot and Duranton (2004) use survey data on workplace communication in France, while Schläpfer et al. (2014) examine mobile phone records for Portugal. Both studies find that the average number of (unique) social interactions increases with population size. However, neither can plausibly isolate the causal impact of density from non-random sorting, as the first paper relies on cross-sectional data and the second paper narrows down to a descriptive analysis. Burley (2015) studies the German Socio-Economic Panel and finds that population density is only positively correlated with an index of social interactions, as long as person specific characteristics are ignored.[1] Our results reinforce this finding: we also show that the positive effect of cities compared to the hinterland vanishes, once targeted sorting of individuals is accounted for. Given the pattern emerging from these

[1] Based on US survey data, Brueckner and Largey (2008) also examine population density and social interactions obtaining consistently negative correlations. These findings are at odds with the other studies, as they suggest that cities are too dense from a social interaction point of view.

four studies, the claim that cities produce more social contacts than the periphery seems unfounded.[2]

Efficiency of Social Interactions – Matching Quality. Another strand of literature argues that cities do not necessarily increase the quantity of social interactions but rather improve their quality / efficiency. In the model of Berliant, Reed and Wang (2006), agents possess differentiated types of knowledge. The effect of cities on the number of social interactions then becomes twofold, as densely populated areas increase the number of random meetings but also make agents more selective regarding matching quality. Hence, while cities do not necessarily affect the number of social interactions, Berliant, Reed and Wang (2006) show that their quality in terms of knowledge complementarity should improve with increasing population density. With the aim of providing an empirical test for the matching channel, Abel and Deitz (2015) study data on job searching of college graduates. They find that larger and thicker labour markets indeed improve the matching between job advertisements and applicants' qualifications. To the best of our knowledge, no study to date has assessed this hypothesis with respect to social interactions. We formulate two tests, one relying on a network formation model, and the other analysing the social adjustment process among movers. Both approaches indicate that urban dwellers indeed benefit from higher quality matches compared to people living in the hinterland.

Efficiency of Social Interactions – Clustering. Borrowing from the network literature, the level of clustering / triangular relations is an additional dimension of efficiency that is sometimes assumed to vary regionally. Granovetter (1973) famously argues that weak ties are often more valuable in terms of information provision than strong ties. He formally defines a weak tie as a social relation between two agents who have no overlap in their personal networks. In contrast, strong ties involve triangular relations that bring about redundancies in the process of information diffusion. Sato and Zenou (2015) claim that cities not only increase the number of social interactions – as discussed above – but also give rise to a disproportionally high number of weak tie relations that are more valuable in the job market. We calculate the clustering coefficient (i.e. the share of triangular relations) of each agent in the data set and test whether the level of clustering systematically varies with population density. We find that personal networks in cities indeed tend to be characterised by lower levels of clustering and thus have a higher fraction of weak ties. This finding suggests that cities may facilitate the diffusion of information, although the average number of social interactions is not necessarily larger than in more sparsely populated areas.

The following section elaborates on the main concepts. Section 3 introduces the data used in the empirical analysis. Section 4 explains the empirical strategy. Section 5 discusses the results. Section 6 concludes.

[2]Other factors that have been shown to impact the level of social interactions are homeownership (e.g. Hilber, 2010) and racial fragmentation (e.g. Alesina and La Ferrara, 2000; Brueckner and Largey, 2008).

2 Cities and Social Interactions: Main Concepts

We consider a directed network with N nodes each representing a unique phone customer which we denote by $i \in \mathcal{N} = \{1, ..., N\}$. Each customer has a place of residence, r, which is assigned either on the municipality or postcode level. The number of nodes at location r is N_r, and so with R denoting the total number of different residences, $N = \sum_r N_r$ holds. Finally, $\mathcal{R}_r$ is the set of individuals living in location r.

A link between nodes i and j is denoted by $g_{ij} = 1$, while the absence of a link is marked as $g_{ij} = 0$. The network can then be characterised by a pair $(\mathcal{N}, \mathcal{G})$ where $\mathcal{G} = [g_{ij}]$ is a $N \times N$ adjacency matrix. As in Graham (2014), we assume that rational agents i and j establish a link if the net surplus from doing so is positive. This yields a random utility model of the form

$$g_{ij} = \mathbf{1}\left(X_{ij}'\eta + \nu_i + \nu_j + U_{ij} \geq 0\right), \tag{1}$$

where X_{ij} is a vector of dyad attributes (i.e. pair specific characteristics), ν_i and ν_j denote agent specific characteristics, and U_{ij} is a randomly distributed component of link surplus. We are particularly interested in the role of dyad attributes, which we divide into three groups: geographical distance or travel time (T_{ij}), the number of friends i and j share ($F_{ij} = \sum_{k=1}^{N} g_{ik}g_{jk}$), and matching ($m(\nu_i, \nu_j, \delta)$). As defined in this study, higher levels of $m(\cdot)$ increase link surplus, which is why we refer to it as matching *quality*. Importantly, it absorbs the spread between Q individual characteristics of agent i and j, $|\nu_i - \nu_j|$, which – depending on the specific attribute $q \in Q$ – may be positively (i.e. $\delta_q > 0$) or negatively correlated (i.e. $\delta_q < 0$) with matching quality. Based on these considerations we define the vector X_{ij} as

$$X_{ij}'\eta = \eta_1 \cdot T_{ij} + \eta_2 \cdot F_{ij}(\mathcal{G}) + \eta_3 \cdot m(\nu_i, \nu_j, \delta). \tag{2}$$

If link-surplus is indeed a function of these three dyad-specific factors, this may have important consequences for the network topography across rural and urban areas. Provided that distance is costly for social interactions, regional differences in population density may determine the *size* of an agent's social network. This is of interest, because social contacts can foster the diffusion of information, promote trust and thereby lower transaction costs, and facilitate learning from peers (Granovetter, 2005; Gui and Sugden, 2005; Jackson, 2014) in addition to having intrinsic value for a person's well-being (Burt, 1987). We further focus on matching and common friends (or clustering), as they have implications for a *network's efficiency*: Matching reflects the quality of a specific contact, which incorporates various dimensions such as productivity enhancing skill complementarity, or shared interests (e.g. Berliant, Reed and Wang, 2006). Clustering governs the informational value of a link, since contacts who share a common friend introduce redundancies and are therefore less valuable in the information diffusion process (Granovetter,

1973). In return, sharing mutual contacts fosters cooperative and pro-social behaviour, because the triangular relation can act as a reputational control and retaliation device (Jackson, 2014).

Network Size and Degree Centrality. We first discuss the relation between population density and the size of an individuals social network, which we measure based on *degree centrality*, formally defined as

$$D_i(\mathcal{G}) = \#\{j : g_{ij} = 1\}. \tag{3}$$

The degree yields the number of distinct peers with whom agent i interacts socially and therefore the number of sources that potentially forward valuable information. Typically, urban economic theory (e.g. Glaeser, 1999; Sato and Zenou, 2015) presumes that cities provide a favourable environment for social interactions and support larger network sizes, as they are more densely populated than rural communities. The underlying argument hinges on the assumption that the costs of social interactions increases with distance. Let us abstract from the matching spread, $m(\nu_i, \nu_j, \delta)$, as well as triangular ties, $F_{ij}(\mathcal{G})$, and focus on the relationship between distance and population density. A stylised argument is as follows: On weekdays an agent i needs to keep her travelling costs low, and she therefore has random encounters only with people in her municipality, $j \in \mathcal{R}_r$. At the weekend, however, the radius of the agent's actions is unbounded, so that she might form ties with people living outside her place of residence, $k \notin \mathcal{R}_r$. Since people spend more time in their residence's vicinity, the probability to acquire social contacts among neighbours, $P_r = P(g_{i,j\in\mathcal{R}_r} = 1)$, is larger than for the rest of the population, that is $P_r > P_{-r} = P(g_{i,k\notin\mathcal{R}_r} = 1)$. In the outlined example, the size of a person's social network positively depends on the population living in the neighbourhood, N_r, so that cities support a larger degree than rural municipalities, i.e.

$$D_i = N_r \cdot P_r + (N - N_r) \cdot P_{-r} \qquad \text{with } \frac{\partial D_i}{\partial N_r} > 0. \tag{4}$$

While this intuitive rationale is appealing, it may be challenged from two angles, namely from biological/anthropological and search strategic points of view.

In evolutionary biology, Dunbar (1992, 1998) has famously advocated and popularised the *social brain hypothesis*. It challenges the field's traditionally dominant view that brains evolved to address ecological problem-solving tasks, such as foraging. Instead the social brain hypothesis attributes the growth in primates' brain sizes to the computational demands of their increasingly complex social systems. Indeed, empirical analyses reveal a close relation between neocortex volume and mean social group size among primates. This has been interpreted as evidence that there is a species-specific upper limit to group size that is set purely by cognitive constraints. For humans, Dunbar (1993) calculates the upper

limit to lie between 100 and 230 social contacts, citing anthropological studies on modern hunter-gatherer societies as evidence that support his prediction. Recent studies explore this hypothesis by analysing patterns among adults' prefrontal cortex volume, cognitive ability, and the size of their social networks (Powell et al., 2012; Stiller and Dunbar, 2007) or by exploiting social media user statistics (Dunbar, 2015). In consideration of the manifold results corroborating the social brain hypothesis, one may note that the size of a person's network is fundamentally restricted by congenital factors. Because the population of practically all Swiss municipalities exceeds the limit for network size as calculated by Dunbar (1993), the number of social interactions may be independent of regional differences in population density.

In equation (4) a random encounter between two persons is equivalent to establishing a link. We now add another layer: After meeting a potential contact, agents can either accept or reject to form a link based on the other person's characteristics. Since forming a link consumes time and cognitive capacity, this introduces a quality-quantity trade-off. Consequently, it may be optimal to reject some potential contacts to wait for a better match (see Berliant, Reed and Wang, 2006). Hence, from a *search strategic* perspective, higher population density may impact network size only marginally, but it may allow for higher selectivity along dyad-specific characteristics. This has important consequences for the analysis of social networks across different regions. Even if densely populated areas improve social networks, the advantages may not be in terms of size but in terms of efficiency. In this respect, *matching quality* between agents i and j, $m(\nu_i, \nu_j, \delta)$, is of key interest, as it determines how well their interests correspond or how fruitful the intellectual exchange between them is. Once we add the strategic component of weighing between quality and quantity to the above mechanics, we would expect a positive effect of population density on matching quality, or network size, or both.

Perimeter of Social Interactions and the Within-Degree. The previous line of reasoning also has implications for the perimeter of a person's network. Essentially, the travel time between two agents can be considered one dimension of matching quality. Assuming that distance is costly when maintaining a link, one would rather form a tie with a neighbour than with an identical person living far away. Following Berliant, Reed and Wang (2006) therefore implies that high-density locations allow people to be more selective regarding the travel distance to their contacts, so that they can minimise travel costs induced through social interactions. Put differently, one may expect that urban dwellers can recruit their contacts within a narrower perimeter, since densely populated cities make high quality matches in a small area possible. In contrast, people in rural areas face a much tighter choice in their neighbourhood, thus they likely prefer to widen the search radius with the objective of improving their network's quality. To analyse these claims, we examine the degree within an individual's neighbourhood or *within-degree*, formally defined as

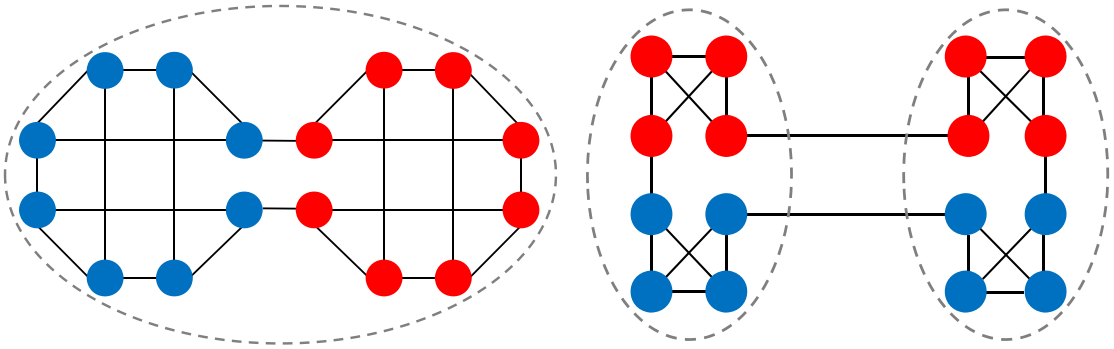
$$DW_i^r(\mathcal{G}) = \#\{i, j \in \mathcal{R}_r : g_{ij} = 1\}. \tag{5}$$

Of course, negligible distance costs would wipe out any differences between cities and rural areas. Costs related to distance may indeed be of secondary importance for a person with naturally few social interactions, whereas highly sociable persons may benefit more from densely populated areas, as recently formalised in a paper by Helsley and Zenou (2014). Consequently, differences in network size may simply be observable due to the sorting of highly sociable types into cities, because they gain disproportionally from low distance costs per contact.

Clustering. In a last step, we discuss regional differences in population density and their implications for clustering in social networks. Clustering is an important network characteristic as it can provide insights into reciprocity and information diffusion. On the one hand, high clustering strengthens reputational concerns and with it the enforcement of social norms and cooperation (e.g. Ali and Miller, 2009), or risk-sharing (e.g. Ambrus, Mobius and Szeidl, 2014). On the other hand, Granovetter (1973) highlights the importance of local bridges for passing on information. An individual with high clustering introduces redundancies in the network, which are inefficient in terms of information diffusion. The *clustering coefficient* for node i is given by

$$C_i = \frac{\sum_{j,k\, j\neq k} g_{jk}}{\sum_{j,k,j\neq k} g_{ij}g_{ik}}, \tag{6}$$

and measures whether an individual's contacts form a tightly knit group ($C_i \to 1$) or are completely separate from each other ($C_i \to 0$). How does population density relate to clustering? There are two potential channels, one mechanical and the other as a consequence of differing preferences. Figure 1 illustrates the mechanical rationale: Panel (a) shows a city with 16 agents, eight blue and eight red. All agents socially interact with three other agents, preferably of the same type. Panel (b) represents a peripheral region with lower population density, therefore the 16 agents are equally split between two municipalities. As in the city, all individuals have a degree of three. Importantly, travelling between the two municipalities is costly, therefore agents prefer to form links with their neighbours. Since every person has only three neighbours of the same type, the network ends up tightly clustered. In contrast, the city makes clustering less likely, because each urbanite can choose among seven agents of the same colour. In the way the example is drawn, the average clustering in the city equals 0, while it amounts to 0.5 in the periphery. As a consequence, the average path length in the city (=2.73) is lower than in the periphery (=3.2), which accelerates the diffusion of information. Thus, low density locations should tend to display higher clustering, simply because residents of these areas face a substantially smaller set of suitable contacts in their direct vicinity compared to urban dwellers. In addition to this purely probabilistic relation between density and



(a) **City:** Average Degree=3, Matching Rate=0.833, Average Clustering=0, Average Path Length=2.73

(b) **Periphery:** Average Degree=3, Matching Rate=0.833, Average Clustering=0.5, Average Path Length=3.2

Figure 1: Clustering in Cities and the Periphery – An Illustrative Example

clustering, preferences for forming links with friends of friends, $F_{ij}(\mathcal{G})$, could be different in cities than in rural areas. Agents face a trade-off in terms of efficient information exchange (i.e. low clustering) and benefits due to stronger reciprocity (i.e. high clustering). The optimal balance may vary regionally due to factors that assign a higher weight to reciprocity or information diffusion. For instance, high quality local institutions may substitute for reciprocity or a dynamic labour market environment may support the value of information diffusion. In addition, clusters may facilitate simultaneous interactions with multiple persons, allowing for larger networks given a certain time constraint. If people living in rural neighbourhoods have more geographically dispersed social networks, clusters of friends could be a strategy to mitigate travel costs. Finally, it has been documented that people living in peripheral areas have a higher proportion of kin ties than urban dwellers (Fischer, 1982). A preference for spending time with relatives most likely increases the clustering in an individual's network as relatives inevitably have an overlap in their circle of acquaintances.

3 Data

The main dataset used in this paper is provided by Switzerland's largest telecommunications operator, *Swisscom AG*, whose market share is 55% for mobile phones and 60% for landlines (ComCom, 2015). The data comprises comprehensive *call detail records (CDR)* of all outgoing calls made by the operator's customers between June 2015 and May 2016. The CDRs include the anonymised phone number of caller and callee, a date and time stamp, a binary indicator for private and business customers, a code for the type of interaction recorded (e.g. call, SMS, MMS), the duration of calls in seconds, and the x-y-coordinates of the caller's main transmitting antenna. We observe finely grained information on about 15 million calls and text messages per day, covering 7.2 million phones,

Figure 2: Degree of Urbanisation – Cities, Hinterland and Periphery

of which 2.4 million are private mobile devices.[3] Along with the anonymised CDRs, the operator also provided monthly updated *customer information* including billing address, language of correspondance (German, French, Italian, English), age, and gender. Table 1 summarises the socio-demographic characteristics of mobile phone customers in our sample, while Table A.3 shows correlations between census data and our customer statistics for various subpopulations. This comparison suggests that the data at hand is highly representative of the Swiss population.

The phone data are complemented by various municipal statistics for 2014 provided by the Federal Statistical Office (FSO), including population figures and the degree of urbanisation as classified by EUROSTAT.[4] Figure 2 shows the regional variation in urbanisation based on the aforementioned measure. We also compute geographical distances between pairs of municipalities and pairs of postcodes using ArcGIS software and shape files for administrative boundaries published by the Federal Office of Topography. Car driving distances between centroids of municipalities and postcode areas were kindly provided by the company *search.ch*. Descriptive statistics for municipalities and postcodes are shown in Table A.1 in the appendix.

The anonymity of Swisscom customers was guaranteed at all steps of the analysis. We never dealt with or had access to uncensored data. A data security specialist retrieved the CDRs from the operator's database and anonymised the telephone numbers using a 64-bit hash algorithm that preserved the international and local area codes. He further removed columns with information on the transmitting antenna before making the data

[3] More specifically, the data set covers 2.4 million private mobile phones, 1.9 million private land lines, 1.1 million corporate mobile phones, and 1.8 million corporate landlines.

[4] See `http://ec.europa.eu/eurostat/ramon/miscellaneous/index.cfm?TargetUrl=DSP_DEGURBA` (last access: 01.06.2016) for more information on the EUROSTAT DEGURBA measure.

available. Once the anonymised data were copied to a fully sealed and encrypted Swisscom workstation, we ran the analysis on site. To utilise information on the transmitting antenna we passed location scripts to Swisscom personnel who executed them for us.

Our primary aim is to observe social networks, but not every instance of phone activity reflects a social interaction in the narrower sense so that the dataset needs to be cleaned beforehand (for a discussion see Blondel, Decuyper and Krings, 2015). In our benchmark analysis, we filter the data as follows: *First*, we restrict the analysis to *calls* between mobile phones. Mobile phones are personal objects and are thus representative of the social network of a single person, while calls from fixed phones possibly resemble overlapping social networks as they are usually shared by multiple users. For the same reason, all results are based on customers who have registered only one active mobile phone number. Customers with multiple active numbers typically include corporate customers, as well as parents acting as invoice recipients for their children. *Second*, we limit the analysis to outgoing calls in order to cover intra-operator and inter-operator activity equally well and to filter out promotional calls by call centres. *Third*, calls with a duration of less than 10 seconds are considered accidental and are therefore excluded from the analysis. *Fourth*, we drop mobile phone numbers that display implausibly low or high monthly usage statistics, with a minimum threshold of 1 minute and a maximum threshold of 56 hours per month, respectively. This removes practically inactive numbers as well as phones used for commercial purposes. *Fifth*, the analysis is limited to *private mobile phones*, so that daily business calls between corporate customers do not create noise in our measures. *Sixth*, some measures require address information for both caller and callee such that inter-operator calls cannot be used in all steps of the analysis. Measures requiring location information for the callee are therefore based on intra-operator calls only, which we weight according to the operator's market share at the callee's billing address. *Finally*, we only use the first 28 days of each month to make the data easily comparable across different time periods.

These steps eliminate approximately 60 percent of the calls recorded for private customers, leaving us with around 60 million calls per month that amount to a total duration of 200 million minutes (for details see Table A.2 in the appendix). We have performed sensitivity checks with regards to all above mentioned dimensions to ensure that our results are robust.

3.1 Descriptive Statistics on Phone Usage and the Social Network

Table 1 shows summary statistics on the mobile phone usage of customers aged 15 to 64 for the filtered data set.[5] The average private mobile-phone users makes about three calls per day with a cumulative duration of nine minutes. Figures 3a and 3b further show that the distributions are markedly right-skewed.

[5]Due to privacy concerns, we worked with decimal age-brackets. This means that a customer aged 24 was assigned to the 20-bracket, while a customer aged 25 belongs to the 30-bracket.

Table 1: Descriptive Statistics, Private Mobile Phone Customers

	Mean	SD	N	Min	Max
Phone Usage, June 2015 – May 2016 (pooled)					
Number of Calls	111.781	109.599	10 399 549	1	10 113
Duration (Minutes)	254.970	295.609	10 399 549	2	3359
Network Characteristics, June 2015 – May 2016 (pooled)					
Degree Centrality	9.202	7.910	10 399 549	1	470
Within-Degree (15 Min. Radius)	7.067	7.231	10 399 549	0	221
Clustering Coefficient	0.092	0.132	10 248 923	0	1
Sociodemographics					
Age	34.964	13.561	866 646	20	60
Female	0.522	–	866 646	0	1
Language: German	0.681	–	866 646	0	1
Language: French	0.270	–	866 646	0	1
Language: Italian	0.043	–	866 646	0	1
Language: English	0.006	–	866 646	0	1

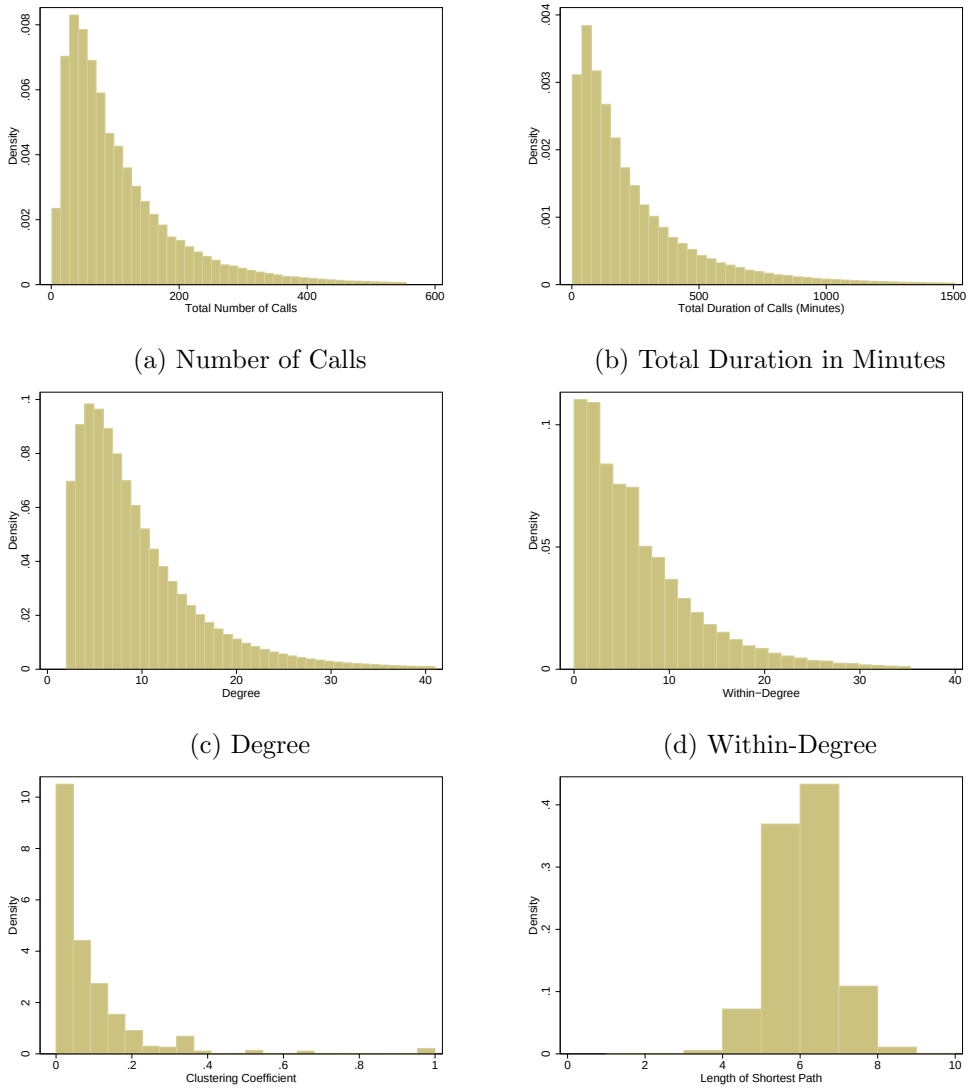
Notes: The table is based on the subsample of customers with phone activity in all 12 months, which we also use in the main analysis. Further filters as described in section 3. Phone usage statistics include in- and outgoing calls. The *within-degree* measures network size within a radius of 15 minutes around an agent's residence.

The network of private mobile phone interactions uncovered by the data exhibits characteristic features of other socially generated networks documented in the literature (Jackson and Rogers, 2007; Watts, 1999): Small diameter and short average path length between pairs, "fat tails" in the degree distribution, and substantial clustering.

To gain insights into the diameter and the average path length, we randomly select 100 individuals and calculate the length of the shortest paths connecting every other private mobile phone users in the data. The mean path length in the sample is 5.6, with the longest path having a length of 12 (1 out of 246 mio.); the histogram plotted in Figure 3f reveals that 88 percent of dyads are separated by 6 or fewer links. This fits strikingly well with the "small-world"-hypothesis first formulated by Milgram (1967) and the early empirical evidence based on a chain letter experiment conducted by Travers and Milgram (1969).

As Figure 3c illustrates, the degree distribution in our social network exhibits "fat tails", so that there are more nodes with relatively high and low degrees, and fewer nodes with medium degrees, than one would find in a network where links are formed uniformly. The average degree in our monthly data is approximately 9, with the vast majority having a degree below 20 and some hub-agents reaching network sizes of 100 links or more. As reported in other studies on social networks, the probability distribution is well fitted ($R^2 = 0.92$) by a power-distribution, $P(D) = cD^{-\varphi}$, with parameter estimates of $\hat{\varphi} = 3.86$ and $\hat{c} = 5.96$.

The clustering coefficient, which measures the tendency of linked nodes to have common neighbours, is, on average, 0.092, with more than 75 percent of the individuals in the dataset having a non-zero clustering coefficient (see Figure 3e). Considering the low



(a) Number of Calls

(b) Total Duration in Minutes

(c) Degree

(d) Within-Degree

(e) Clustering Coefficient

(f) Dyad Distances

Figure 3: Histograms of Phone Usage Statistics & Network Characteristics for June 2015.

density of our network (≈ 0.00002), the observed clustering is evidently larger than in a benchmark network where links would have been generated by an independent random process.

4 Identification

In order to analyse the impact of geography and location characteristics on the structure of social interactions we conduct two complementary identification strategies. The first aims to identify factors that predict the likelihood of individuals i and j forming a link and is referred to as *network formation*. In particular, this approach allows us to study the effects of distance between i's and j's place of residence on the probability that they form a link.

It further enables inference on the preference for triadic relations. The presence of network overlap may influence the likelihood that i and j establish a link as the returns may be higher or lower if it involves mutual contacts. Moreover, we study whether homophily – the process of matching on common characteristics – is prevalent in the data.

The second approach, to which we refer as *network topography*, estimates the effect of local characteristics on individual-level network measures. This relates to the equilibrium outcome of network formation at different places and allows us to examine the impact of location specific attributes on social networks.

Sorting of individuals with specific characteristics can affect the results of both approaches. As we observe the full social network for one year we can exploit changes in the address of mobile phone customers; this enables us to isolate the role of systematic sorting and to obtain causal estimates of geography and population density on network formation and network topography.

4.1 Network Formation

We observe the social network's adjacency matrix $\mathcal{G}_t = [g_{ij,t}]$ in each month $t \in \{1, ..., 12\}$. Following Graham (2014), we specify the probability that two nodes i and j form a link as

$$g_{ij,t} = \mathbf{1}(\beta g_{ij,t-1} + T'_{ij,t}\eta_1 + F'_{ij,t-1}\eta_2 + Z'_{ij}\rho + \phi_1 D_i + \phi_2 D_j + m(\xi_i, \xi_j, \delta) + U_{ij,t} \geq 0) \quad (7)$$

where vector $T_{ij,t}$ measures the distance between agent i and j based on their residence and workplace, $F_{ij,t-1}$ is a vector of dummies to discretise the number of contacts agents i and j share in common, Z_{ij} is a vector of dyad-specific time invariant covariates, D_i and D_j capture static differences in sociability based on both parties' logarithmised long-term degree, and $m(\xi_i, \xi_j, \delta)$ is a symmetric matching function of unobserved node specific heterogeneity.[6] We assume that $U_{ij,t}$ is independent and identically distributed and has mean zero such that we can estimate a linear probability model of the form:

$$g_{ij,t} = \beta g_{ij,t-1} + T'_{ij,t}\eta_1 + F'_{ij,t-1}\eta_2 + Z'_{ij}\rho + \phi_1 D_i + \phi_2 D_j + m(\xi_i, \xi_j, \delta) + U_{ij,t}. \quad (8)$$

In particular, the distance measures represented by vector $T_{ij,t}$ comprise the log travel time between agents i's and j's residence as well as a dummy for same workplace. The latter equals one if they predominantly use antennas within the same 5 km radius during business hours. We discretise the number of common friends, such that we obtain two dummy variables contained in $F_{ij,t-1}$: The first indicator equals one, if agents i and j share at least one common social contact, while the second indicators equals one if agents

[6]Note that the number of mutual contacts, $F_{ij,t-1}$, enters with a lag. This implies that agents form/maintain/dissolve links myopically, as if all features of the previous period's network remain fixed. Assuming this structure, eliminates contemporaneous feedback, which can problematic for inference (see Graham, 2014).

i and j share at least two common contacts.[7] The dyad-specific covariates in vector Z_{ij} include three dummy variables indicating same age, same gender and same language.

The model in (8) also accounts for matching based on unobservables as reflected by $m(\xi_i, \xi_j, \delta)$. Those that favourably match in terms of unobservable characteristics ξ feature a higher likelihood to form a link. These unobservables may bias our estimates of the cross sectional model in (8). If individuals with common unobservable attributes are more likely to cluster regionally and thus live closer together, our distance measure will be negatively correlated with the error term. A within-transformation will take out time invariant factors that affect the matching quality, i.e.

$$\begin{aligned} g_{ij,t} - \bar{g}_{ij} =& \beta(g_{ij,t-1} - \bar{g}_{ij}) + (T_{ij,t} - \bar{T}_{ij,t})'\eta_1 + (F_{ij,t-1} - \bar{F}_{ij})'\eta_2 + U_{ij,t} - \bar{U}_{ij}, \quad \text{or} \\ \ddot{g}_{ij,t} =& \beta\ddot{g}_{ij,t-1} + \ddot{T}_{ij,t}\eta_1 + \ddot{F}'_{ij,t-1}\eta_2 + \ddot{U}_{ij,t}. \end{aligned} \tag{9}$$

In equation (9) the transformed residual, $\ddot{U}_{ij,t}$, is necessarily correlated with the lagged dependent variable, $\ddot{g}_{ij,t-1}$, because both are a function of $\bar{U}_{ij}$. Therefore, OLS estimates of equation (9) are not consistent for the parameters of interest. We therefore follow Angrist and Pischke (2009) and estimate models including the lagged dependent variable but not the fixed effects, as in equation (10a.), and then compare the results to estimates obtained from a fixed effect regression without the dynamic component, as in equation (10b.):

$$\begin{aligned} &\text{a.} \quad g_{ij,t} = \beta g_{ij,t-1} + T'_{ij,t}\eta_1 + F'_{ij,t-1}\eta_2 + Z'_{ij}\rho + \phi_1 D_i + \phi_2 D_j + U_{ij,t} \\ &\text{b.} \quad \ddot{g}_{ij,t} = \ddot{T}_{ij,t}\eta_1 + \ddot{F}'_{ij,t-1}\eta_2 + \ddot{U}_{ij,t}. \end{aligned} \tag{10}$$

These two models have a useful bracketing property, that bounds the causal effect of interest. With respect to the geographical distance between two agents, we expect that the fixed effect estimates are upwardly biased, while the lagged dependent model yields a downwardly biased estimate (see Angrist and Pischke, 2009, p.245–247). We also estimate equation (10a.) within a Logit framework in order to account for the dichotomous nature of the data.

A practical issue that arises with estimating the outlined network formation models is the size of the adjacency matrix that potentially includes $(2 \cdot 10^6)^2$ unique pairs of agents. It is neither computationally feasible to estimate the models based on all these pairs nor necessary for obtaining consistent estimates of the parameters of interest as is shown by Manski and Lerman (1977), and Cosslett (1981). Since we have complete information on the network we can use a stratified sample and adjust the estimates with the respective sampling weights. Our choice-based sample results from an endogenous stratified sampling scheme where each stratum is defined according to the individual responses, that is the

[7] We discretise the number of mutual friends, because the continuous measure yields imprecise (yet significant) estimates. We also tried specifications with three or more common friends dummies, which turned out insignificant.

binary values taken by the response variable $g_{ij,t}$.[8] This sampling structure requires the availability of prior information on the marginal response probabilities which is in our setting available due to the full observation of $\mathcal{G}_t$.

4.2 Network Topography

We estimate the effect of location characteristics on the individual-level network measures formally defined in section 2: degree, within-degree, and clustering coefficient. Below, we lay out the estimation strategy for degree centrality noting that specifications for all other network measures follow analogously.

Following the earlier notation, the econometric models involve measures of degree centrality, D_{it}, as dependent variable and location specific covariates at the place of residence denoted by L_r. Hence, we specify the benchmark model as

$$D_{ir,t} = \alpha + L'_{r,t}\beta + X'_{ir,t}\gamma + \lambda_t + \lambda^l_r + \epsilon_{ir,t}, \tag{11}$$

where $X_{ir,t}$ is a vector of individual characteristics (i.e. commuting distance, language, dummy for belonging to language minority, gender, and age), λ_t stands for month fixed effects, and λ^l_r denotes language region fixed effects. The location vector $L_{r,t}$ includes indicators for EUROSTAT's harmonised definition of functional urban areas which distinguish between the urban core, the hinterland and peripheral regions. Alternatively, we measure local density using the number of private mobile phone customers within 15 minutes travel time from the respective place. Unlike municipal population statistics this measure has the advantage that it is independent from administrative boundaries.

In a next step, we address the issue of individual sorting on unobservables across locations. If the most sociable individuals systematically sort into high-density places, equation (11) would yield upwardly biased estimates of the density externality. Compared to the pooled OLS specification, we add an individual fixed effect in order to disentangle the density externality and the sorting effect, i.e.

$$D_{ir,t} = \mu_i + L'_{r,t}\beta + X'_{ir,t}\gamma + \lambda_t + \lambda^l_r + \epsilon_{ir,t}. \tag{12}$$

Note that this model identifies the effects on the basis of movers i.e. those who changed their place of residence between July 2015 and April 2016. These are about 147'000 individuals in the unfiltered data or 6% of the operator's private customers (see Table A.4). One concern in introducing fixed effects is that movers may differ systematically from the population. Like reported in other studies that adopt a similar identification strategy (e.g. D'Costa and Overman, 2014), movers in our data are on average younger than non-

[8] The main motivation behind this approach is usually the possibility of oversampling rare alternatives, which can improve the accuracy of the econometric analysis but also reduce survey costs. However, in our case we undersample those dyads with $g_{ij,t} = 0$ in order to enhance computational efficiency. One disadvantage is that most specification tests for non-linear models are not computable with sample weights.

movers. Apart from age, Table A.5 shows that differences in both individual characteristics as well as phone usage behaviour and network properties are sufficiently small between both groups.

5 Results

The results section is structured as follows: We begin by discussing the main results for the network formation model. Our focus is on the question of whether distance is costly to social interactions (section 5.1). In a second step, we analyse differences in network size across regions, to test the hypothesis that cities promote social interactions (section 5.2). We then proceed to investigate, whether population density affects the efficiency of networks. To draw conclusions regarding network efficiency, we analyse the perimeter of social networks (section 5.3), examine regional differences in matching quality (section 5.4), and finally aim to gain insights regarding clustering (section 5.5).

5.1 The Role of Distance and Other Determinants in Tie Formation

It is instructive to begin by looking at plain descriptives. Figure 4 plots the share of ties along the share of potential contacts by radius. Considering that almost 50 percent of bilateral ties are formed within a 5 km perimeter that covers on average less than 1 percent of the population, this illustrates the rapid decline of social interactions across space.

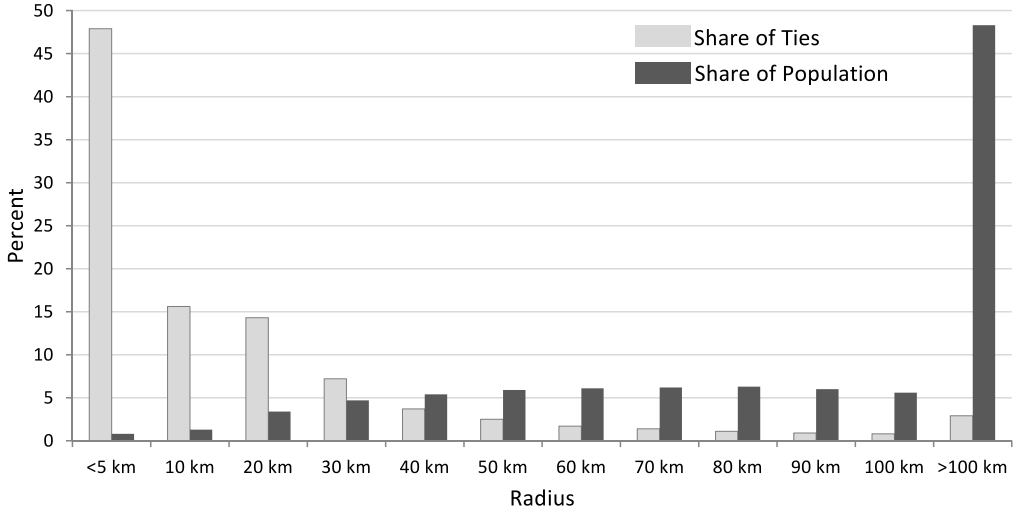


Figure 4: Share of Social Ties and Population by Radius

Of course, this approach does not account for biases due to spatial sorting of similar types. We therefore proceed to the network formation models, outlined in the previous section. Table 2 presents the result for the linear probability model. All coefficients were multiplied by 10'000 and therefore can be interpreted as basis points. This means that a

Table 2: LPMs of Network Formation, Monthly Data for June 2015–May 2016.

	Pooled OLS		Panel FE		Lagged Dependent Var.	
	(1)	(2)	(3)	(4)	(5)	(6)
Ln(Travel Time$_{ij,t}$)	-0.112***	-0.942***	-0.024***	-0.094***	-0.053***	-0.479***
	(0.000)	(0.053)	(0.000)	(0.019)	(0.000)	(0.024)
Ln(Travel Time$_{ij,t}$)^2		0.104***		0.010***		0.053***
		(0.006)		(0.002)		(0.003)
Same Workplace$_{ij,t}$		0.166***		0.071***		0.100***
		(0.030)		(0.002)		(0.014)
Same Language$_{ij,t}$		0.017***				0.009***
		(0.001)				(0.001)
>0 Common Contacts$_{ij,t-1}$		213.822***		11.840***		100.943***
		(10.101)		(0.928)		(4.866)
>1 Common Contacts$_{ij,t-1}$		2257.176***		145.633***		1024.429***
		(331.296)		(35.656)		(159.448)
g$_{ij,t-1}$					5231.433***	4973.641***
					(2.929)	(34.689)
Const.	0.545***	2.079***	0.135***	0.224***	0.256***	1.060***
	(0.001)	(0.114)	(0.002)	(0.038)	(0.000)	(0.052)
R^2	0.001	0.054	0.001	0.001	0.275	0.288
Further Controls	No	Yes	No	No	No	Yes
Pair FE	No	No	Yes	Yes	No	No
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Groups	–	–	2,584,869	2,582,702	–	–
Observations	30,996,082	27,238,673	30,996,082	27,238,673	28,411,817	27,238,673

Notes: The *sample* covers movers who used their phone every month at least once. All *coefficients* of the linear probability models are multiplied by 10000, and therefore can be interpreted as basis points. *Further controls* include the degree for both agents (log), dummies for same gender and same age, as well as the absolute age difference between agents i and j. Standard errors in parentheses. + p<0.10, * p<0.05, ** p<0.01 *** p<0.001.

coefficient equalling one translates to a marginal increase in $P(g_{ij,t} = 1)$ of a hundredth percentage point. The first two columns display pooled OLS estimations, the middle columns report pair fixed effects models, and the last two columns show lagged dependent variable specifications. In all models estimated, the travel time between two agents enters negatively, implying that *distance* is indeed costly when forming and maintaining a link. Columns (2), (4) and (6) reveal that tie formation is actually a convex function in distance; the log of travel time enters strongly negative, while the squared term is positive. Their relative magnitudes suggest that the negative effect of distance completely fades at approximately 90 minutes driving distance.

In addition to being neighbours, *working in the same area* also increases the likelihood that two persons form a link. The coefficient for the dummy variable "Same Workplace", which equals one if agents i and j predominantly use antennas within the same 5 km radius during business hours, ranges between 0.07 and 0.1. Hence, working in close proximity increases the probability of forming a tie by about 0.1 basis point, which is roughly ten times the estimated effect of speaking the same principal language. Taken together, distance in terms of both residence and workplace are costly to social interactions.

In order to analyse preferences for triadic closure or *clustering*, we discretise the number of common friends, such that we obtain two dummy variables: one indicating that agents i and j share at least one common social contact, and the other indicating that they share

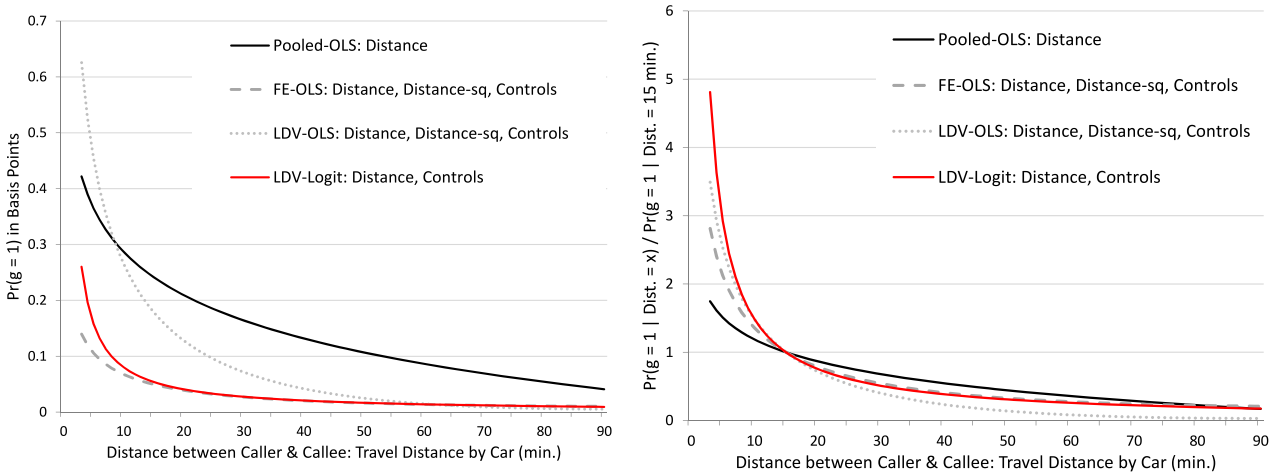
Table 3: Logit Models of Network Formation, Monthly Data for June 2015–May 2016.

	Pooled Logit		Lagged Dependent Var.	
	(1)	(2)	(3)	(4)
Ln(Travel Time$_{ij,t}$)	-1.410***	-0.877***	-1.131***	-0.976***
	(0.002)	(0.049)	(0.001)	(0.005)
Same Workplace$_{ij,t}$		0.893***		1.085***
		(0.161)		(0.013)
Same Language$_{ij,t}$		1.813***		1.685***
		(0.057)		(0.005)
>0 Common Contacts$_{ij,t-1}$		7.363***		4.786***
		(0.122)		(0.070)
>1 Common Contacts$_{ij,t-1}$		2.323***		-0.018
		(0.352)		(0.071)
g$_{ij,t-1}$			12.218***	9.868***
			(0.003)	(0.029)
Const.	-7.357***	-12.951***	-8.958***	-11.170***
	(0.010)	(0.249)	(0.007)	(0.026)
Pseudo R^2	0.138	0.376	0.492	0.527
Further Controls	No	Yes	No	Yes
Pair FE	No	No	No	No
Month FE	Yes	Yes	Yes	Yes
Observations	30,996,082	27,238,673	28,411,817	28,411,817

Notes: The *sample* covers movers who used their phone every month at least once. *Further controls* include the degree of both agents (log), dummies for same gender and same age, as well as the absolute age difference between agents i and j. Standard errors in parentheses. + p<0.10, * p<0.05, ** p<0.01 *** p<0.001.

at least two common contacts. The coefficients for both "Common Contact" variables are highly significant. Column (2) shows that the probability of forming a link with another person increases by up to 22 percentage points, if one shares at least two common contacts. As one would expect, the estimates are considerably smaller in column (4), which controls for matching quality by employing dyad-specific fixed effects. Nonetheless, the additional link-surplus of 1.5 percentage points due to triangular relations – as obtained in the most conservative specification – is quantitatively substantial. Agents clearly value triadic relations, which explains the evidently non-random clustering in this network, as discussed in section 3.1.

Overall *matching quality* between two agents is not observable, but the regressions in column (2) and column (6) account for socio-demographic (dis)similarities that are incorporated in the matching concept, namely dummies for same language, same gender and same age, as well as the absolute age difference between customers i and j. If we abstract from potential omitted variable bias and assume that $m(\cdot)$ is a linear and additive function, the interpretation of the estimated coefficients in terms of matching is as follows: By definition $\frac{\partial E[g_{ij}|m(\cdot)]}{\partial m(\cdot)} > 0$, therefore $sign(\hat{\rho_q}) = sign(\delta_q)$ holds. Accordingly, a positive (negative) sign not only implies an increase in the probability that two agents socially interact, but also a positive (negative) relation in terms of matching quality. Our results unambiguously point toward homophily, which is the well documented tendency of individuals to bond with similar others (e.g. Currarini, Jackson and Pin, 2009; McPherson, Smith-Lovin and Cook, 2001). For instance, individuals who share the same principal language are on average more likely to form a tie than individuals with different language



(a) Level

(b) Ratio, Base Distance = 15 min.

Figure 5: Probability to Form a Tie: Prediction Based on OLS and Logit Models
Notes: Same Workplace=0, Common Contacts=0, Degree=mean, Same Gender=1, Same Age=1, Age Diff=0, $g_{ij,t-1}$=0, FE=0.

preferences. The same holds true regarding age and gender (results not shown).

The OLS results suggest that spatial proximity, the presence of common friends, and demographic similarity increase the likelihood that two individuals interact. We also estimate *Logit models* to accommodate for the binary dependent variable and check the robustness of these results. Since the incidental parameter problem can induce severe bias in the Logit fixed effects estimates (e.g Lancaster, 2000), we only show results for the pooled Logit model and the lagged dependent variable model. Moreover, the squared distance term is excluded due to convergence issues, which should be a minor problem given the Logit estimator's inherent non-linearity. The results in Table 3 are qualitatively almost identical to the OLS results in Table 2, except for one of the common friends dummies, which turns out insignificant in column (4). Hence, in terms of qualitative interpretation the main results are very robust with respect to modelling choice.

We now inspect the functional relation between distance and tie formation in more detail. Figure 5a displays the predicted probability for $g_{ij} = 1$ based on various specifications. Figure 5b plots the relative probability for $g_{ij} = 1$ compared to the base probability at a distance of 15 minutes travel time. Although the models differ somewhat regarding the level prediction, they consistently reveal a convexly decreasing relation between link formation and distance. Overall, the graphs illustrate that the effect of distance on social interactions is highly localised; the probability of forming a link is about twice as large for neighbours than for people living 10 minutes apart. This probability continues to fall quickly up to a distance of 30 minutes, beyond which the negative effect of travel time flattens out.

Summarising this comprehensive evidence, we have been able to demonstrate that distance is highly detrimental to social interactions. If distance between two individuals

did not impose costs on their social exchange, it would be difficult to argue that regional differences in population density should impact the topography of social networks. In such a – with respect to distance – frictionless world, cities and rural villages would offer an identical environment for social interactions as all people could choose from the same pool of potential contacts without there being any costs due to remoteness.

In what follows, we examine whether distance costs indeed lead to significant differences in the topography of social networks across urban and rural areas. First, we examine the consequences regarding network size, and then we turn our attention to network efficiency.

5.2 Cities and Network Size

A number of urban economic theories argue that cities are favourable to social interactions and support larger networks. As discussed in section 2, the underlying idea is that people living in densely populated areas encounter more potential contacts, and accordingly establish a larger number of valuable social ties. So far, we have presented evidence suggesting that distance is indeed costly to forming and maintaining a tie, which is a necessary condition for the hypothesis of larger networks in cities.

In order to directly test the hypothesis, we estimate a series of pooled OLS models, which are reported in Table 4. We use two sets of key explanatory variables, including the trichotomous classification for urbanisation by EUROSTAT (i.e. urban core, hinterland, periphery) as well as a continuous measure for population density. The latter is defined as the log of the population living within a 15-minute radius of an individual's postcode area. Network size is measured on a monthly basis as degree centrality, i.e. the number of unique contacts an individual calls during one month.

Columns (1) and (4) contain the results for the discretised measure of urbanisation, the former excluding and the latter accounting for individual controls in the regression. Agents who live in the hinterland or periphery have on average a smaller network than city residents. The correlations are statistically highly significant, with an average difference of -1.1 to -1.7 percent when comparing the urban core to the periphery, and -2.4 to -2.5 percent when comparing the urban core to the hinterland.

The continuous population density measure in columns (2) and (5) is negatively correlated with network size. This unexpected result is due to non-linearities, as the results in columns (1) and (4) already indicate; although the hinterland has a higher population density than peripheral municipalities, the hinterland coefficient is significantly smaller than the periphery coefficient. When a squared-term is included, the results indeed reveal a convex relation between population density and network size, with the marginal effect of population density turning positive around its mean value.

Overall, these findings lend support to the hypothesis that dense urban areas facilitate social interactions, backing earlier studies that report a positive correlation between the level of urbanisation, the volume of phone calls, and network size (e.g. Charlot and Du-

Table 4: Regional Differences in Network Size, Monthly Data for June 2015–May 2016.

	Pooled OLS					
	(1)	(2)	(3)	(4)	(5)	(6)
Hinterland (vs. Cities)	-0.024***			-0.025***		
	(0.001)			(0.001)		
Periphery (vs. Cities)	-0.011***			-0.017***		
	(0.001)			(0.001)		
Ln(Pop. Density)		-0.012***	-0.223***		-0.008***	-0.222***
		(0.000)	(0.002)		(0.000)	(0.002)
Ln(Pop. Density)^2			0.012***			0.012***
			(0.000)			(0.000)
R^2	0.011	0.012	0.013	0.067	0.067	0.068
Further Controls	No	No	No	Yes	Yes	Yes
Language Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	10,117,645	10,117,522	10,117,522	9,353,794	9,353,679	9,353,679

Notes: The *sample* covers customers who used their phone every month at least once. *Further controls* include commuting distance, language, dummy for belonging to language minority, gender, and age. Standard errors in parentheses. + p<0.10, * p<0.05, ** p<0.01 *** p<0.001.

ranton, 2004; Schläpfer et al., 2014). So far it is unclear, however, whether the effect has a causal interpretation or is driven by the sorting of high sociability types to urban centres.

In a next step, the regressions include individual fixed effects to back out any person specific characteristics and thereby eliminate the sorting channel. Consequently, inference is now based on customers who changed their billing address during the 12 months period covered. Columns (1) to (3) of Table 5 display results for the baseline fixed effects regression, while columns (4) to (6) show a robustness check based on people who changed their residence by at least 30 minutes driving time. The results stand in stark contrast to the plain OLS regressions and clearly reject the hypothesis that cities have a causal impact on network size. All coefficients related to regional differences in population density are practically zero and statistically insignificant.

Figure B.1 in the appendix plots the degree of movers over time. It shows that agents expand their social network in the three months prior to moving, and then revert to their initial level within two months. To test the robustness of our results with respect to this dynamic, we re-estimate the fixed effects models for movers who changed their residence by at least 30 minutes driving time and successively exclude periods around the moving month. Table B.1 in the appendix shows that only 2 out of 20 coefficients turn out statistically significant at the 10 percent level. Hence, these additional results do not alter the conclusion from the benchmark analysis in Table 5.[9]

[9]One further concern may be that urban dwellers use messenger apps more frequently than people in rural areas, which could lead to a downward bias in the population density / city dummy estimates. Although we can not rebut such concerns with absolute certainty, they seem unsubstantiated for two reasons. *First*, messenger apps and mobile phone calls are most likely *complements* not *substitutes*. We decompose messenger usage along gender and language region, based on a survey conducted by *comparis.ch* in 2014. It shows that messenger apps are more often used among men than women and are more widespread in French-speaking than German-speaking regions. The same ranking unfolds for network size. If anything, this indicates that the two media are complements not substitutes. Additionally, a paper on workplace communication by Charlot and Duranton (2006) shows that telephone usage is complementary

Table 5: Regional Differences in Network Size, Monthly Data for June 2015–May 2016.

	FE: Full Sample			FE: Moving Distance > 30min.		
	(1)	(2)	(3)	(4)	(5)	(6)
Hinterland (vs. Cities)	0.000			-0.006		
	(0.003)			(0.006)		
Periphery (vs. Cities)	0.000			-0.001		
	(0.004)			(0.007)		
Pop. Density		-0.002	-0.001		-0.002	-0.006
		(0.001)	(0.012)		(0.002)	(0.017)
Pop. Density^2			-0.000			0.000
			(0.001)			(0.001)
R^2	0.011	0.011	0.011	0.011	0.011	0.011
Further Controls	Yes	Yes	Yes	Yes	Yes	Yes
Language Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Groups	60,514	60,514	60,514	16,874	16,874	16,874
Observations	669,825	669,825	669,825	185,676	185,663	185,663

Notes: The *sample* covers movers who used their phone every month at least once. *Further controls* include commuting distance and a dummy for belonging to language minority. Standard errors in parentheses. + p<0.10, * p<0.05, ** p<0.01 *** p<0.001.

It seems, then, that the correlation between population density / urbanisation and network size is fully driven by the sorting of above-average sociable people to the urban core and cannot be attributed to the positive externalities of people living close together. A variance decomposition, which computes the contributions of individual fixed effects, local fixed effects, and time specific factors to the total variance of $D_{i,t}$, also supports the conclusion that regional differences play a small role in explaining differences in network size. Individual components contribute 73.0 percent to the overall variance of degree centrality, while local factors only explain 2.3 percent. The remaining variation can be attributed to time specific factors (0.3%) and to the residual (24.4%), i.e. individual and time variant components.

This analysis provides evidence that the correlation between population density and network size is primarily driven by the sorting of highly sociable people to urban centres. Sociability may thereby refer to the mental capability of maintaining social ties, as suggested by the social brain hypothesis (e.g. Dunbar, 1998), and/or to personality traits, as advocated by Asendorpf and Wilpers (1998). This raises the question of why people with an above-average sociable predisposition move to cities. One evident explanation could be that cities provide a favourable environment for social interactions, which does not manifest itself in terms of network size but rather with respect to network efficiency. If this were the case, individuals with a preference for and capability of maintaining large networks would disproportionally benefit from moving to cities, which could explain the sorting pattern uncovered in the above analysis.

to all other modes of communication studied, including face-to-face communication, letter correspondence, email traffic, and internet usage. *Second*, we conduct a series of robustness checks, in which we control for an individual's communication preference based on his monthly text message–call volume ratio. These robustness checks do not alter the results.

5.3 Cities and the Perimeter of Social Networks

We begin the discussion of network efficiency by examining variations in network perimeters across regions. Everything else being equal, an agent is better off the less distant his social contacts live, simply because he will incur lower travel costs. Since people residing in cities have a larger pool of potential contacts within close proximity, one would expect them to recruit their social contacts within a narrower perimeter to minimise travel costs.

We analyse the impact of population density on the perimeter of an individual's network in three steps. First, we discuss descriptive evidence based on a density plot for social ties by radius and location type (i.e. cities versus hinterland/periphery). Second, we use the network formation model to test whether urban dwellers value distance differently than people living in less densely populated areas. Third, we calculate the within-degree, which measures network size within a radius of 15 minutes around an agent's residence, to analyse whether it varies systematically with population density.

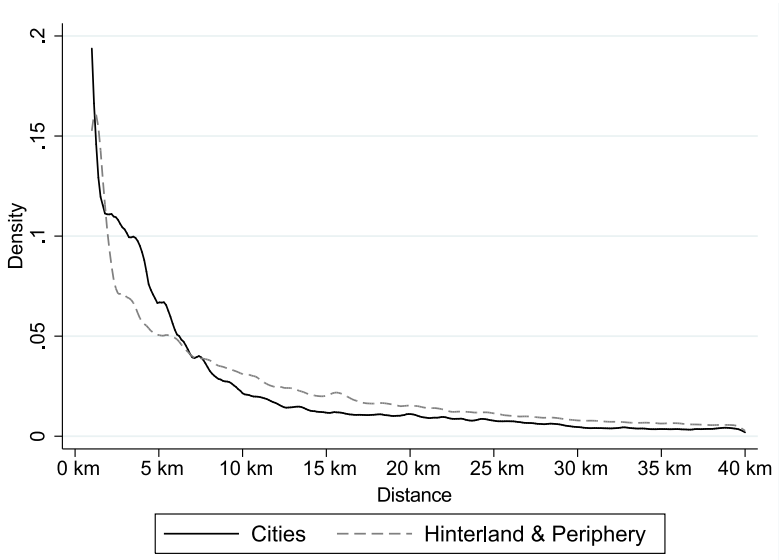


Figure 6: City versus Hinterland / Periphery – Density Plot for Social Ties by Radius.
Notes: The density plot starts at 1 km; links spanning shorter distances (mostly links within the same postcode) were assigned a value of 1 km.

Figure 6 plots the density of social ties by radius and location type. In comparison to individuals living in the hinterland or periphery, urban dwellers evidently have a larger mass of social contacts within a 7 km radius, and fewer contacts beyond. This supports the hypothesis that living in a city can lower the costs incurred from social interactions with distant contacts.

In order to examine this claim further we use the network formation model and interact the log of distance with either population density or the city dummy. The top panel of Table 6 reports the output of the augmented network formation model, with columns (1) and (2) displaying the OLS results and columns (3) and (4) showing the pair fixed effect

Table 6: Regional Differences in the Perimeter of Social Networks, Monthly Data for June 2015–May 2016.

a. Network Formation	Pooled OLS		Panel FE	
Cities & Distance	(1)	(2)	(3)	(4)
Ln(Distance$_{ij,t}$)	-0.068***	-0.069***	-0.016***	-0.016***
	(0.003)	(0.003)	(0.001)	(0.001)
Ln(Distance$_{ij,t}$) × City$_{i,t}$	-0.001***		-0.001**	
	(0.000)		(0.000)	
Ln(Distance$_{ij,t}$) × Pop. Density$_{i,t}$		-0.001***		-0.001***
		(0.000)		(0.000)
R^2	0.054	0.054	0.001	0.001
Further Controls	Yes	Yes	Yes	Yes
Pair FE	No	No	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Groups	–	–	2,582,702	2,582,702
Observations	27,238,673	27,238,673	27,238,673	27,238,673
b. Network Topography	Pooled OLS		Panel FE	
Cities & Within-Degree (15 min.)	(1)	(2)	(3)	(4)
Hinterland (vs. Cities)	-0.111***		-0.123***	
	(0.001)		(0.010)	
Periphery (vs. Cities)	-0.208***		-0.231***	
	(0.001)		(0.012)	
Population Density		0.086***		0.143***
		(0.000)		(0.004)
R^2	0.049	0.056	0.010	0.018
Further Controls	Yes	Yes	Yes	Yes
Individual FE	No	No	Yes	Yes
Language Region FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Groups	–	–	60,514	60,514
Observations	9,353,794	9,353679	669,825	669,812

Notes: The *sample* covers movers who used their phone every month at least once. *a. Controls in network formation models:* Dummies for same workplace, same language, common contacts, degree of both agents (pooled OLS), same gender (pooled OLS), same age (pooled OLS), and the absolute age difference between agents i and j (pooled OLS). *b. Controls in network topography models:* Commuting distance, language minority dummy, gender (pooled OLS) and age (pooled OLS). Standard errors in parentheses. + p<0.10, * p<0.05, ** p<0.01 *** p<0.001.

estimates. All specifications suggest that urban dwellers incorporate distance costs more strongly in their valuation than people living in peripheral areas. The interaction terms yield statistically significant negative effects, but are quantitatively relatively small.

Finally, we resort to our network topography model using the within-degree, DW_i^r, as dependent variable. The bottom panel of Table 6 reports the outputs of this approach, with columns (1) and (2) displaying the OLS results and columns (3) and (4) showing the individual fixed effect estimates that account for the sorting of highly sociable individuals to urban areas. As hypothesised, the within-degree is largest in urban areas and positively correlated with population density. This holds true for both the plain OLS estimates, as well as the individual fixed effects results. According to our causal estimates from the individual fixed effects specification, urban dwellers have on average a 10 percent larger within-degree than individuals residing in the hinterland, and a 23 percent larger within-degree than people living in peripheral areas. The results also show that doubling

population density leads, on average, to a 6.8 percent higher within-degree.[10] While population density is hardly relevant for overall network size, it has considerable explanatory power regarding the number of close-range contacts. The variance decomposition also reveals that regional factors explain more than twice as much of the within-degree variance (4.9%) than the variance in network size (2.3%).

Densely populated areas evidently shrink the perimeter of an individual's social network, in the sense that a larger fraction of her social contacts are likely to live in close proximity. Considering that distant social contacts are costly, this consequently suggests that urban dwellers bear fewer costs from social interactions than people living in sparsely populated areas. This could – at least partly – explain why sociable people sort into cities, as they disproportionally benefit from this channel and therefore have a higher willingness to pay for housing in cities than less sociable types. This result may also be interpreted as better matching in cities, because geographical distance is essentially one dimension of matching quality. We further explore matching quality across regions in the following section.

5.4 Cities and Matching

Since matching quality cannot be directly observed, we propose two indirect tests for the hypothesis that matching quality improves with population density. In one test we resort to the network formation model, while the second test is based on the network topography approach.

We begin with the *network formation model*, or more specifically with the fixed effect specification given in equation (10b.): The pair fixed effect absorbs any dyad-specific constant factors that either raise or lower the surplus of interaction for the involved agents. Hence, it primarily captures matching quality, $m(\cdot)$, which governs the value obtained from forming a link with another person. If agents living in cities indeed benefit from better matching quality, we would expect that fixed effects associated with their actually formed links are higher than the equivalent fixed effects calculated for agents living in rural areas. To test this claim, we first estimate equation (10b.), and then regress the predicted pair fixed effects for the subsample of active links (i.e. $g_{ij} = 1$) on population density at agent i's place of residence. Because we focus on movers to back out any distance-related effects, the estimates yield the impact of population density weighted by duration of stay.[11] We obtain strong positive and significant effects for population density in column (2), and negative effects for residents of peripheral municipalities in column (1). Restricting the sample to customers with a minimum driving distance of 30 minutes between their old and new addresses does not affect the results. This backs the claim that densely populated

[10] As for the degree, we re-estimate the fixed effects models and successively exclude periods around the moving month. Table B.2 in the appendix shows that this does not affect the results.

[11] As a robustness check, we also restrict the sample to movers who change their residence but stay within the same class of municipalities, i.e. moving from city to city or from hinterland to hinterland. As Table B.4 in the appendix reveals, this does not alter the results.

Table 7: Regional Differences in the Matching Quality

a. Network Formation	Full Sample		Moving Distance > 30min.	
Cities and Matching (Predicted FE)	(1)	(2)	(3)	(4)
Hinterland$_{i,t}$	-55.595***		-59.265***	
	(5.928)		(11.127)	
Periphery$_{i,t}$	-145.315***		-117.816***	
	(6.451)		(11.832)	
Pop. Density$_{i,t}$		34.954***		17.834***
		(1.856)		(2.815)
Constant	2492.994***	2085.201***	2466.674***	2232.319***
	(492.036)	(115.281)	(250.231)	(84.219)
R^2	0.001	0.001	0.001	0.001
Observations	11,616,147	11,692,984	3,089,595	3,116,907
b. Network Topography	Full Sample		Moving Distance > 30min.	
Cities and Matching (Social Adaption)	(1)	(2)	(3)	(4)
City$_{post}$	0.327***		0.090	
	(0.046)		(0.056)	
City$_{pre}$	-0.449***		-0.275***	
	(0.033)		(0.044)	
Pop. Density$_{post}$		0.227***		0.076***
		(0.011)		(0.013)
Pop. Density$_{pre}$		-0.326***		-0.138***
		(0.014)		(0.020)
Constant	1.009***	1.801***	0.685***	1.256***
	(0.097)	(0.176)	(0.037)	(0.194)
R^2	0.047	0.078	0.259	0.263
Further Controls	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes
Language Region FE	Yes	Yes	Yes	Yes
Observations	28,871	28,871	7,887	7,801

Notes: *Dependent Variable in Panel a.:* Predicted dyad specific fixed effect from network formation model outlined in equation (10b). *Dependent Variable in Panel b.:* The number of *post-move* contacts at the *post-move* place of residence over the number of *post-move* contacts at the *pre-move* place of residence. *Controls in Panel b.:* Number of contacts at new address prior to moving, commuting distance, dummy for belonging to language minority, gender and age . Standard errors in parentheses. + p<0.10, * p<0.05, ** p<0.01 *** p<0.001.

areas lead to favourable matching outcomes.

In the next step, we reassess the hypothesis by returning to the *network topography approach*. If people change their residence, we would expect them to keep up with some of their previous contacts and replace others with individuals living in their new neighbourhood. Since distance makes social interactions costly, only highly valuable contacts at the old place of residence are worthwhile to maintain. Furthermore, if one encounters very good matches at the new place of residence, the replacement of pre-existing ties with new contacts should advance more quickly. We therefore examine whether this social adjustment process systematically varies with population density at the pre- and post-move residence. Specifically, we estimate

$$\frac{DW_{i,post}^{r_{post}}}{DW_{i,post}^{r_{pre}}} = \alpha + \beta_{r_{post}} \cdot L_i^{r_{post}} + \beta_{r_{pre}} \cdot L_i^{r_{pre}} + X_i'\gamma + \varrho DW_{i,pre}^{r_{post}} + \epsilon_i, \tag{13}$$

where the ratio $DW_{i,post}^{r_{post}}/DW_{i,post}^{r_{pre}}$ reflects the number of *post-move* contacts at the

post-move place of residence over the number of *post-move* contacts at the *pre-move* place of residence. The main explanatory variables are population density and the trichotomous classification for urbanisation at mover i's new address ($L_i^{r_{post}}$) and old address ($L_i^{r_{pre}}$), complemented with a measure for the number of pre-move contacts at the new address ($DW_{i,pre}^{r_{post}}$), and individual level characteristics X_i.[12] The results reported in the bottom panel of Table 7 are based on address changes between October 2015 and January 2016, a pre-move window covering June 2015 to August 2015, and a post-move window covering March 2016 to May 2016. As hypothesised, the fastest social adjustment process is observed for people who move from the periphery to the city, while movers who lived in urban areas before changing their address keep comparatively large shares of their pre-move contacts. Since maintaining spatially distant contacts is costly, this suggests that contacts formed in cities generate on average a higher surplus and are therefore more likely to be maintained. Hence, this test further supports the hypothesis that densely populated areas improve matching quality.

So far, our results suggest that high population density in cities does not lead to *larger* social networks, but rather improves their efficiency in terms of narrower perimeters and matching quality. We are not aware of any paper providing evidence on regional differences in social matching quality, which is a key factor underlying the main agglomeration forces as formally discussed in Duranton and Puga (2004).

5.5 Cities and Clustering

The final network property that we examine is clustering. Agents face a trade-off in terms of efficient information exchange (i.e. low clustering) and benefits related to reciprocity (i.e. high clustering). The optimal balance may vary regionally due to factors that alter this trade-off. Additionally, one would expect that more populous neighbourhoods display lower average clustering, simply because randomly established links are less likely to form triadic structures when the pool of potential contacts grows larger. To test the first claim, we resort to the network formation model. Even if there is no evidence that urban dwellers value triadic relations differently than people living in rural communities, the mechanical relation between population density and clustering may lead to measurable regional differences. If this is the case, the network topography approach should be able to uncover them.

We begin with the network formation model and interact the dummy for common contacts with either population density or the city dummy. In order to back out spurious clustering due to the grouping of similar types, we focus on the pair fixed effects specification. The top panel of Table 8 reports the results for these regressions. In both specifications, the interaction terms are negative and statistically significant at the 10

[12] Instead of controlling for the pre-move contacts at the new address, we also re-estimate the model for a subsample of customers that move to a location where they have no prior contacts, i.e. $DW_{i,pre}^{r_{post}} = 0$. This does not alter the conclusion, as the results in Table B.4 (Panel b.) in the appendix show.

Table 8: Regional Differences in the Transitivity of Social Networks, Monthly Data for June 2015–May 2016.

a. Network Formation			Panel FE	
Cities & Common Friends			(3)	(4)
> 0 Common Contacts$_{ij,t-1}$			17.337***	16.477***
			(1.268)	(1.058)
> 0 Common Contacts$_{ij,t-1}$ × City$_{i,t}$			-3.681^{+}	
			(2.009)	
> 0 Common Contacts$_{ij,t-1}$ × Pop. Density$_{i,t}$				-1.745^{+}
				(0.957)
R^2			0.001	0.001
Further Controls			Yes	Yes
Pair FE			Yes	Yes
Month FE			Yes	Yes
Groups			2,582,702	2,582,702
Observations			27,238,673	27,238,138
b. Network Topography	Pooled OLS		Panel FE	
Cities & Clustering	(1)	(2)	(3)	(4)
Hinterland (vs. Cities)	0.010***		0.002*	
	(0.001)		(0.001)	
Periphery (vs. Cities)	0.014***		0.002**	
	(0.001)		(0.001)	
Population Density		-0.004***		-0.001**
		(0.001)		(0.000)
R^2	0.022	0.022	0.001	0.001
Further Controls	Yes	Yes	Yes	Yes
Individual FE	No	No	Yes	Yes
Language Region FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Groups	–	–	60,507	60,507
Observations	9,252,183	9,252,183	664,343	664,330

Notes: The *sample* covers movers who used their phone every month at least once. *a. Controls in network formation models:* Dummies for same workplace, same language, common contacts, degree of both agents (pooled OLS), same gender (pooled OLS), same age (pooled OLS), and the absolute age difference between agents i and j (pooled OLS). *b. Controls in network topography models:* Commuting distance, dummy for belonging to language minority, gender (pooled OLS) and age (pooled OLS). Standard errors in parentheses. $^{+}$ p<0.10, * p<0.05, ** p<0.01 *** p<0.001.

percent level. Hence, this analysis suggests that sharing a common link is valued less by urban dwellers than by residents of peripheral areas. Magnitude wise the impact is fairly substantial, as it amounts to approximately 20 percent of the effect attributed to the common contact dummy. While sharing a common contact increases the probability of forming and maintaining a link by 17.3 basis points, the effect is only 13.7 basis points among city residents.

Given the results of the network formation analysis, we expect lower clustering in cities than in peripheral areas. First, city residents seem to value triadic relations less than people living in peripheral areas. Second, the larger pool of potential contacts lowers the likelihood of triadic relations, at least if there is some randomness in the link formation process. The bottom panel of Table 8 displays the results of the network topography analysis with clustering as the dependent variable. Both the pooled OLS regressions in columns (1) and (2), as well as the fixed effects specifications in columns (3) and (4)

suggest that cities attenuate network clustering. The effect ranges between -0.010 and -0.014 in the pooled OLS regressions, which is roughly 11 to 15 percent of the sample mean. The difference between city and hinterland / periphery drops by 80 percent in the fixed effects specifications, but remains significant at the 5 percent level or higher.[13]

Despite the evidence that population density has no impact on the number of social interactions, cities may facilitate the diffusion of information due to below-average clustering. This can have important consequences for local labour markets, as discussed in Sato and Zenou (2015), for example. In conjunction with the findings on network size, matching quality and distance costs, this suggests that cities may encourage not a larger number but rather more valuable social interactions.

6 Conclusion

The results of this study suggest that that cities provide a superior environment for social interactions, which is fundamentally important to the mechanics of classic agglomeration forces. Contrary to many theoretical models, the advantages of densely populated areas do not translate into larger social networks but rather into improvements in terms of matching quality, smaller distance costs, and a favourable structure for information diffusion (i.e. lower clustering).

Evidently, modern communication technologies do not render cities obsolete. Our analysis has illustrated that they remain important as catalyst of valuable social exchange and, consequently, as potential engines of growth. From a policy perspective, this result provides micro-level evidence for the positive externalities of densely populated areas, which should be taken into account, for example, in the design of zoning policies, or the pricing of mobility.

There are many potential extensions of the work described in this paper. First and foremost, a quantification of the effects in monetary terms would be insightful, and – in our opinion – would be the first attempt to plausibly identify the causal link between density, social interactions, and a measure of productivity / output. Second, we focused exclusively on private social interactions, thus it would be fruitful to examine whether the same conclusions apply to networks from business communication. Third, we barely scratched the surface of the information available in the mobility data recorded from transmitting antennas. Such data would allow, for instance, to thoroughly test the influential claim by social scientist Robert Putnam (2000) that commuting causes an erosion of social capital.

[13] Successively excluding periods around the moving month, as done in Table B.3, yields occasionally insignificant results for the Hinterland dummy, but overall the same pattern as in the benchmark model emerges.

References

Abel, Jaison and Richard Deitz. 2015. "Agglomeration and Job Matching among College Graduates." *Regional Science and Urban Economics* 51:14–24.

Alesina, Alberto and Eliana La Ferrara. 2000. "Participation in Heterogeneous Communities." *Quarterly Journal of Economics* 115(3):847–904.

Ali, Nageeb and David Miller. 2009. "Enforcing Cooperation in Networked Societies." Society for Economic Dynamics, *mimeo.*

Ambrus, Attila, Markus Mobius and Adam Szeidl. 2014. "Consumption Risk-Sharing in Social Networks." *American Economic Review* 104(1):149–182.

Angrist, Joshua and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics. An Empiricist's Companion.* Princeton: Princeton University Press.

Asendorpf, Jens and Susanne Wilpers. 1998. "Personality Effects on Social Relationships." *Journal of Personality and Social Psychology* 74(6):1531–1544.

Berliant, Marcus, Robert Reed and Ping Wang. 2006. "Knowledge Exchange, Matching, and Agglomeration." *Journal of Urban Economics* 60:69–95.

Blondel, Vincent, Adeline Decuyper and Gautier Krings. 2015. "A Survey of Results on Mobile Phone Datasets and Analysis." *EPJ Data Science* 4:1–55.

Brueckner, Jan and Ann Largey. 2008. "Social Interaction and Urban Sprawl." *Journal of Urban Economics* 64:18–34.

Burley, Jessica. 2015. "The Built Environment and Social Interactions: Evidence from Panel Data." University of Toronto, *mimeo.*

Burt, Ronald. 1987. "A Note on Strangers, Friends and Happiness." *Social Networks* 9(4):311–331.

Cairncross, Frances. 2001. *The Death of Distance: How the Communication Revolution Is Changing Our Lives.* Cambridge: Harvard Business School Press.

Charlot, Sylvie and Gilles Duranton. 2004. "Communication Externalities in Cities." *Journal of Urban Economics* 56:581–631.

Charlot, Sylvie and Gilles Duranton. 2006. "Cities and Workplace Communication: Some Quantitative French Evidence." *Urban Studies* 43(8):1365–1394.

ComCom, Eidgenössische Kommunikationskommission. 2015. "Tätigkeitsbericht der ComCom 2015." Published online `http://www.comcom.admin.ch/dokumentation/00564/index.html?lang=de` (01.06.2016).

Cosslett, Stephen. 1981. "Maximum Likelihood Estimator for Choice-Based Samples." *Econometrica* 49:1289–1316.

Currarini, Sergio, Matthew Jackson and Paolo Pin. 2009. "An Economic Model of Friendship: Homophily, Minorities, and Segregation." *Econometrica* 77(4):1003–1045.

D'Costa, Sabine and Henry Overman. 2014. "The Urban Wage Growth Premium: Sorting or Learning." *Regional Science and Urban Economics* 48:168–179.

Dunbar, Robin. 1992. "Neocortex Size as a Constraint on Group Size in Primates." *Journal of Human Evolution* 20:469–493.

Dunbar, Robin. 1993. "Coevolution of Neocortical Size, Group Size and Language in Humans." *Behavioral and Brain Sciences* 16:681–735.

Dunbar, Robin. 1998. "The Social Brain Hypothesis." *Evolutionary Anthropology* 9(10):178–190.

Dunbar, Robin. 2015. "Do Online Social Media Cut Through the Constraints that Limit the Size of Offline Social Networks?" *Royal Society Open Science* 3:1–9.

Duranton, Gilles and Diego Puga. 2004. Micro-Foundations of Urban Agglomeration Economies. In *Handbook of Regional and Urban Economics*, ed. John Vernon Henderson and Jacques Thisse. Amsterdam: Elsevier pp. 2063–2115.

Fischer, Claude. 1982. *To Dwell among Friends. Personal Networks in Town and City.* Chicago: Chicago University Press.

Gaspar, Jess and Edward Glaeser. 1998. "Information Technology and the Future of Cities." *Journal of Urban Economics* 43:136–156.

Glaeser, Edward. 1999. "Learning in Cities." *Journal of Urban Economics* 46:254–277.

Graham, Bryan. 2014. "Methods of Identification in Social Networks." NBER Working Paper No. 20414.

Granovetter, Mark. 1973. "The Strength of Weak Ties." *American Journal of Sociology* 78(6):1360–1380.

Granovetter, Mark. 2005. "The Impact of Social Structure on Economic Outcomes." *Journal of Economic Perspectives* 19(1):33–50.

Gui, Benedetto and Robert Sugden. 2005. Why Interpersonal Relations Matter for Eccnomics. In *Economics and Social Interaction*, ed. Benedetto Gui and Robert Sugden. New York: Cambridge University Press pp. 1–23.

Helsley, Robert and Yves Zenou. 2014. "Social Networks and Interactions in Cities." *Journal of Economic Theory* 150:426–466.

Hilber, Christian. 2010. "New Housing Supply and the Dilution of Social Capital." *Journal of Urban Economics* 67:419–437.

Ioannides, Yannis, Henry Overman, Esteban Rossi-Hansberg and Kurt Schmidheiny. 2008. "The Effect of Information and Communication Technologies on Urban Structure." *Economic Policy* 23(54):201–242.

Jackson, Matthew. 2008. *Social and Economic Networks*. Princeton: Princeton University Press.

Jackson, Matthew. 2014. "Networks in the Understanding of Economic Behavior." *Journal of Economic Perspectives* 28(4):3–22.

Jackson, Matthew and Brian Rogers. 2007. "Meeting Strangers and Friends of Friends: How Random are Social Networks?" *The American Economic Review* 97(3):890–915.

Jacobs, Jane. 1969. *The Economy of Cities*. New York: Random House.

Lancaster, Tony. 2000. "The Incidental Parameter Problem since 1948." *Journal of Econometrics* 95(2):391–413.

Levy, Moshe and Jacob Goldenberg. 2014. "The Gravitational Law of Social Interaction." *Physica A* 393:418–426.

Lucas, Robert. 1988. "On the Mechanics of Economic Development." *Journal of Monetary Economics* 22:3–42.

Manski, Charles and Steven Lerman. 1977. "The Estimation of Choice Probabilities from Choice Based Samples." *Econometrica* 45(8):8.

Marshall, Alfred. 1890. *Principles of Economics*. London: Macmillan.

McPherson, Miller, Lynn Smith-Lovin and James Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27:415–444.

Milgram, Stanley. 1967. "The Small-World Problem." *Psychology Today* 1(1):61–67.

Powell, Joanne, Penelope Lewis, Neil Robers, Marta Garcia-Finana and Robin Dunbar. 2012. "Orbital Prefrontal Cortex Volume Predicts Social Network Size: An Imaging Study of Individual Differences in Humans." *Proceedings of the Royal Statistical Society B* 279:1–6.

Putnam, Robert. 2000. *Bowling Alone. The Collapse and Revival of American Community*. New York: Simon and Schuster Paperbacks.

Sato, Yasuhiro and Yves Zenou. 2015. "How Urbanization Affect Employment and Social Interactions." *European Economic Review* 75:131–155.

Schläpfer, Markus, Luis Bettencourt, Sebastian Grauwin, Mathias Raschke, Rob Claxton, Zbigniew Smoreda, Geoffrey West and Carlo Ratti. 2014. "The Scaling of Human Interactions with City Size." *Journal of the Royal Society Interface* 11(98):1–9.

Stiller, James and Robin Dunbar. 2007. "Perspective-Taking and Memory Capacity Predict Social Network Size." *Social Networks* 29:93–104.

Travers, Jeffrey and Stanley Milgram. 1969. "An Experimental Study of the Small World Problem." *Sociometry* 32(4):425–443.

Watts, Duncan. 1999. "Networks, Dynamics, and the Small-World Phenomenon." *American Journal of Sociology* 105(2):493–527.

A Appendix: Data

A.1 Descriptive Statistics – Municipalities and Postcode Areas

Table A.1: Main Descriptives for Municipalities and Postcode Areas

	Mean	SD	Min	Max
		Municipal Level (N=2322)		
Area in km^2	17.412	31.434	0.327	438.562
Population (from 2010 Census)	2396	3397.175	12	384786
Market Share of Swisscom	0.577	0.096	0.090	0.997
Degree of Urbanization				
Core	0.035	–	0	1
Periphery	0.336	–	0	1
Hinterland	0.629	–	0	1
Main Language				
German	0.628	–	0	1
French	0.295	–	0	1
Italian	0.065	–	0	1
Rhaeto-Romanic	0.012	–	0	1
Distance: Municipality i to j (km)	107.611	58.955	0.581	348.644
Travel Time: Municipality i to j (min.)	134.004	66.897	0.692	433.696
		Postcode Level (N=3201)		
Area in km^2	12.927	19.215	0.014	242.904
# Customers within 15 min. Radius	14683	16818.31	50	107549
Distance: Postcode i to j (km)	111.931	59.501	0.336	353.852
Travel Time: Postcode i to j (min.)	142.804	69.033	0.283	453.508

Sources: Municipal and postcode areas from Swisstopo; municipal population, language shares, and degree of urbanisation from Federal Statistical Office; car travel times from *search.ch*; number of private mobile phone customers from *Swisscom*. Data from postcodes and municipalities with less than 50 customers were deleted due to privacy concerns.

A.2 Phone Usage Statstics

Table A.2: Call Duration (in Mio. Minutes) between June 2015 to May 2016

	Phone Activity (in Mio.)					Call Duration (in Mio. Minutes)			
	MP-Calls	SMS	Landline	**Total**	*Filtered*	MP-Calls	Landline	**Total**	*Filtered*
Jun. 2015	166.3	90.9	64.3	**321.6**	*66.0*	351.2	296.2	**647.4**	*222.4*
Jul. 2015	157.3	91.9	57.8	**307.0**	*62.0*	324.8	271.1	**595.9**	*202.2*
Aug. 2015	153.6	89.0	59.7	**302.3**	*60.3*	337.0	283.6	**620.6**	*211.3*
Sep. 2015	153.8	85.2	61.9	**300.9**	*61.6*	343.0	294.2	**637.2**	*216.9*
Oct. 2015	133.6	76.3	59.9	**269.8**	*53.7*	307.5	284.8	**592.3**	*192.6*
Nov. 2015	138.1	77.7	62.1	**277.9**	*56.5*	333.1	298.5	**631.6**	*208.7*
Dec. 2015	154.1	79.1	61.6	**294.8**	*62.0*	347.4	298.1	**645.5**	*218.5*
Jan. 2016	155.7	78.5	62.0	**296.2**	*61.0*	376.0	312.4	**688.4**	*235.5*
Feb. 2016	167.6	77.5	60.6	**305.7**	*66.3*	393.3	299.6	**692.9**	*246.7*
Mar. 2016	163.3	74.9	58.6	**296.8**	*65.4*	378.1	286.8	**664.9**	*240.3*
Apr. 2016	164.2	70.7	59.9	**294.8**	*65.7*	378.8	286.1	**664.9**	*241.1*
Mai 2016	161.1	68.6	55.9	**285.7**	*64.9*	353.5	264.6	**618.1**	*228.3*

Notes: These figures base on phone usage statistics of 2.4 million private mobile phones.

A.3 Descriptive Statistics – Individual Level

Table A.3: Share of Variance in Census Population Explained by Number of Customers

	All	Male	Female	German	French	Italian
Age All	0.987	0.984	0.988	0.992	0.990	0.893
Age 20	0.945	0.946	0.944	0.960	0.946	0.916
Age 30	0.953	0.955	0.951	0.953	0.973	0.765
Age 40	0.968	0.963	0.971	0.983	0.993	0.875
Age 50	0.985	0.982	0.984	0.993	0.988	0.914
Age 60	0.990	0.988	0.987	0.994	0.984	0.922

Notes: These figures base on customer information of active phones during June 2015 and the most recent census conducted by the Federal Statistical Office in 2010.

Table A.4: Number of Private Mobile Phone Customers with a Change in Residence

			DEGURBA Classification of Movers			
Month	All	Distance > 30min	City to Hint./Peri.	Hint./Peri. to City	Within Hint./Peri.	No Change
July	13880	4461	1468	1858	2864	7690
August	14212	4572	1431	1930	2923	7928
September	15636	4842	1584	2044	3160	8848
October	15673	4795	1572	2052	3229	8820
November	14820	4612	1537	1977	3070	8236
December	14053	4202	1396	1836	3229	7592
January	13292	4432	1194	2207	2708	7183
February	13705	4333	1275	2033	2807	7590
March	15171	4671	1501	2060	3181	8429
April	15838	4873	1529	2111	3234	8964

Notes: Movers are identified based on address changes in the customer database. Columns 3 to 6 document the moving pattern along the DEGURBA classification.

Table A.5: Comparing Non-movers to Movers, Main Descriptive Statistics

	Non-Movers			Movers			
	Mean	SD	N	Mean	SD	N	*Difference*
Phone Usage Statistics, June 2015 – May 2016 (pooled)							
Number of Calls	110.525	109.039	9 564 636	126.170	114.84	834 913	*-15.646*
Duration (Minutes)	250.840	293.322	9 564 636	302.285	316.835	834 913	*-51.445*
Network Characteristics, June 2015 – May 2016 (pooled)							
Degree Centrality	9.164	7.912	9 564 636	9.633	7.875	834 913	*-0.468*
Within-Degree	7.163	7.266	9 564 636	5.971	6.721	834 913	*1.192*
Clustering Coefficient	0.092	0.134	9 423 136	0.081	0.114	825 787	*0.011*
Sociodemographics - Private Mobile Phones							
Age	35.307	13.734	797 053	31.038	10.624	69 593	*4.269*
Female	0.522	–	797 053	0.527	–	69 593	*-0.005*
Language: German	0.680	–	797 053	0.703	–	69 593	*-0.023*
Language: French	0.271	–	797 053	0.251	–	69 593	*0.020*
Language: Italian	0.043	–	797 053	0.039	–	69 593	*0.004*
Language: English	0.006	–	797 053	0.007	–	69 593	*-0.001*

Notes: The table is based on the subsample of customers with phone activity in all 12 months, which we also use in the main analysis. Further filters as described in section 3. Phone usage statistics include in- and outgoing calls. The *within-degree* measures network size within a radius of 15 minutes around an agent's residence.

B Appendix: Analysis

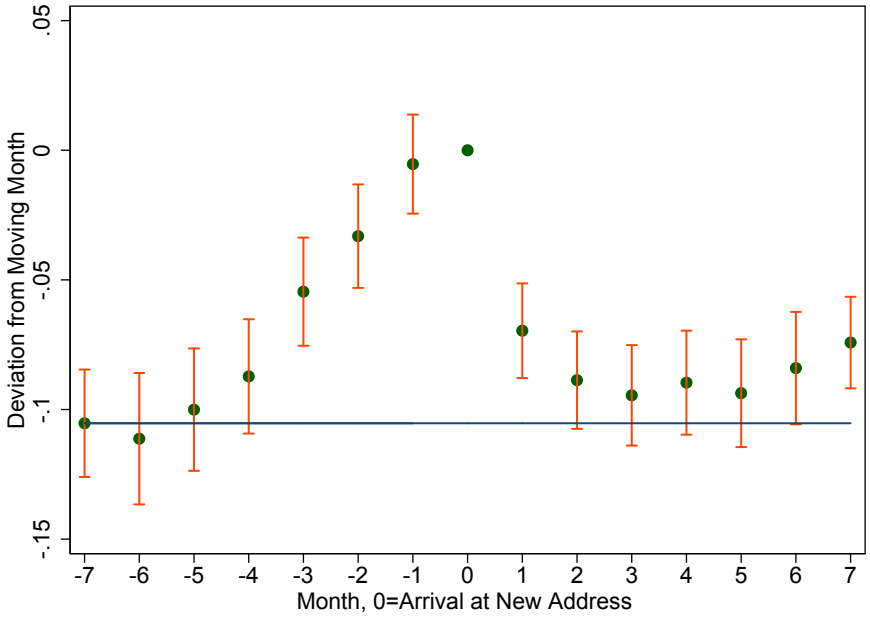


Figure B.1: The Degree prior and after Moving

B.1 Robustness Checks: Degree

Table B.1: Robustness – Cities and Network Size, June 2015–May 2016.

		Excluding Months around Change of Residence, i.e. $t = 0$				
Moving Distance at least 30 min.	All Months (0)	$t \neq 0$ (1)	$-2 \leq t \geq 2$ (2)	$-3 \leq t \geq 3$ (3)	$-4 \leq t \geq 4$ (4)	$-5 \leq t \geq 5$ (5)
Hinterland (vs. Cities)	-0.006	-0.007	-0.008	-0.017^{+}	-0.012	-0.015
	(0.006)	(0.006)	(0.007)	(0.009)	(0.012)	(0.020)
Periphery (vs. Cities)	-0.001	0.001	0.002	-0.001	-0.005	-0.027
	(0.007)	(0.007)	(0.009)	(0.011)	(0.015)	(0.024)
R^2	0.011	0.011	0.011	0.010	0.009	0.011
Groups	16,874	16,868	16,808	16,743	16,681	16,535
Observations	185,644	167,761	138,883	113,106	90,018	69,675
Population Density	-0.006	-0.008	-0.011	-0.048^{+}	-0.030	-0.094
	(0.017)	(0.019)	(0.023)	(0.027)	(0.038)	(0.061)
Population Density^2	0.000	0.000	0.000	0.002	0.002	0.005
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.003)
R^2	0.011	0.011	0.011	0.010	0.009	0.008
Groups	16,874	16,868	16,808	16,743	16,680	16,534
Observations	185,644	167,749	138,872	113,097	90,011	69,670
Further Controls	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes
Language Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *sample* covers movers (minimum moving distance 30min) who used their phone every month at least once. Column (1) excludes the moving month; column (2) excludes the moving month and the first month prior and after moving; and so on. *Further controls* include commuting distance and a dummy for belonging to language minority. Standard errors in parentheses. $^{+}$ p<0.10, * p<0.05, ** p<0.01 *** p<0.001.

B.2 Robustness Checks: Within-Degree

Table B.2: Robustness – Cities and the Within-Degree, June 2015–May 2016.

		Excluding Months around Change of Residence, i.e. $t = 0$				
Moving Distance at least 30 min.	All Months (0)	$t \neq 0$ (1)	$-2 \leq t \geq 2$ (2)	$-3 \leq t \geq 3$ (3)	$-4 \leq t \geq 4$ (4)	$-5 \leq t \geq 5$ (5)
Hinterland (vs. Cities)	-0.038*	-0.039*	-0.049*	-0.058*	-0.049	-0.084+
	(0.018)	(0.018)	(0.021)	(0.025)	(0.032)	(0.046)
Periphery (vs. Cities)	-0.132***	-0.133***	-0.149***	-0.148***	-0.148***	-0.194***
	(0.020)	(0.021)	(0.024)	(0.028)	(0.036)	(0.053)
R^2	0.012	0.015	0.017	0.016	0.015	0.012
Groups	16,874	16,868	16,808	16,743	16,681	16,535
Observations	185,676	167,761	138,883	113,106	90,018	69,675
Population Density	0.076***	0.077***	0.082***	0.085***	0.087***	0.087***
	(0.006)	(0.006)	(0.007)	(0.008)	(0.010)	(0.015)
R^2	0.016	0.019	0.021	0.020	0.018	0.013
Groups	16,874	16,868	16,808	16,743	16,680	16,534
Observations	185,663	167,749	138,872	113,097	90,011	69,670
Further Controls	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes
Language Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *sample* covers movers (minimum moving distance 30min) who used their phone every month at least once. Column (1) excludes the moving month; column (2) excludes the moving month and the first month prior and after moving; and so on. *Further controls* include commuting distance and a dummy for belonging to language minority. Standard errors in parentheses. + p<0.10, * p<0.05, ** p<0.01 *** p<0.001.

B.3 Robustness Checks: Clustering

Table B.3: Robustness – Cities and Clustering, June 2015–May 2016.

		Excluding Months around Change of Residence, i.e. $t = 0$				
	All Months (0)	$t \neq 0$ (1)	$-2 \leq t \geq 2$ (2)	$-3 \leq t \geq 3$ (3)	$-4 \leq t \geq 4$ (4)	$-5 \leq t \geq 5$ (5)
Hinterland (vs. Cities)	0.002+	0.002+	0.002	0.004*	0.003	0.002
	(0.001)	(0.001)	(0.001)	(0.002)	(0.003)	(0.004)
Periphery (vs. Cities)	0.002+	0.003*	0.003*	0.006**	0.006*	0.008+
	(0.001)	(0.001)	(0.002)	(0.002)	(0.003)	(0.004)
R^2	0.001	0.001	0.001	0.001	0.001	0.01
Groups	16,870	16,863	16,802	16,735	16,670	16,518
Observations	183,896	166,130	137,489	111,965	89,099	68,953
Population Density	-0.001+	-0.001+	-0.001+	-0.001+	-0.001+	-0.003*
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)
R^2	0.001	0.001	0.001	0.001	0.001	0.001
Groups	16,870	16,863	16,802	16,735	16,669	16,517
Observations	183,896	166,118	137,478	111,956	89,092	68,948
Further Controls	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes
Language Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *sample* covers movers (minimum moving distance 30min) who used their phone every month at least once. Column (1) excludes the moving month; column (2) excludes the moving month and the first month prior and after moving; and so on. *Further controls* include commuting distance and a dummy for belonging to language minority. Standard errors in parentheses. + p<0.10, * p<0.05, ** p<0.01 *** p<0.001.

B.4 Robustness Checks: Matching

Table B.4: Robustness – Regional Differences in the Matching Quality, Robustness

a. Network Formation	Full Sample		Moving Distance > 30min.	
Cities and Matching (Predicted FE)	(1)	(2)	(3)	(4)
$\text{Hinterland}_{i,t}$	-101.581***		-44.090	
	(-6.316)		(-0.905)	
$\text{Periphery}_{i,t}$	-381.461***		-252.426***	
	(-20.221)		(-4.952)	
Pop. $\text{Density}_{i,t}$		88.509***		56.085***
		(17.412)		(6.019)
Constant	5447.207***	4439.930***	5346.683***	4718.905***
	(414.239)	(88.771)	(121.435)	(54.765)
R^2	0.002	0.001	0.001	0.001
Observations	1,631,708	1,646,566	30,6798	313,072
b. Network Topography	Full Sample		Moving Distance > 30min.	
Cities and Matching (Social Adaption)	(1)	(2)	(3)	(4)
City_{post}	0.308***		0.202***	
	(0.046)		(0.056)	
City_{pre}	-0.231***		-0.275***	
	(0.033)		(0.045)	
Pop. Density_{post}		0.184***		0.099***
		(0.013)		(0.013)
Pop. Density_{pre}		-0.145***		-0.103**
		(0.016)		(0.031)
Constant	0.738***	0.304^{+}	0.754**	0.670***
	(0.124)	(0.176)	(0.282)	(0.033)
R^2	0.017	0.041	0.018	0.005
Further Controls	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes
Language Region FE	Yes	Yes	Yes	Yes
Observations	5,718	5,718	3,108	3,194

Notes: *Dependent Variable in Panel a.:* Predicted dyad specific fixed effect from network formation model outlined in equation (10b). *Dependent Variable in Panel b.:* The number of *post-move* contacts at the *post-move* place of residence over the number of *post-move* contacts at the *pre-move* place of residence. *Controls in Panel b.:* Number of contacts at new address prior to moving, commuting distance, dummy for belonging to language minority, and Romansh region), gender and age . Standard errors in parentheses. + p<0.10, * p<0.05, ** p<0.01 *** p<0.001.

Center for Regional Economic Development (CRED)

University of Bern

Schanzeneckstrasse 1

P.O.Box

CH-3001 Bern

Telephone: +41 31 631 37 11

E-Mail: info@cred.unibe.ch

Website: http://www.cred.unibe.ch

The Center for Regional Economic Development (CRED) is an interdisciplinary hub for the scientific analysis of questions of regional economic development. The Center encompasses an association of scientists dedicated to examining regional development from an economic, geographic and business perspective.

Contact of the authors:

Konstantin Büchel

University of Bern

Schanzeneckstrasse 1

P.O.Box

CH-3001 Bern

Telephone: +41 31 631 49 97

Email: konstantin.buechel@vwi.unibe.ch

Maximilian von Ehrlich

University of Bern

Schanzeneckstrasse 1

P.O.Box

CH-3001 Bern

Telephone: +41 31 631 80 75

Email: maximilian.vonehrlich@vwi.unibe.ch

This paper can be downloaded at:

http://www.cred.unibe.ch/forschung/publikationen/cred_research_papers/index_ger.html