Segregation and "Silent Separation":

Using Large-Scale Network Data to Model the Determinants of Ethnic Segregation

Joshua E. Blumenstock
University of Washington

Ott Toomet
Tartu University

Abstract

Ethnic segregation is a striking feature of many contemporary societies. Yet, due in part to constraints of existing data, empirical efforts to study the determinants of segregation have been limited. In this paper, we exploit a novel source of data to model the impact of migration and urbanization on segregation in a small society.

The focus of our empirical analysis is Estonia, a country whose complex geo-political history has led to a status quo where two dominant ethnic groups – ethnic Estonians and ethnic Russians – coexist in a stratified and segregated society. We analyze the complete mobile phone records of hundreds of thousands of Estonians, which allows us to observe the ethnicity of each individual (Russian or Estonian), the complete history of locations visited by each individual, and every phone-based interaction that takes place over the network. Together, these features offer a unique perspective on the structure of ethnic segregation in Estonia, as well as the role that migration plays in processes of segregation.

Initial results indicate that the ethnic composition of an individual's physical neighborhood is highly correlated with the ethnic composition of the individual's social network: people who are physically surrounded by co-ethnics are more likely to be in phone contact with co-ethnics. We further find that patterns of segregation are significantly different for migrants than for the at-large population: migrants are more likely to interact with coethnics than non-migrants, but are less sensitive to the ethnic composition of their immediate neighborhood than non-migrants. Interpreted these results with a nested search-based model of friendship formation, we test between different determinants of ethnic segregation in Estonia.

Background and Motivation

Ethnic diversity, segregation, and fractionalization have long been thought to play a critical role in the socioeconomic structure and overall stability of our societies. Much of the literature agrees that ethnic diversity hinders economic development (Easterly and Levine 1997; Collier 1998) through reduced investment in human capital, inefficient labor markets, increased violence and corruption, as well as broader patterns of inequality, prejudice, and discrimination (cf. Cutler and Glaeser 2007; Bayard et al. 1999; Miguel and Gugerty 2005). As a result, a great deal of attention has focused on polices to promote integration and interaction in heterogeneous societies (Castles, Miller, and Ammendola 2005).

The extent and nature of segregation in urban areas is intimately connected to patterns of migration, and to the manner in which migrants are able to integrate into the destination community. Migrants often choose to migrate to areas where their networks are stronger (Munshi 2003), existing residents often strive to keep out dissimilar immigrants (Schelling 1971), and political and institutional forces often prevent integration of migrants across ethnic lines (Yinger 1986; Clark 1986). More recent evidence suggests that ethnic concentration can hamper social interaction (Vermeij, Van Duijn, and Baerveldt 2009; Semyonov and Glikman 2009), while the relative

size of immigrant populations can affect social integration (Martinovic, Van Tubergen, and Maas 2009; Vervoort and Dagevos 2011).

While migration is thus commonly believed to play a key role in determining patterns of segregation, there exists only scant empirical evidence to illustrate how this process occurs. This gap in the literature is due largely to the coarse data that has historically been used to study segregation, as the vast majority of empirical studies rely on census and household survey datasets, which can detect cross-sectional levels of residential segregation, but which typically lack thorough modules on migration and social interaction. Posner (2004), for instance, notes that "most measures of ethnic diversity... are inappropriate for testing [the effects] of ethnic diversity" (p.849). A recent review article by Blattman and Miguel (2010) similarly concludes with a "plea for new and better data" (p.3).

Context

The context for our research is Estonia, a country with a long and complex history of ethnic strife and resettlement. Prior to World War II, roughly 94% 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% 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. Critical to our current analysis, linguistic identity 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 and Kulu 2003). 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.

Data

Our empirical analysis is based on a large repository of mobile communications data that has been collected by one of the major cellular service providers in Estonia. We have negotiated an agreement with the cellular provider in order to access this dataset for the purposes of this project. These data contain records of several hundred million interactions between roughly one million individuals over a period of six years. Three features of this dataset are critical to the current study: [1]

i. **Language Preference:** Although each subscriber in the dataset is anonymous, we observe the preferred language of each subscriber. In Estonia, a bilingual economy where native Estonians speak Estonian and native Russians speak Russian, this allows us to very accurately infer the ethnicity of each subscriber.
ii. **Social Network Structure:** The data consists of hundreds of millions of interactions between individuals, where each interaction contains the unique identifier of the calling party and the receiving party. We reconstruct the complete social graph by modelling these interactions as directed edges in a network, where each individual is a node and each edge is a communication event.
iii. **Geographic Location:** For each call or SMS event in the dataset, we additionally observe the time and approximate location of the subscriber involved in the event. The location is specified at the level of the cell tower, which gives us a spatial resolution ranging from a few hundred meters in dense urban settings to several kilometers in rural areas.

[1] The privacy and anonymity of all individuals in this dataset is of paramount importance to both the mobile operator and to the researchers working on this project. The researchers never touch the raw data, which remains on the operator's facility in Estonia, and which is completely anonymized even at the operator's facility. In practice, we develop statistical scripts and computer programs using "fake" data, send these scripts to the operator, and they return the output from these scripts.

Methods

A.1. Measuring segregation and migration

Segregation: We measure two distinct types of segregation in our data: *social segregation*, defined by the set of contacts with whom each individual communicates on the phone network; and *geographic segregation*, defined by the extent to which individuals of different ethnicities reside or work in the same geographic location. The language information collected by the operator, allows us to assign an ethnicity to each individual in the dataset (either Estonian or Russian). We then define ethnic homophily as $h_i = s_i/(s_i + d_i)$, where s denotes the number of coethnics in the network of i, and d denotes the number of individuals of different ethnicity in i's network.

Migration: Given a continuous sequence of locations for all individuals over several years, we can identify the home and work locations for cellphone users using the anchor-point methodology of Ahas et al. (2010). Using methods developed in Blumenstock (2012), we are able to identify *migrants* in the population. This technique allows us to measure the date and trajectory of each migration, and construct a binary indicator variable M_{it} that indicates for each individual i whether he or she migrated at time t.

A.2. Integration and segregation of migrants

Using the above measure of homophily h_{it} as an indicator of the extent to which individual i prefers to associate with coethnics over non-coethnics at time t, and taking M_{it} to be a binary indicator of whether i migrates at time t, a simple fixed-effects model can be used to study the effect of the migration on i's preference for coethnics:

$$h_{it} = \alpha_1 + \sum_{s=0}^{T} \beta_s M_{i(t-s)} + \mu_i + \pi_t + \epsilon_{it} \qquad (1)$$

where μ_i is an individual-specific fixed effect that accounts for the fact that each individual may have a different preference for segregation at baseline, and π_t is a time-specific fixed effect to reduce bias caused by common trends across all individuals over time. The coefficients on the β_s thus indicate the extent to which migrations in the past T periods (for s={0,1,...,T}) lead to changes in current homophily for the average migrant. By analogy, adding heterogeneous effects to Model (1) permits us to identify which types of migrants are more or less likely to be integrated into communities of coethnics and non-coethnics.

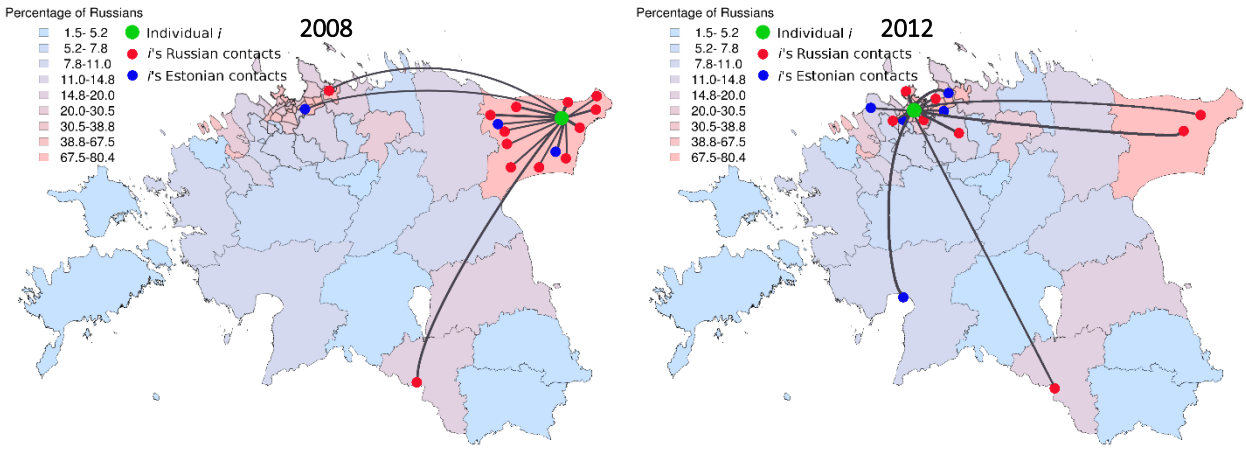

Figure 1: Location of a single Russian migrant (green dot) in Estonia, pre-migration (left panel), and post-migration (right panel). The migrant's network of Russian (red dots) and Estonian (blue dots) contacts changes significantly after migration.

Empirical Results (Preliminary)

Using a sample of data from the operator, we have begun to perform this portion of the analysis. A simple visualization of the core results can be seen in Figure 2, which depicts changes in the structure of a representative migrant's social network. Estimating a simple form of Model (1), we find the following preliminary results:

1. We find that the ethnic composition of an individual's physical neighborhood is highly correlated with the ethnic composition of the individual's social network (Table 1, column 1). Namely, people who are physically surrounded by co-ethnics are more likely to be in phone contact with co-ethnics. We model this relationship nonparametrically in Figure 3 by plotting, for 50,000 different individuals, the relationship between social homophily and the fraction of co-ethnics in the immediate geographic region. The effect is stronger for Estonians, but is positive and statistically significant for both ethnic groups.
2. In general, migrants are more likely to interact with coethnics than non-migrants (Table 1, column 2). However, migrants are *less sensitive* to the ethnic composition of their immediate neighborhood than non-migrants (Table 1, column 2, interaction term). Russian migrants are even less sensitive to physical surroundings, and the effect is strongest for recent migrants (results not shown).
3. The geographic structure of an individual's network is dependent upon the ethnic composition of the immediate neighborhood, and the past history of migration. In particular, people who are surrounded by co-ethnics are more likely to have phone contacts who live close by, and are less likely to be in contact with people living more than 10km away. For migrants the effect is the opposite (Table 1, cols. 3 and 4).

A Simple Model of Segregation

To interpret these results, we develop a nested search-based model of friendship formation. We and assume each individual i's likelihood to form contact with another individual j depends on the pairwise physical (Euclidean) distance d_{ij}^{p}, network (geodesic or shortest-path) distance d_{ij}^{n}, and ethnic distance d_{ij}^{e}, where $d_{ij}^{e} = 0$ if i and j are co-ethnic and 1 otherwise. The utility of an agent thus depends on the numbers of contacts of the same and different types, with penalties imposed for forming (or maintaining) contact with distant individuals. Aggregating decisions through this matching process allows us to solve for different steady-state equilibria that depend on the geographic concentration of different ethnicities, as well as preference parameters that specify the extent to which individuals weight different types of distance (cf. Currarini, Jackson, and Pin 2009).

Table 1: Relationship between ethnic neighborhood composition, homophily, and geographic structure of network

	(1) Homophily	(2) Homophily	(3) # Local Contacts	(4) # Distant Contacts
pctCoethnic	0.327***	0.364***	0.281***	-0.149***
	(0.01)	(0.02)	(0.04)	(0.05)
Migrant		0.137***	-0.128***	0.122***
		(0.02)	(0.02)	(0.02)
Migrant*PctCoethnic		-0.147***		
		(0.03)		
(Intercept)	0.685***	0.657***	0.187***	0.679***
	(0.01)	(0.02)	(0.03)	(0.04)
N	22439	4703	8075	8075

Notes: "pctCoethnic" indicates the percentage of people living in *i*'s vicinity of the same ethnicity. "Migrant" indicates whether *i* is a migrant, and the third row denotes the interaction. An individual's homophily h_i is the dependent variable in columns (1) and (2), while the number of local (within 10km) and distant (>10km) network connections are the dependent variables in columns (3) and (4), respectively. Standard errors in parenthesis. *** p < 0.001, **p < 0.01, *p < 0.05.

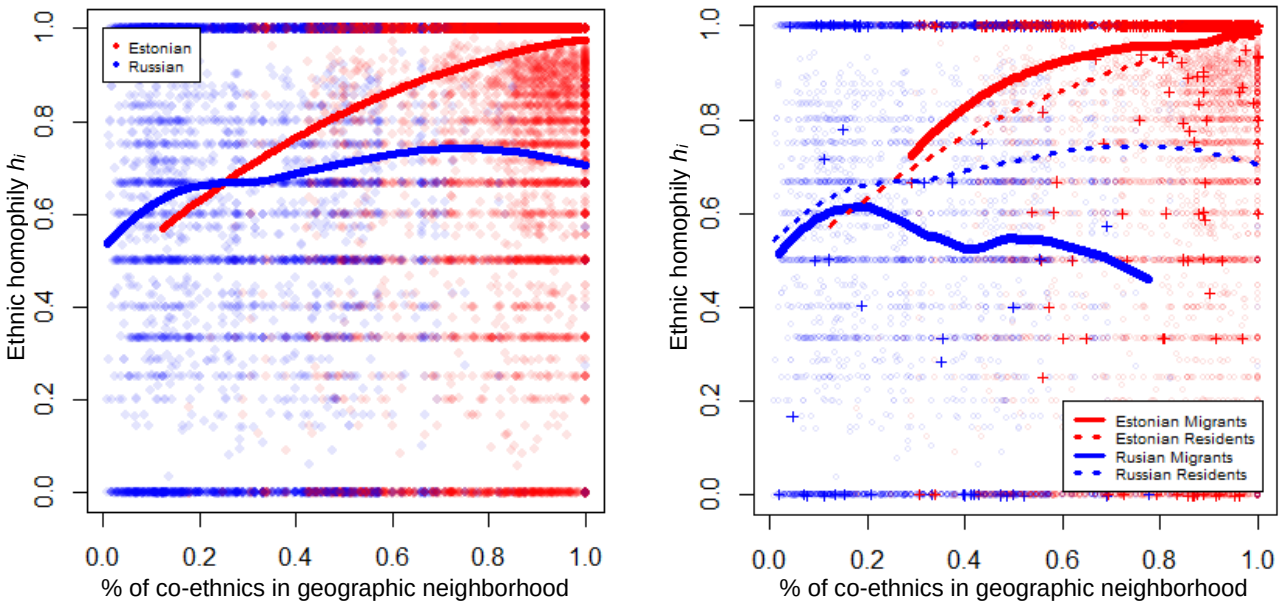

Figure 2: Ethnic homophily as a function of geographic segregation for 50,000 individuals. Left figure disaggregates by ethnic group for 50,000 individuals, right figure further disaggregates for migrants.

References

Ahas, Rein, Siiri Silm, Olle Järv, Erki Saluveer, and Margus Tiru. 2010. "Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones." *Journal of Urban Technology* 17 (1): 3–27.

Bayard, Kimberly, Judith Hellerstein, David Neumark, and Kenneth Troske. 1999. "Why Are Racial and Ethnic Wage Gaps Larger for Men Than for Women? Exploring the Role of Segregation". Working Paper 6997.

Blattman, Christopher, and Edward Miguel. 2010. "Civil War." *Journal of Economic Literature* 48 (1) (March): 3–57. doi:10.2307/40651577.

Blumenstock, Joshua E, and Lauren Fratamico. 2013. "Social and Spatial Ethnic Segregation: A Framework for Analyzing Segregation With Large-Scale Spatial Network Data." In *The 4th Annual Symposium on Computing for Development*. Cape Town, South Africa.

Blumenstock, Joshua Evan. 2012. "Inferring Patterns of Internal Migration from Mobile Phone Call Records: Evidence from Rwanda." *Information Technology for Development* 18 (2): 107–125.

Blumenstock, Joshua, Ye Shen, and Nathan Eagle. 2010. "A Method for Estimating the Relationship Between Phone Use and Wealth." QualMeetsQuant Workshop at the 4th International IEEE/ACM Conference on Information and Communication Technologies and Development.

Castles, Stephen, Mark J. Miller, and Giuseppe Ammendola. 2005. "The Age of Migration: International Population Movements in the Modern World" New York: The Guilford Press.

Clark, William AV. 1986. "Residential Segregation in American Cities: A Review and Interpretation." *Population Research and Policy Review* 5 (2): 95–127.

Collier, Paul. 1998. "The Political Economy of Ethnicity." In *Annual World Bank Conference on Development Economics*, 387–405.

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

Cutler, David M., and Edward L. Glaeser. 2007. "Social Interactions and Smoking". Working Paper 13477. 1050 Massachusetts Avenue, Cambridge, MA 02138, USA: NBER.

Easterly, William, and Ross Levine. 1997. "Africa's Growth Tragedy: Policies and Ethnic Divisions." *The Quarterly Journal of Economics* 112 (4): 1203–1250.

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

Miguel, Edward, and Mary Kay Gugerty. 2005. "Ethnic Diversity, Social Sanctions, and Public Goods in Kenya." *Journal of Public Economics* 89 (11): 2325–2368.

Munshi, K. 2003. "Networks in the Modern Economy: Mexican Migrants in the US Labor Market." *Quarterly Journal of Economics* 118 (2): 549–599.

Schelling, Thomas C. 1971. "Dynamic Models of Segregation." *Journal of Math. Sociology* 1 (2): 143–186.

Semyonov, Moshe, and Anya Glikman. 2009. "Ethnic Residential Segregation, Social Contacts, and Anti-minority Attitudes in European Societies." *European Sociological Review* 25 (6): 693–708.

Tammaru, Tiit, and Hill Kulu. 2003. "The Ethnic Minorities of Estonia: Changing Size, Location, and Composition." *Eurasian Geography and Economics* 44 (2): 105–120.

Vermeij, Lotte, Marijtje AJ Van Duijn, and Chris Baerveldt. 2009. "Ethnic Segregation in Context: Social Discrimination Among Native Dutch Pupils and Their Ethnic Minority Classmates." *Social Networks* 31 (4): 230–239.

Vervoort, Miranda, and Jaco Dagevos. 2011. "The Social Integration of Ethnic Minorities: An Explanation of the Trend in Ethnic Minorities' Social Contacts with Natives in the Netherlands, 1998–2006." *Journal of Ethnic and Migration Studies* 37 (4): 619–635.

Yinger, John. 1986. "Measuring Racial Discrimination with Fair Housing Audits: Caught in the Act." *The American Economic Review*: 881–893.