NEIGHBORHOOD AND NETWORK SEGREGATION: ETHNIC HOMOPHILY IN A 'SILENTLY SEPARATE' SOCIETY*

Joshua Blumenstock, University of Washington
Ott Toomet, Tartu University

This Version: September 2015

Abstract

We examine the relationship between geography and ethnic homophily in Estonia, a linguistically divided country. Analyzing the physical locations and cellular communications of tens of thousands of individuals, we document a strong relationship between the ethnic concentration of an individual's geographic neighborhood and the ethnic composition of the people with whom he interacts. The empirical evidence is consistent with a theoretical model in which individuals prefer to form ties with others living close by and of the same ethnicity. Exploiting variation in the data caused by migrants and quasi-exogenous settlement patterns, we find suggestive evidence that the ethnic composition of geographic neighborhoods has a causal influence on the ethnic structure of social networks.

Keywords: residential segregation, homophily, social segregation, minorities

JEL codes: J15, J61

1 Introduction

Ethnic segregation is a prominent feature of most contemporary and historical societies. Such fractionalization has been tied to patterns of economic development and growth, investment in human capital, the efficiency of labor markets, violence and corruption, as well as broader patterns of inequality, prejudice, and discrimination (cf. Easterly and Levine 1997, Collier 1998, Cutler and Glaeser 2007, Bayard, Hellerstein, Neumark, and Troske 1999, Miguel and Gugerty 2005).

*The authors are grateful for thoughtful comments from Mark Ellis, Ira Gang, Matthew Jackson, Štěpán Jurajda, Ramona Angelescu-Naqvi, Xu Tan, and participants of Tartu 2014 Xmas seminar. We also thank Siiri Silm for help with geodata analysis. We gratefully acknowledge financial support from GDN RRC and ESRC and Estonian Science Foundation grants IUT2-17, IUT20-49, and 9247. All errors are our own.

Ethnic homophily, or an individual's preference for co-ethnics, plays an important role in shaping patterns of ethnic segregation. When individuals prefer to associate with others of the same type, this influences where people choose to live, where people choose to work, and with whom they choose to interact (Massey 1985). At the same time, the reverse process may also obtain: when people are physically surrounded by others of their own type, they may choose to associate with them to a higher degree The relationship between neighborhood segregation, as defined by the structure of geographic communities, and network segregation, as determined by actual patterns of interaction, is thus theoretically ambiguous. Empirical adjudication is similarly complicated by the difficulty of separately observing both types of segregation on a single population.

Here, we study the relationship between neighborhood segregation and network segregation using a novel dataset that allows us to disentangle the ethnic composition of an individual's physical surroundings from the ethnic composition of people with whom he interacts. The context for this study is Estonia, a country with a long and complex history of ethnic strife and resettlement. Prior to World War II, roughly 94 percent of the Estonian population was ethnically Estonian; however, the Estonia's incorporation into the USSR created a large influx of Russian immigrants into Estonia, and by 1989 roughly 39 percent of the Estonian population was ethnically Russian. Due to Stalin's brutal regime, and the anti-Russian backlash that followed Estonian independence, strong feelings of animosity exist between the two groups. Modern-day Estonian society has been described by Heidmets (1998) as one of "silent separation" where the two ethnic groups occupy the same physical spaces but rarely interact.

Empirically, we analyze a large dataset of anonymized mobile phone communications data that allows us to observe, for tens of thousands of individuals, all locations inhabited be each individual over a five-year period, as well as all phone-based interactions with other individuals in the dataset. Critical to our current analysis, we also observe the language spoken by each mobile phone subscriber. In Estonia, linguistic preference remains a core component of ethnic identity, with most ethnic Russians choosing to speak the Russian language, and most ethnic Estonians choosing to speak the Estonian language (Tammaru 2001). Using these data, we are thus able to separately measure, for each individual, the extent to which she is physically surrounded by coethnics, and the ethnic composition of her social network.

We observe a strong and robust relationship between neighborhood and network homophily. In other words, while the average individual is more likely to interact with coethnics than others, those individuals who are physically surrounded by coethnics are even more likely to interact

with coethnics. This relationship exceeds what would be expected under a naive model of geographically constrained random attachment, where individuals randomly interact with those in their geographic network irrespective of ethnicity. This effect persists even when controlling for a range of demographic characteristics.

We further find suggestive evidence that the ethnic composition of an individual's network exerts a causal influence on the social connections formed by the same individual. In particular, when we study the homophilic tendencies of migrants,[1] we find that they are *less* sensitive to their physical surroundings; while they still interact more with coethnics the more they are surrounded by coethnics, this relationship is less pronounced than it is in the population of people who never migrate. It appears that this differential effect for migrants is primarily driven by the fact that migrants remain connected to their place of origin, and do not immediately form connections in their new neighborhood.

The above results are consistent with a simple model of social preferences where geographically close coethnic ties are most likely to form. If interethnic ties are possible, the number of interethnic friends will depend on the "local" ethnic composition. The relationship is weaker for migrants who have just recently arrived in the region and have not yet had time to adapt their networks to the new local environment.

In analyzing the behavior of migrants, we cannot rule out the possibility that individuals with different homophilic tendencies select into migration. However, we exploit several plausibly exogenous sources of variation to help assuage such concerns of endogeneity. First, we use the high-resolution mobility data to infer the exact date of migration, and look separately at homophily for migrants of different cohorts. Here, we find that recent migrants are indeed more likely to be in contact with their origin community than migrants who have lived in the destination community for several years; these recent migrants are also less likely to interact with individuals in their new neighborhood.

As a second robustness test, we analyze the relationship on Soviet-era housing estates. As the housing market was virtually non-existent during the communist period, the residential location was largely exogenous and the segregation preferences play little role accordingly. The find the effect being almost as strong on these estates as in the full sample, suggesting that neighborhoods that play a substantially stronger role in determining the social networks.

Our paper thus documents the strong relationship between neighborhood and network seg-

[1] Migrants are identified in our data based on the set of geolocated mobile phone towers used to route their calls. Migrants constitute roughly 10% of the sample population, and our results are robust to a variety of plausible ways of classifying migrants.

regation, and provides suggestive evidence on causality. Taken further, these results imply that physical integration may lead to social integration.

The remainder of the paper is organized as follows: Section 2 describes the background and institutions of Estonia, the country we analyze. Section 3 provides a simple theoretical framework for interpreting our results. The data and the empirical approach are described in greater detail in Section 4, and Section 5 discusses the results. In Section 6 we discuss and interpret the results in greater detail before concluding.

2 Estonia: Background and Context

The focus of our empirical analysis is Estonia, a country uniquely suited for the empirical analysis of ethnic segregation. Before World War II, roughly 94 percent of the population was ethnic Estonians, with the remainder largely comprised of ethnic Russians (Katus 1990). During WWII, Estonia was incorporated into the Soviet Union, and fell under the brutal Stalinist regime for nearly a decade. As part of the post-war reconstruction and industrialization effort, it experienced large-scale immigration from other parts of the Soviet Union, mainly from Russia. This process resulted in an increase of the population of the country to 1.57 million by 1989, 39% of whom were ethnic minorities (Tammaru and Kulu 2003).

Most of the migrants from the Soviet Union were Russian-speaking, and were regarded by the ethnic Estonian population as the "Russians" and associated with the harsh regime. Relations between the two dominant ethnic groups deteriorated rapidly, and by the 1970s the society was sharply divied along linguistic lines. Estonians and Russians attended different schools, worked in different establishments and followed different media outlets (Kalmus and Pavelson 2002, Vihalemm 2010). Russian language functioned as the *lingua franca* of the Soviet Union and most Estonians possessed a working knowledge of the language. However, the command of the Estonian language was poor among the minorities (Kulu and Tammaru 2004). Despite the official Soviet policy, Estonians never considered themselves a part of the Soviet nation, and distinguished clearly between in-group (i.e. "Estonians") and out-group (i.e. "Russian") members. This linguistically divided society where two ethnic communities live rather parallel lives has thus been characterized as a"silently separated society" (Heidmets 1998).

After the fall of the USSR, Estonia began a nation-building that has been widely regarded as discriminatory towards ethnic Russians (Pettai 2002). First, the newly elected parliament granted citizenship only to nationals of the pre-WWII republic and to their offspring (Everly

1997). As a result, a sizeable part of the current minority population does not have Estonian citizenship. Second, the Estonian language was made the sole official language of the country, causing a gradual deterioration of Russian language skills among Estonians, particularly among the younger generation. However, a large percentage of Russians are still not able to communicate in Estonian (Kulu and Tammaru 2004). For this reason there is no universally shared language in the country today. The shift in the roles of the languages was also accompanied by relative deterioration of the economic position of Russian speakers (Leping and Toomet 2008). In this way the historic animosity between the two language groups, the high levels of segregation in many important spheres, and the lack of a *lingua franca* contribute to the low number of interethnic contacts and general lack of social integration today. While the attitudes toward the other ethnic groups have been improving through the previous decade, interethnic engagement is still relatively infrequent (Lauristin, Uus, and Seppel 2011). The tensions do occasionally rise to the surface as, for instance, during the large-scale riots in Tallinn in the spring of 2007.[2]

The fall of the Soviet regime also had a signifant impact on patterns of migration and settlement. The Russian immigration of the Soviet era come to a rapid halt, while both urbanization and sub-urbanization gathered momentum. The main mechanism that shaped the ethnic composition of the urban neighborhoods, including in the capital Tallinn, is related to historic immigration and residential construction. Between 1950 and 1989, the population of the country rose by more than 40%, from 1,097,000 to 1,565,000, mainly through immigration from elsewhere in the USSR (Tammaru 2001). In the absence of a housing market, immigrants were usually granted flats in newly built, standardized, high-rise housing estates (Kährik and Tammaru 2010) which nowadays provide accommodation for a large part of the total population. These are often dominated by ethnic Russians, whereas Estonians are over-represented in pre-WWII (and also in the small post-1991) housing stock, and also in detached houses. In this way, the current ethnic composition across urban neighborhoods largely reflects the immigration patterns during the construction periods, rather than factors such as socio-economic status. Recent suburbanization, and the fact that a substantial part of the immigrant population left after the collapse of Soviet Union, has not radically changed this picture (Hess, Tammaru, and Leetmaa 2012).

[2]The riots were caused by the relocation of a Soviet World War II monument, popularly referred to as the "Bronze Soldier", from central Tallinn to a military cemetery. From the perspective of ethnic Estonians, the monument was considered to glorify oppressive Soviet rule, while for the Russian-speaking population it was a symbol of victory over the Nazis in the "Great Patriotic War." See Schultze (2011).

3 Theoretical Framework

We develop a simple model that includes two different groups of people, two regions, and migration between these regions. The model allows us to describe the expected number of ties within and across groups, and illustrates two important results: first, homophily, or the preference for individuals of the same group, is positively correlated with the neighborhood ethnic composition; and second, this relationship is stronger for people who remain in a single location and weaker for people who migrate. We formalize these results as propositions below and explain the intuition.

We consider a world containing two regions, A and B, and two (ethnic) groups, 0 and 1. The population in region A is n^A and in region B it is n^B. There are $n_0^A = \pi^A n^A$ group-0 members and $n_1^A = (1-\pi^A)n^A$ group-1 members in region A where π^A is the fraction of group-0 members in region A. The expressions for region B are analogous.

Assume that ties between two individuals, located at geographic distance d^g and "ethnic distance" d^e is created by a Poisson process with intensity

$$\mu = \phi(d^g) \cdot \chi(d^e). \tag{1}$$

The first term $\phi(\cdot)$ describes how the intensity depends on the geographic distance and $\chi(\cdot)$ describes the dependence on "ethnic distance", where $d^e = 0$ for members of the same group and $d^e = 1$ for members of the different group. As we only have two regions and two groups, we can label the corresponding function values as ϕ^0 and ϕ^1 for local ties and distant ties, and χ^0 and χ^1 for in-group and out-group ties. We assume "short" ties arise more easily: $\phi^0 > \phi^1$ and $\chi^0 > \chi^1$. Ties are destroyed with Poisson process with intensity δ, independent of their "length". Assuming that the number of actual ties is much smaller than the number of potential ties, we have the following expression for the expected number of individual ties, ν. For instance, for a group 0 member in region A we have:

$$\frac{\mathrm{d}\,\nu}{\mathrm{d}\,t} = \phi^0(\chi^0 n_0^A + \chi^1 n_1^A) + \phi^1(\chi^0 n_0^B + \chi^1 n_1^B) - \delta\nu. \tag{2}$$

The first term describes creation of new local ties, the second term that of distant ties, and the last term the destruction of ties.

If people who remain in one region have lived there long enough, their expected number of ties correspond to these in the steady-state, ν^* and homophily (for group 0) $h^* = \nu_0^*/(\nu_0^* + \nu_1^*)$.[3]

[3] Homophily is a measure of exposure dimension of segregation (Massey and Denton 1988). Here we focus on

[4] Here ν_0^* and ν_1^* are ties to group 0 and group 1, i.e. in-group and out-group ties for group 0.

Proposition 1. *Homophily in social ties is positively related to the ethnic composition of local geographic neighborhoods:* $\frac{\partial h^*}{\partial \pi^A} > 0$.

Proof. See Appendix A. □

Intuitively, as our model explicitly allows for a greater likelihood of local tie formation, the local population composition influences substantially the actual homophily.

As migrants do not posses the the steady-state equilibrium networks, we solve the dynamic equation (2) and have the following result:

Proposition 2. *Conditional on the average level of neighborhood homophily, the homophily of migrants is less sensitive to the neighborhood composition than that of non-migrants.*

Proof. See Appendix A. □

This effect exists because local network ties develop over time. In-migrants, arriving from neighborhoods of different ethnic composition, have only partially adapted to the ethnic composition of their new neighborhoods.

4 Data and Empirical Approach

To analyze the relationship between neighborhood and network segregations, we exploit a large set of data on mobile phone use in Estonia. This dataset permits us to simultaneously observe the locations of thousands of individuals over a period of several years, the ethnicity of those individuals, as well as the extent to which those individuals interacts with others of the same or different ethnicity.

4.1 Data

We employ cellphone usage data from the largest mobile service provider in Estonia, EMT, which has roughly 60% market share. We obtained two related datasets from this operator.

Passive Positioning Data: The first dataset contains the locations of each individual over the period from 2007–2012. As is typical for such positioning data, we do not observe the actual

homophily based on the percentage of contacts in an individual's network, but our empirical results are robust to other common definitions of homophily.

[4] Empirically we observe $\mathbb{E}\, h^* = \mathbb{E}[\nu_0^*/(\nu_0^* + \nu_1^*)]$ instead of $\mathbb{E}\, \nu_0^*/(\mathbb{E}\, \nu_0^* + \mathbb{E}\, \nu_1^*)$. However, under mild assumptions these two expressions are equal. See Appendix A.

location but rather the Cell Global Identity (CGI), i.e. the network antenna which processed the outgoing call.[5] This gives us a spatial resolution of a few hundred meters in dense urban environments, and up to five kilometers in rural areas. The data include the time of each call activity and the corresponding location (CGI). Every network user (as identified by a SIM card with a unique phone number) is assigned a random identification tag, making it possible to track the same user over time.[6] Based on timing, location and regularity of the calls, we attach a place of residence to each cellphone (Ahas, Silm, Järv, Saluveer, and Tiru 2010). We focus on yearly modal place of residence in order to avoid places that are too unstable, or seasonal migration.

Call Graph: The second dataset contains a complete 10-day call graph, which allows us to observe in a fixed window who is communicating with whom. This call data records (CDR) are similar to the passive positioning data, but for each call or SMS event we also observe the ID of the second party.

For each subscriber in our dataset, we additionally observe whether the subscriber prefers the Estonian or Russian language. Since the correlation between ethnicity and language is almost complete (Kulu and Tammaru 2004) we use language as a proxy for ethnic background. All of these data use shared anonymized identifiers which allow us to link long-term location information to the network communication data, and in this way to relate the segregation in communication network to segregation in space.

We perform our empirical analysis on a random sample of 48,781 individual mobile phone subscribers. Of these, 42,604 are Estonian and 6,178 are Russian; 46,835 have a known residence location, and 18,716 live in the metropolitan area.

4.2 Empirical Approach

In our empirical analysis, we examine the relationship between the ethnic composition of an individual's immediate physical neigborhood, and the ethnic composition of his call graph. The passive positioning data allows us to determine where individuals live, which in turn makes it possible to observe the physical neighborhood. We will begin by defining physical neighborhood

[5] In a cellular network, a "cell" roughly corresponds to an area where all the network traffic goes through a single antenna. Usually, several antennas are located in one transmission tower and are oriented in different directions. We know the location of the transmission towers and the direction of the antennas. Based on this information, we can construct "typical" cell boundaries; however, the actual boundaries may fluctuate due to network load, obstacles and noise.

[6] The individuals and real phone numbers cannot be identified using the tag in our data. The collection, storage, and processing of the data complies with all European Union requirements regarding the protection of personal data (European Commission 2002). Approval was also obtained from the Estonian Data Protection Inspectorate and the University of Washington Human Subjects Division.

along political boundaries, but our results are robust to several alternative definitions.[7] Similarly, we will initially measure the ethnic composition of the call graph as the fraction of contacts of the same ethnicity, but our results obtain when we define network homophily as the fraction of communication events (i.e., the weighted call graph).

Below, we will separately analyze the relationship between neighborhood and social network homophily, as well as the extent to which people are connected to current and historical regions of residence. In both cases we show the nonparametric relationships graphically, then test the statistical relationship in a regression specification.

The basic regression equation for estimating the relationship between social network homophily h_i and neighborhood own-group percentage P_i for individual i is

$$h_i = \alpha_0 + \beta E_i + \gamma P_i + \eta P_i \cdot M_i + \epsilon_i \tag{3}$$

where E_i indicates the ethnicity of individual i, and M_i indicates whether i is a migrant.[8] Note that as P_i is defined at region level, we cluster the standard errors within regions.

Later, we will also introduce several variants on model (). First, to analyze the geographic structure of networks in greater detail, we will split the connections into local and distant ones, depending on whether these cross a county border. For migrants, we will also separately analyze the extent to which their current social network is comprised of people residing in their current location, or the location from which they migrated. Additionally, to assess the robustness of our results, we will restrict our sample to specific types of individuals, for instance those who were likely to be assigned to their current place of residence through a Soviet-era natural experiment. Finally, we will also disaggregate migration by cohort, to determine whether the effects of migration are different for recent migrants. Formally, if b is the proportion of given type connections

[7] We perform our analysis using different types of spatial units. Calculation of neighborhood homophily is based on city tracts inside of the capital city. These are spatial units, based on access roads and housing type. Elsewhere in the metropolitan area we rely on municipalities, the area contains 18 suburban municipalities. Finally, outside of the metropolitan area we use counties, there are 14 of these outside of the metropolitan area. Counties are of roughly equal size (though of very unequal population) and broadly correspond to commute-to-work area around an urban center. We choose such an approach to account for different population density and also to take into account the uneven distribution of network antennas. Finally, we analyze migration and geographic tie distance at county level.

[8] As people show a heterogeneous pattern of spatial mobility, we use several definitions of migrants. The strictest definition requires a valid residence region for all 6 years, and only a single move during this period. This gives us 2,614 migrants. The most flexible definition allows up to three missing yearly locations and up to two moves (we analyze the last of these). This gives us 6,592 migrants. All our central results are robust with respect to the definition of migrants.

and YSM is years since migration, we estimate:

$$b_i = \alpha_0 + \beta \cdot YSM_i + \epsilon_i. \tag{4}$$

5 Results

5.1 Neighborhood and Network Homophily

We start by presenting the relationship between the regional homophily and the average network homophily for the residents of these neighborhoods. Figure 1 shows the nonparametric relationship between geographic neighbordhood composition and observed homophily in the call graph. Each point indicates the average values for a municipality, with each municipality appearing once for Estonians (hollow circles) and once for Russians (filled circles). The horizontal axis indicates the proportion of a given ethnicity in the municipality and the vertical axis indicates the average homophily of that ethnicity in that region. We see a clear positive relationship for both Estonian speakers and Russian speakers. We also see that while the slope for both groups is similar, Russian speakers are substantially less homophilous.

Next we estimate the same relationship at the individual level, using variants of model (), described above. Table 1 presents four different specifications, where in addition to neighborhood composition we add different combinations of migrant and minority status. All models confirm the visual impression that network composition is strongly related to that of neighborhoods. In the first column, we simply regress individual network homophily h_i on the ethnic composition of the neighborhood P_i. The estimates indicate that a 10 percentage point increase in co-ethnics in a geographic neighborhood corresponds to a 3.5 percentage point increase in co-ethnics in the call network (column 1). This figure is highly significant and robust to the inclusion of several control variables (columns 2-4).

In Column 2 of Table 1, we note that migrants are in general more homophilous than non migrants (by 12 percentage points), but that critically the relationship between the neighborhood and the network is weaker. For migrants, a 10 percentage point incresae in co-ethnic share is only associated with a $10 \times (0.36 - 0.12) = 2.4$ percentage point increase in co-ethnics in the call network.

Model 3 adds controls for ethnicity. We see that while Russians are generally less homophilous than Estonians, they are more sensitive to the neighborhood environment: here, 10 percentage points more co-ethnics in the neighborhood corresponds to $10 \times (0.20 + 0.22) = 4.4$ percentage

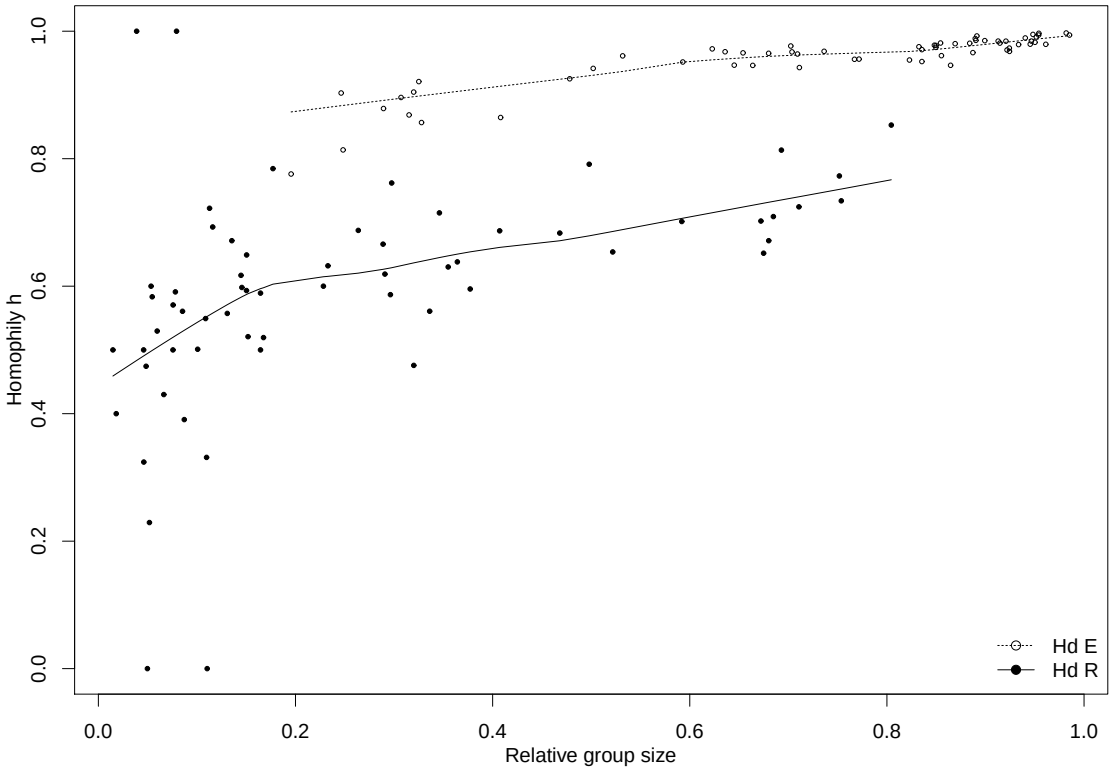


Figure 1: Average homophily as a function of group size across regions. Empty circles denote Estonian-speakers, filled circles Russian speakers. The dashed and solid lines are corresponding smoothed averages.

points higher share in call network. The lower average homophily levels are presumably related to smaller country-wide numbers of Russian speakers while the figure suggests that the stronger correlation is related to neighborhoods with very low group size. The housing type may also play a role as that Russian speakers are overrepresented in Soviet-era high-rise estates.

The above results are evidence of the strong relationship between neighborhood and network segregation. Whether this relationship is causal, however, is not clear. The model presented in Section 3 assumed that neighborhood composition would influence tie formation, and these results are consistent with the two Propositions from that model. Below, we provide additional empirical evidence that appears to indicate there is indeed a casual effect of physical segregation on network homophily.

Outcome: individual homophily h (percentage of co-ethnic contacts in call network)

	Model 1	Model 2	Model 3	Model 4
% coethnics	0.35^{***}	0.36^{***}	0.20^{***}	0.21^{***}
	(0.00)	(0.00)	(0.00)	(0.01)
Migrant		0.12^{***}		0.10^{***}
		(0.02)		(0.02)
Migrant×% coethnics		-0.12^{***}		-0.11^{***}
		(0.02)		(0.02)
Russian			-0.32^{***}	0.34^{***}
			(0.01)	(0.00)
Russian×% coethnics			0.22^{***}	0.22^{***}
			(0.01)	(0.01)
Migrant×Russian				-0.04
				(0.03)
Migrant×Russian×% coethnics				0.02
				(0.06)
Intercept	0.67^{***}	0.66^{***}	0.80^{***}	0.68^{***}
	(0.00)	(0.00)	(0.00)	(0.00)
R^2	0.14	0.14	0.22	0.22
Num. obs.	40819	40819	40819	40819

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 1: Full sample results: relationship between individual homophily and neighborhood ethnic composition.

5.2 Soviet-Era Neighborhoods as a Natural Experiment

If the decision to migrate were exogenous, the fact that the network structure of migrants is less strongly correlated with the geographic struture of their current surroundings could be interpreted as evidence of a causal relationship in which physical surroundings determine network structure. However, since migration and patterns of settlement are not generally exogenous, it may also be the case that people select into migration, and in partcular that people who care less about their physical surroundings are the ones who choose to migrate. To disentangle these two possibilities, we examine the same relationship on a sample of individuals for whom the current location choice is more plausibly exogenous.

As discussed in Section 2, the constant shortage of housing and lack of choice in the housing market created quasi-exogenous variation in patterns of settlement. During the period of highest Russian immigration, 1970s and 80s, individuals had little choice over where to live in cities. We treat this as a natural experiment and limit our analysis here to neighborhoods that are dominated by Soviet-era housing estates. We focus on the capital city Tallinn only. Note that

our setup does not constitute a perfect experiment as we do not know who in our sample did actually live in these neighborhoods during the Soviet period. However, we exclude all the migrants into these areas we are able to identify in the sample.

Outcome: individual homophily h (percentage of co-ethnic contacts in call network)

	Model 1	Model 2	Model 3
% coethnics	0.05	0.20^{***}	0.17^{***}
	(0.04)	(0.04)	(0.04)
Russian		-0.25^{***}	-0.33^{***}
		(0.01)	(0.05)
Russian×% coethnics			0.13
			(0.09)
Intercept	0.85^{***}	0.83^{***}	0.85^{***}
	(0.02)	(0.02)	(0.02)
R^2	0.00	0.16	0.16
Num. obs.	2362	2362	2362

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Soviet-era neighborhoods: relationship between individual homophily and neighborhood ethnic composition.

The results (Table 2) are rather similar to the estimates in the full sample. In particular, the estimate for co-ethnics, 0.17, in Model 3 is statistically indistinguishable to that of Model 3 in the full sample, 0.20 (in Table 1). However, as the sample is smaller, the standard errors are correspondingly larger. The fact that in this sample, who were plausibly exogenously settled into their currently location, we see the same strong relationship between neighborhood and network segregation, lends support to the causal nature of the relationship, i.e., that social networks are at least in part determined through the residential neighborhoods, and that a portion of network segregation is caused by geographic segregation.

5.3 Local and Distant Ties

The homophily-related predictions in Section 3 are based on the fact that local ties are created more easily than distant ones. Unfortunately, as our observations of network structure are based on a single 10-day period, we cannot directly observe the process of tie formation. However, we do observe the geographic structure of ties, which allows us to make two observations that are consistent with our model of tie formation. First, for for migrants, we observe that distant ties are primarily linked to their former place of residence: most of the distant ties were formed when

the distant place still was "local". Second, when we separate the migrants into cohorts by time since migration (we can observe migration 1-6 years ago), we observe an increasing number of local ties and a decreasing number of ties to the previous home region.

Specifically, we split the connections to local and distant ones based on counties (Estonia consist of 15 counties, roughly similar in terms of area but very unevenly populated). We compare the number of ties in the current county of residence, in the previous county of residence (for migrants only), and in all other counties. We select residents of the metropolitan area as of 2012. Table 3 presents their average number of contacts (based on the 10-day callgraph) in selected counties: the Metro area, Ida-Viru (code 44), Pärnu (67), and Tartu (78).[9] The rows correspond to the previous (2011) residence: *Metro* are those who were living in the metropolitan area in 2011 as well, i.e. "stayers"; 44 are those who lived in Ida-Viru and hence they are recent migrants to the metro area, and analogously for the other rows. Columns represent the county of contact.[10] In case of Estonian-Estonian ties (left panel), we see that those who have been in the metropolitan area both for 2011 and 2012 (row labeled "Metro") clearly posses the largest number of the connections in that area. The average number of connections in other counties (columns labeled "44", "67" and "78") is very small. For movers (rows labeled "44", "67" and "78") the picture is different. All of them possess a substantial number of connections in the metro area (after all, they are living there as of 2012) while the number of contacts to their previous county of residence is also relatively large (left panel, main diagonal). However, the number of contacts in the other counties is negligible, exactly as in case of those who never migrate. The Russian-Russian ties (right panel) paint a similar picture. There are too few observations for any inference on interethnic ties (not shown).

To summarize, Table 3 strongly suggests that ties form locally. People who never migrate are almost exclusively connected to their current county while migrants have a substantial number of connections to their previous county.

5.4 Evidence on Tie Creation and Destruction

Analyzing differences in tie structure by year of migration allows us to indirectly test the theory of tie formation posited above. Figure 2 shows the relationship between migration year and the geographic structure of the current social network. The figure indicates that the number of contacts in the current county (circles) is lower for the recent migrants while the number

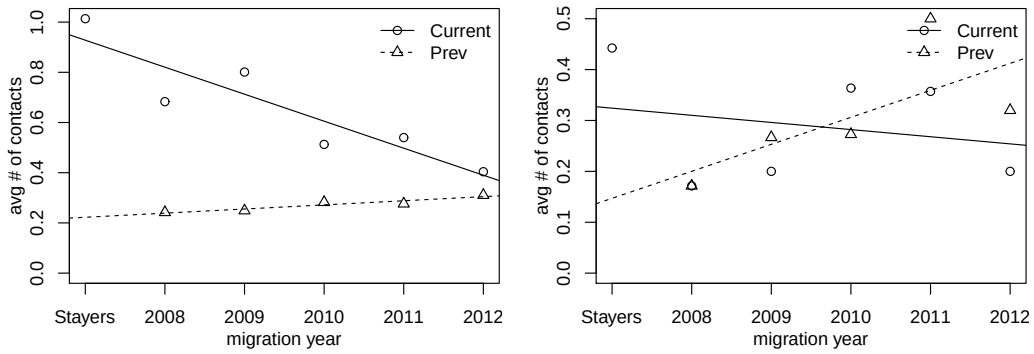
[9] The results for other counties are qualitatively similar.

[10] The numbers are low because we do not observe valid county of residence for the contacts outside of the sample.

Residence 2011	connections to Metro	44	67	78	connections to Metro	44	67	78
	Estonian-Estonian				Russian-Russian			
Metro	0.91	0.01	0.03	0.05	0.45	0.03	0.00	0.00
44	0.35	0.28	0.02	0.02	0.20	0.39	0.00	0.02
67	0.45	0.00	0.32	0.06	0.00	0.00	0.17	0.00
78	0.72	0.00	0.01	0.39	0.14	0.00	0.00	0.29

Notes: Residents of the metropolitan (capital) area 2012 depending on their 2011 residence (in rows), and their number of contacts (degree) in columns. The county codes are 44 = Ida Viru; 67 = Pärnu; 78 = Tartu.

Table 3: Number of contacts in the current residence county, previous residence county, and other counties. Estonian-Estonian and Russian-Russian ties.



(a) Estonian-Estonian (b) Russian-Russian

Figure 2: Average number of contacts to the current and previous county of residence. The lines represents the corresponding linear fit (with stayers excluded).

of contacts in the previous county (triangles) is decreasing. The figure for Estonian-Estonian ties (left panel) is less noisy and suggests that the ties to the former county are rather resilient and may be related to family or others in-kin (Phithakkitnukoon, Calabrese, Smoreda, and Ratti 2011).

Table 4 gives similar results using individual-level regressions. We estimate the percentage of connections to the current and previous county of residence as a function of years since migration. The table indicates that the share of contacts in the former place of residence falls by about 4 percentage points per year and are replaced by a corresponding growth in connections in the current place of residence.

In summary, our contact distance analysis strongly suggests that ties arise locally over time and also fade away over time when individuals move elsewhere. These outcomes fit to our theoretical framework and suggest that neighborhood population composition is an important

Outcome: percentage of contacts in the region		
	Region:	
	Home 2007	Home 2012
Years since migration	-0.04^{***}	0.04^{***}
	(0.01)	(0.01)
Intercept	0.40^{***}	0.44^{***}
	(0.03)	(0.03)
R^2	0.01	0.01
Num. obs.	1203	1203

Notes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4: 2007 and 2012 connections percentage, by migrant status and ethnicity

determinant of social networks.

6 Discussion

Our theoretical framework suggests that a number of our findings are compatible with the causality running from neighborhoods to networks. The ethnic composition in residential neighborhoods is related to that in social networks, and the relationship is stronger for stayers and weaker for migrants. In addition, the relationship is similar in Soviet-era high-rise estates, the neighborhoods that were populated in a period with little residential choice. We also show that cellular calls are mostly connecting individuals living in the same commuting area, and if stretching a longer distance, these are likely connections to the former place of residence. All these outcomes suggest that ties typically arise between individuals who are living close to each other. Most importantly, the trade-off between ethnic and geographic distance results in out-group ties, likelihood of which increases along the number of out-groups living in the same neighborhood.

Our central outcome, the positive correlation between network and neighborhood composition can be explained in other ways as well. First, networks may influence neighborhood choice, in particular individuals with certain amount of inter-ethnic contacts may choose to live in a correspondingly mixed environment. However, in this case one expects migrants to be equally sensitive to the neighborhood composition than the non-migrants, and second, one also expects the relationship to be substantially weaker in the Soviet-era neighborhoods. Neither of these predictions is true. Unfortunately, the current cross-sectional network data does not allow to test this hypothesis directly.

Alternatively, both network and neighborhood composition may be jointly determined by a third variable, such as segregation preferences. If this is true, we would not expect the geographic distance to be a strong determinant of networks, and in addition, we would also expect the relationship to be much weaker in Soviet-era neighborhoods.

7 Conclusion

We use mobile telecommunication data to analyze the relationship between physical segregation and social network network. The data originates from Estonia, a linguistically divided country where the relationship between the corresponding ethnic groups has been characterized by distrust and animosity. These data allow us to compute network homophily and analyze its relationship with the place of residence, separately for people who migrate and those who do not. We document a strong positive relationship between ethnic composition in the residential neighborhood and network homophily. We also show that communication networks are largely local, except for migrants who possess a substantial number of contacts in their previous place of residence. As more time passes since the date of migration, migrants have more contacts in their new location and fewer contacts in their previous location.

A simple theoretical framework suggests that all these outcomes can be explained by neighborhood ethnic composition being an important determinant of social network homophily. While we cannot conclusive rule out alternative explanations, we test several additional specifications that are consistent with this causal relationship.

More speculatively, our results suggest that physical integration might help generate social integration. By contrast, ethnic or racial ghettos may harm social integration, even if such separation might increase the efficiency of local labor markets (Edin, Fredriksson, and Åslund 2003, Damm 2009). Of course, our context, where two well-established ethnic groups share similar socioeconomic and cultural backgrounds, may not be generalizable to contexts where the ethnic groups are separated by much more than just language. Nonetheless, we believe these results provide empirical support for policies designed to promote physical integration.

A Proofs

Expected Homophily The observed communication process does not include the complete network data. A number of links are missing, either because they connect to out-of-sample peers, or because no calls were made during our 10 days of network sampling. Here we show that under independent link sampling, $\mathbb{E}\, h^* = \mathbb{E}[\nu_0^*/(\nu_0^* + \nu_1^*)] = \mathbb{E}\,\nu_0^*/\,\mathbb{E}(\nu_0^* + \nu_1^*)$.

Look at individual i. Assume that in the complete network she has degree $N_i = S_i + D_i$ (communication links to different alters), comprising of S_i same type links and D_i different type links. Assume that the alters are observed independently with probability p. Hence, the observed number of contacts $\tilde{N}_i$ is a random variable and can be expressed as a sum of N_i realizations of independent Bernoulli random variables $A_{ij} \sim \text{Bernoulli}(p)$ where $A_{ij} = 1$ denotes that the alter j of individual i is observed. The observed number of same type alters is $\tilde{S}_i = \sum_{j \in \mathfrak{S}} A_{ij}^S$ and that of different type alters $\tilde{D}_i = \sum_{j \in \mathfrak{D}} A_{ij}^D$ ($\mathfrak{S}$ and $\mathfrak{D}$ denote the set of same and different type friends of i). The true homophily is $H_i = S_i/N_i$ while we observe

$$\tilde{H}_i = \frac{\tilde{S}_i}{\tilde{N}_i} = \frac{\sum_{i \in \mathfrak{S}} A_i^S}{\sum_{i \in \mathfrak{S}} A_i^S + \sum_{i \in \mathfrak{D}} A_i^D}. \tag{5}$$

As A^S and A^D have equal i.i.d distribution, we can apply lemma 3 by Heijmans (1999) and conclude that

$$\mathbb{E}\left[\tilde{H}_i | \tilde{N}_i > 0\right] = \frac{\mathbb{E}\,\tilde{S}_i}{\mathbb{E}\,\tilde{S}_i + \mathbb{E}\,\tilde{D}_i} = \frac{S_i}{S_i + D_i}. \tag{6}$$

Accordingly, we can easily base our inference on the individual homophily on the observed ties in the data.

Propositions

Proposition 1 For stayers, individuals who spend long time in one region, we observe the expected equilibrium number of ties. For group 0:

$$\nu^* = \frac{1}{\delta}\left[\phi^0(\chi^0 n_0^A + \chi^1 n_1^A) + \phi^1(\chi^0 n_0^B + \chi^1 n_1^B)\right] = \tag{7}$$

$$= \frac{1}{\delta}\left\{\phi^0[\chi^0 \pi^A + \chi^1(1-\pi^A)]n^A + \phi^1(\chi^0 \pi^B + \chi^1(1-\pi^B)]n^B\right\}. \tag{8}$$

This relationship can be used to express the individual network homophily in equilibrium:

$$h^* = \frac{\nu_0^*}{\nu^*} = \frac{\nu_0^*}{\nu_0^* + \nu_1^*} = \frac{\nu_0^{A*} + \nu_0^{B*}}{\nu_0^{A*} + \nu_1^{A*} + \nu_0^{B*} + \nu_1^{B*}} =$$
$$= \frac{\frac{1}{\delta}\left(\phi^0\chi^0\pi^A n^A + \phi^1\chi^0\pi^B n^B\right)}{\frac{1}{\delta}\left\{\phi^0[\chi^0\pi^A + \chi^1(1-\pi^A)]n^A + \phi^1(\chi^0\pi^B + \chi^1(1-\pi^B)]n^B\right\}} =$$
$$= \frac{\phi^0\chi^0\pi^A n^A + \phi^1\chi^0\pi^B n^B}{\phi^0[\chi^0\pi^A + \chi^1(1-\pi^A)]n^A + \phi^1[\chi^0\pi^B + \chi^1(1-\pi^B)]n^B}, \quad (9)$$

where ν_0 and ν_1 are ties to group 0 and 1 respectively, ν^A and ν^B are ties to regions A and B, and * denotes the corresponding equilibrium values.

Look at the region A with the local ethnic composition π^A. As the number of ties in region B, ν^{B*}, does not depend on π^A, we have

$$\frac{\partial h^*}{\partial \pi^A} = \frac{\partial}{\partial \pi^A}\left(\frac{\nu_0^{A*} + \nu_0^{B*}}{\nu_0^{A*} + \nu_1^{A*} + \nu_0^{B*} + \nu_1^{B*}}\right) =$$
$$= \frac{\frac{\partial \nu_0^{A*}}{\partial \pi^A}}{\nu^*} - \frac{\nu_0^*}{(\nu^*)^2}\frac{\partial \nu^A}{\partial \pi^A} = \frac{1}{\delta}\left[\frac{\phi^0}{\nu^*}\chi^0 n^A - \frac{\phi^0}{\nu^*}h(\chi^0 - \chi^1)n^A\right]. \quad (10)$$

The first term in brackets describes the growth of ν_0^* while π^A grows, the second one the corresponding growth of ν^*. As $\chi^0 > \chi^1 > 0$ and $0 \leq h \leq 1$, the derivative is positive. It increases in the local interaction rate ϕ^0 and in the local population size.

Proposition 2 Look at movers from the region A to B. Assume they initially posses the ties, corresponding to the equilibrium in A. At time $t = 0$ they relocate to B. As ties are neither created nor destroyed instantly, we have at the moment of move $\nu_0^B(0) = \frac{1}{\delta}\phi^1\chi^0\pi^B n^B$ which is not the equilibrium value. Solving the differential equation (2) we find

$$\nu_0^B(t) = \frac{1}{\delta}\phi^0\chi^0\pi^B n^B + \mathrm{e}^{-\delta t}\left[\frac{1}{\delta}\phi^1\chi^0\pi^B n^B - \frac{1}{\delta}\phi^0\chi^0\pi^B n^B\right] =$$
$$= \frac{1}{\delta}\left[\phi^0 + \mathrm{e}^{-\delta t}(\phi^1 - \phi^0)\right]\chi^0\pi^B n^B \equiv \frac{1}{\delta}\phi_{10}(t)\chi^0\pi^B n^B. \quad (11)$$

Analogous expressions for the other types of ties will be

$$\begin{aligned} \nu_0^A(t) &= \frac{1}{\delta}\phi_{01}(t)\chi^0\pi^A n^A \qquad \nu_1^A(t) = \frac{1}{\delta}\phi_{01}(t)\chi^0(1-\pi^A)n^A \\ \nu_1^B(t) &= \frac{1}{\delta}\phi_{10}(t)\chi^0(1-\pi^B)n^B, \end{aligned} \quad (12)$$

where $\phi_{01}(t) = \phi^1 + \mathrm{e}^{-\delta t}(\phi^0 - \phi^1)$. Note that $\phi^0 > \phi_{01}(t) > \phi^1$ and $\phi^0 > \phi_{10}(t) > \phi^1 \quad \forall t > 0$.

The relationship between migrant's homophily and the new local ethnic composition, $\frac{\partial h(t)}{\partial \pi^B}$, is

$$\frac{\partial h(t)}{\partial \pi^B} = \frac{\partial}{\partial \pi^B}\left(\frac{\nu_0^B(t) + \nu_0^A(t)}{\nu_0^B(t) + \nu_1^B(t) + \nu_0^A(t) + \nu_1^A(t)}\right) = \\ = \frac{\frac{\partial \nu_0^B(t)}{\partial \pi^B}}{\nu(t)} - \frac{h(t)}{\nu(t)}\frac{\partial \nu(t)}{\partial \pi^B} = \frac{1}{\delta}\left[\frac{\phi_{10}(t)}{\nu(t)}\chi^0 n^B - \frac{\phi_{10}(t)}{\nu(t)}h(t)(\chi^0 - \chi^1)n^B\right]. \quad (13)$$

This is positive, by similar argumentation as used for the steady-state equilibrium.

Next, we show that $\phi^0/\nu^* > \phi_{10}(t)/\nu(t)$. It is determined by the sign of

$$\begin{aligned}\delta \cdot \left[\phi^0 \nu(t) - \phi_{10}(t)\nu^*\right] = \\ = \phi^0\left[\phi_{01}(t)\chi^0 n_0^A + \phi_{01}(t)\chi^1 n_1^A + \phi_{10}(t)\chi^0 n_0^B + \phi_{10}(t)\chi^1 n_1^B\right] - \\ - \phi_{10}(t)\left[\phi^1\chi^0 n_0^A + \phi^1\chi^1 n_1^A + \phi^0\chi^0 n_0^B + \phi^0\chi^1 n_1^B\right] = \\ = \phi_{01}(t)\phi^0(\chi^0 n_0^A + \chi^1 n_1^A) + \phi_{10}(t)\phi^0(\chi^0 n_0^B + \chi^1 n_1^B) - \\ - \phi_{10}(t)\phi^1(\chi^0 n_0^A + \chi^1 n_1^A) - \phi_{10}(t)\phi^0(\chi^0 n_0^B + \chi^1 n_1^B) = \\ = \mathrm{e}^{-\delta t}((\phi^0)^2 - (\phi^1)^2) > 0 \quad (14)\end{aligned}$$

where we have used the definition of $\phi_{10}(t)$ and $\phi_{01}(t)$. Accordingly, if homophily levels are comparable, $h(t) = h^*$, then the $\frac{\partial}{\partial \pi^B}h^* > \frac{\partial}{\partial \pi^B}h(t)$ as $\phi^0 > \phi_{10}(t)$. Intuitively, the stayers' networks are primarily determined by the local population composition while the other regions weight more in the movers' networks.

References

Ahas, R., S. Silm, O. Järv, E. Saluveer, and M. Tiru (2010): "Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones," *Journal of Urban Technology*, 17(1), 3–27.

Bayard, K., J. Hellerstein, D. Neumark, and K. Troske (1999): "Why are Racial and Ethnic Wage Gaps Larger for Men than for Women? Exploring the Role of Segregation," Working Paper 6997, National Bureau of Economic Research.

Collier, P. (1998): "The political economy of ethnicity," in *Annual World Bank Conference on Development Economics*, pp. 387–405.

Cutler, D. M., and E. L. Glaeser (2007): "Social interactions and smoking," Working Paper 13477, NBER, 1050 Massachusetts Avenue, Cambridge, MA 02138, USA.

Damm, A. P. (2009): "Ethnic Enclaves and Immigrant Labor Market Outcomes: Quasi-Experimental Evidence," *Journal of Labor Economics*, 27(2), pp. 281–314.

Easterly, W., and R. Levine (1997): "Africa's growth tragedy: policies and ethnic divisions," *Quarterly Journal of Economics*, 112(4), 1203–1250.

Edin, P.-A., P. Fredriksson, and O. Åslund (2003): "Ethnic Enclaves and the Economic Success of Immigrants—Evidence from a Natural Experiment," *The Quarterly Journal of Economics*, 118(1), 329–357.

European Commission (2002): "Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)," *Official Journal of European Communities*, L201, 37–47.

Everly, R. (1997): "Ethnic assimilation or ethnic diversity? Integration and Estonia's citizenship law," in *The Integration of non-Estonians into Estonian Society: History, problems and trends*, ed. by A. Kirch, pp. 106–121. Estonian Academy Publishers, Tallinn.

Heidmets, M. (1998): "The Russian Minority: Dilemmas for Estonia," *Trames*, 3(2), 264–72.

Heijmans, R. (1999): "When does the expectation of a ratio equal the ratio of expectations?," *Statistical Papers*, 40(1), 107–115.

Hess, D. B., T. Tammaru, and K. Leetmaa (2012): "Ethnic differences in housing in post-Soviet Estonia," *Cities*, 29(5), 327–333.

Kalmus, V., and M. Pavelson (2002): "Schools in Estonia as Institutional Actors and as a Field of Socialisation," in *The Challenge of the Russian Minority: Emerging Multicultural Democracy in Estonia*, ed. by M. Lauristin, and M. Heidmets, pp. 227 – 236. Tartu University Press, Tartu.

Katus, K. (1990): "Demographic trends in Estonia throughout the centuries," *Yearbook of Population Research in Finland*, 28, 50–66.

Kährik, A., and T. Tammaru (2010): "Population composition in new suburban settlements of the Tallinn metropolita area," *Urban Studies*, 45, 1055–1078.

Kulu, H., and T. Tammaru (2004): "Diverging views on integration in Estonia, determinants of Estonian language skills among ethnic minorities," *Journal of Baltic Studies*, 35, 378–401.

Lauristin, M., M. Uus, and K. Seppel (2011): "Kodakondsus, kodanikuühiskond ja rahvussuhted (Citizenship, civil society and ethnic relations)," in *Integratsioonimonitor 2011*, pp. 9–50. Kultuuriministeerium, Tallinn, EE.

Leping, K.-O., and O. Toomet (2008): "Emerging ethnic wage gap: Estonia during political and economic transition," *Journal of Comparative Economics*, 36(4), 599–619.

Massey, D. S., and N. A. Denton (1988): "The Dimensions of Residential Segregation," *Social Forces*, 67(2), pp. 281–315.

Miguel, E., and M. K. Gugerty (2005): "Ethnic diversity, social sanctions, and public goods in Kenya," *Journal of public Economics*, 89(11), 2325–2368.

Pettai, I. (2002): "Mutual tolerance of Estonians and non-Estonians (in Estonian)," in *Estonia and Estonians in comparative perspective (in Estonian)*, pp. 213–233. Tartu University Press.

Phithakkitnukoon, S., F. Calabrese, Z. Smoreda, and C. Ratti (2011): "Out of Sight Out of Mind–How Our Mobile Social Network Changes during Migration," in *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pp. 515 –520.

Schultze, J. (2011): "Contact and crisis in interethnic relations," in *The Russian Second Generation in Tallinn and Kohtla-Järve: The TIES Study in Estonia*, ed. by R. Vetik, and J. Helemäe, pp. 165–182. Amsterdam University Press, Amsterdam, NL.

Tammaru, T. (2001): "Suburban growth and suburbanisation under central planning: The case of Soviet Estonia," *Urban Studies*, 38(8), 1341–1357.

Tammaru, T., and H. Kulu (2003): "The ethnic minorities of Estonia: changing size, location, and composition," *Eurasian Geography and Economics*, 44(2), 105–120.

Vihalemm, T. (2010): "To learn or not to learn? Dilemmas of linguistic integration of Russians in Estonia," *Russian Minorities in the Baltic States*, 2, 74–98.