Modeling Homophily in Social Networks: Evidence from Hungarian Schools

by

Dániel Ferenc Molnár

submitted to

Central European University

Department of Economics

in partial fulfillment of the requirements for the degree of Master of Arts in Economics

Supervisor: Prof Sergey Lychagin

Budapest, Hungary

2017

Abstract

In my thesis I examine homophily in Hungarian schools. I use a dataset of friendship networks of students in eighth grade. I find positive inbreeding homophily for both Roma and non-Roma students. Estimating a linear probability model shows that ethnicity, gender and parental education are strong determinants of friendship formation. I construct a micro-level model of friendship formation based on the model by Mayer and Puller (2008). I find that same-ethnicity preferences are responsible for part of the bias in friend composition. Using the calibrated model I conduct a counterfactual experiment of classroom desegregation. I find that this would lead to a reduction in the same-ethnicity bias of friend composition.

My thesis has four sections. Section 1 gives an introduction, presents the relevant literature and the background of my topic. Section 2 describes the data. Section 3 shows the methods of analysis and the results. Section 4 concludes.

Keywords: social networks, racial homophily, Roma, Hungary

Acknowledgements

I would like to thank my supervisor, Prof Sergey Lychagin for his help and comments.

I am grateful to my family and friends for all their love and support during my studies.

Table of contents

List of tables

List of figures

1. Introduction

The reader may well know the Latin phrase *"Similis simili gaudet."* or the English saying *"Birds of a feather flock together."* Both express the notion that people like the company of similar individuals. This phenomenon has been studied in the social sciences, and it is usually referred to as homophily. Homophily may exist along many dimensions, but my thesis focuses on racial homophily. Racial homophily is present when individuals prefer the company of individuals of the same race or ethnicity. My research question is the following: Do same-ethnicity preferences explain ethnic homophily in the social networks of Hungarian students? I answer this question by analyzing the data and using a model of friendship formation. Before this, I present the relevant literature of analyzing social networks and modeling homophily in the following paragraphs.

Baerveldt et al. (2004) studied school segregation in the Netherlands. Their data comprised friendship links and supporting interactions reported by the students. They differentiated four ethnicities based on parents' ethnicities: Dutch, Turkish, Moroccan and Surinamese. They found evidence of ethnic boundaries. Students of all four ethnicities would interact the most with students of the same ethnicity. But Surinamese students were also likely to interact with Dutch students, which can be explained by cultural or linguistic proximity (Suriname used to be a Dutch colony).

Foster (2005) examined the behavior of undergraduate students. Originally students were randomly assigned to housing. Later it was observed how they sorted into residences and what friendship links they formed. Foster's analysis shows that the original campus location, high school SAT scores and ethnicity are the main determinants of friendship formation.

Currarini et al. (2009) proposed a formal model of homophily. Their model is based on microeconomic optimization by the agents. Agents may choose how long they take part in the process of looking for friends, and they have utility functions based on the composition of their friends (same vs. different type). Their model generates homophily if there is type-sensitivity of preferences and bias in the matching process. In Currarini et al. (2010), they used this model to analyze friendship patterns in US high schools. They found strong ethnic biases in the matching process for Asians and Blacks. Also, Blacks and Hispanics had the strongest bias in preferences.

With the advent of social media, such as Facebook.com, researchers gained access to information about friendship networks that had not been possible before. The following three papers all used data from Facebook.com to study friendship formation. Lewis et al. (2008) analyzed Facebook data of students from a college in the USA. They found that all minorities (Black, Hispanic, Asian) had more heterogeneous networks than Whites. They also found evidence of homophily along tastes in music, movies and books. In a follow-up to the previous study, Wimmer and Lewis (2010) used exponential random graph modeling to analyze the data further. They found evidence of homophily along race, ethnicity and finer microethnic subgroups as well.

The third study using Facebook data is the one by Mayer and Puller (2008). They used data from ten colleges in Texas. They used a linear probability model to find that race, socioeconomic background, gender and hobbies are significant determinants of friendship links. Then they used administrative data from Texas A&M to complement the Facebook data. They created a formal model of friendship formation where individuals meet randomly, then meet friends of friends, and conditional on meeting becoming friends is driven by a

utility function based on racial and socioeconomic similarity. They calibrated this model to match the moments of the real data. The calibrated model shows that preferences play a major role in producing homophily in the networks.

Hajdu et al. (2015) studied the friendship networks of Roma students in Hungary. They found that Roma students who were high achievers had significantly more non-Roma friends and the same number of Roma friends, compared to Roma students who were low achievers.

The literature that I presented all support the existence of racial homophily. There are multiple different modeling approaches, but in my opinion the most promising is the use of big data from social media platforms (e.g. Facebook.com).

2. Data

The data that I examine was previously used by Hajdu et al. (2015). I got the data and permission to use it from the authors. The dataset contains information about 3681 Hungarian students enrolled in 8th grade. The data was collected in April 2010. They surveyed 88 elementary schools, but they excluded schools from Budapest.

The dataset contains information about students' friendship links and background characteristics. Each student could nominate five male and five female friends. This allows me to reconstruct the friendship network within each class. There are 208 classes in total.

Please see Table 1 for class-level descriptive statistics. It can be seen that there is variation in class size and ethnic composition among the classes. I will use this variation in ethnic composition to analyze homophily.

Table 1: Class-level statistics

	mean	SD	min	max
class size	17.88	5.31	2	29
Roma (%)	24.44	24.11	0	1
female (%)	46.63	14.09	0	77.27

I present some student-level descriptive statistics in Table 2. These statistics are based on the survey responses of the individual students. All variables apart from age and GPA are dummy variables. The mean age is 14.75 years. The typical age of 8th grade students in Hungary is 14-15 years. The relatively big standard deviation is caused by students who failed and had to repeat grades. GPA is the average off all subjects taken. Having more than 150 books at home is meant to proxy family background and education. The variables for college education of parents has a similar purpose. It is striking that 84% of the students surveyed had no parent with college education. The last variable is a dummy which is one if the student comes from a neighborhood where "most people are well-off" or "most people are rich."

Table 2: Student-level statistics

	mean	SD
Roma	0.2174	0.4126
female	0.4766	0.4995
age	14.7519	0.6362
GPA	3.5886	0.8971
having 150+ books at home	0.3328	0.4713
neither parent has college education	0.8390	0.3676
one parent has college education	0.1094	0.3121
both parents have college education	0.0516	0.2212
well-off neighborhood	0.2207	0.4148

3. Analysis and results

In this section I present three methods of analysis and their results: the inbreeding homophily index, linear probability models of friendship, and a micro-level model of friendship formation.

3.1. Homophily and inbreeding homophily

It is important to quantify the extent of homophily in these networks. I use the homophily index and the inbreeding homophily index as defined by Coleman (1958). Suppose that the group being examined consists of disjoint subgroups. In my case the group is a class, and the subgroups are the Roma and non-Roma students in that given class. Let subgroups be indexed by *i*. Then the homophily index of subgroup *i* is

$$H_i = \frac{s_i}{s_i + d_i}$$

where s_i is average number of friends from the same subgroup and d_i is the average number of friends from different subgroups. However, the homophily index may give a misleading picture because it does not take into account the population share of subgroup *i*. Let the population share be denoted by w_i. Then the inbreeding homophily index of subgroup *i* is

$$IH_i = \frac{H_i - w_i}{1 - w_i}$$

Here a positive value shows that individuals have more in-group friends than what would be expected based on the subgroup's size in the population.

I calculated the inbreeding homophily index of both subgroups in the whole sample:

- $IH_{Roma} = 0.3123$

- $IH_{non\text{-}Roma} = 0.3987$

There is positive inbreeding homophily in both subgroups. Inbreeding homophily is somewhat higher for the non-Roma subgroup. Potential explanations of this are given in the next subsection.

3.2. Linear probability model

In order to investigate the potential causes of homophily, I estimate two linear probability models following Mayer and Puller (2008). The sample includes every possible pair of students within every classroom. This gives an effective sample size of 34,323. The dependent variable is 1 if the two students are friends in the dataset, and it is 0 otherwise. I used MATLAB to estimate two linear probability models with bootstrap standard errors. Students were resampled only within their own class, and there were 500 repetitions. The results are summarized in Table 3.

Column 1 contains the linear probability model where I control for various measures of similarity between two students. “both Roma” and “Roma and non-Roma” control for the ethnicity of the students. The omitted category is “both non-Roma”. “same town” indicates that both students have the same town as their permanent address. I included this variable for the following reason: if students commute from multiple smaller towns to the same school in a bigger town, it might be that they are friends because they knew each other earlier or they commute together. The variables “collegeXY” indicate that one of the students has X parents with college education and the other student has Y parents with college education. The omitted category is “college00”, i.e. both students have no college educated parents.

Table 3: Linear probability models

	(1)			(2)		
friendship	**beta**		**SE**	**beta**		**SE**
constant	0.4996	***	0.0079	0.1309	***	0.0072
common friends				0.0812	***	0.0007
both Roma	0.1238	***	0.0100	0.1367	***	0.0092
Roma and non-Roma	-0.0818	***	0.0066	-0.0331	***	0.0055
same gender	0.0943	***	0.0055	0.1031	***	0.0044
same town	0.0296	***	0.0057	0.0061		0.0057
age difference in months	-0.0030	***	0.0004	-0.0002		0.0004
GPA difference in points	-0.0881	***	0.0036	-0.0572	***	0.0033
both having 150+ books	0.0336	***	0.0082	0.0092		0.0071
college22	0.0814	**	0.0362	0.0476		0.0339
college11	0.0766	***	0.0204	0.0273		0.0171
college21	0.0110		0.0201	-0.0256		0.0183
college20	-0.0248	**	0.0099	-0.0324	***	0.0085
college10	0.0024		0.0070	-0.0060		0.0062
both from well-off neighborhood	0.0414	***	0.0114	0.0105		0.0105
N	34,323			34,323		
R^2	0.0482			0.2579		
The omitted categories are "both non-Roma" and "college00".						
Standard errors are bootstrap standard errors (B=500).						
***: p<0.01						
**: p<0.05						
*: p<0.10						

Before I analyze the estimated coefficients, I must emphasize that these coefficients cannot be given a causal interpretation. They just explain some of the variation in the presence of friendship links. The model in Column 1 shows that two Roma students are 12 percentage points more likely to be friends. There is also a significant negative coefficient on the indicator of different ethnicity. Students of the same gender or coming from the same town also have a higher chance of being friends. Dissimilarity in age and GPA have a negative coefficient. Parental education is also significant.

The picture is different if the regression also controls for the number of common friends. Please see Column 2 of Table 3 for the results. The only factors that remain significant are ethnicity, gender and dissimilarity in parental education. This model has $R^2 = 0.2579$, which is higher than the R^2 of the model in Column 1, but still most of the variance in friendship links remains unexplained. This is consistent with the theory that there are hidden factors influencing friendship formation, or that most of it is driven by pure luck (Foster 2005; Mayer and Puller 2008).

3.3. Micro-level model of friendship formation

3.3.1. Setup

In this subsection I present a model which simulates meeting and becoming friends. It is based on the model by Mayer and Puller (2008). I implemented the model and the simulations in Python.

The model works the following way. In the beginning no students are friends in the class. Then students meet each other with a given probability (P_{meet}). Meeting is independent across pairs of students. If two students meet, they become friends if the friendship function takes on a positive value. The friendship function is

$$f(X_{ij}, \varepsilon_{ij}) = X'_{ij}\beta + \varepsilon_{ij}$$

where X_{ij} contains measures of similarity and dissimilarity between students i and j, β contains parameters (including an intercept) and ε_{ij} is a shock term (iid standard normal). This happens in each classroom. Afterwards the resulting friendship networks are observed.

3.3.2. Calibration and results

The model outlined above can be thought of as a function of the form

$$\mu = g(P_{meet}, \beta, \omega)$$

where μ contains moments of the resulting networks, P_{meet} and β are parameters, and ω includes all random shocks (random values for the meeting process and shocks for the friendship function). Let the shocks be fixed. The model can be calibrated by trying different parameter values and checking which produce moments that are the closest to the true values of the moments. I implemented this in Python the following way: parameter values were chosen at random, then a solver function with the Nelder-Mead algorithm tried to find the local minimum. This was repeated two hundred times. From all these local minima I chose the one with the closest match.

The model that I used had four parameters and four moments. (I based this on the model by Mayer and Puller (2008), but their model used fourteen parameters and fourteen moments because they had a much richer set of data.) The four parameters were the probability of meeting (P_{meet}), the intercept of the friendship function (β_0), the coefficient on the indicator for "both Roma" (β_{RR}), and the coefficient on the indicator for "both non-Roma" (β_{NN}). The targeted moments were the mean and variance of the number of friends, the percentage of friends of Roma who are Roma, and the percentage of friends of non-Roma who are non-Roma. My results are summarized in Table 4.

Column 1 shows the moments that are observed in the real data. Column 2 shows the calibrated model. The model fits nicely the moments of the data. It can be seen that much

Table 4: Model results

	(1) Data	(2) Model	(3) Model without preferences	(4) Experiment
Parameters				
P_{meet}		0.300	0.498	0.300
β_0		0.351	-0.062	0.351
β_{RR}		2.655	0.000	2.655
β_{NN}		0.562	0.000	0.562
Moments				
mean	8.59	8.61	8.66	8.34
var	11.28	11.27	11.20	10.61
RR	0.46	0.46	0.34	0.25
NN	0.87	0.87	0.84	0.82

greater same-ethnicity preference is needed for the Roma subgroup than for the non-Roma subgroup.

I also calibrated the model without same-ethnicity preferences. This time I only calibrated the probability of meeting and the intercept of the friendship function, while the targeted moments were the mean and variance of number of friends. This is summarized in Column 3 of Table 4. This is essentially the case of random friendship formation without same-ethnicity preferences. This version of the model reproduces the targeted moments, and the same-ethnicity bias in the composition of friends is also smaller for both subgroups.

Mayer and Puller (2008) used their calibrated model to conduct counterfactual experiments. I conduct a counterfactual experiment with my version of the model. I rerun the simulations using a modified dataset where each class has the same share of Roma students (21.74%). The counterfactual experiment is meant to simulate the effect of desegregation. The results are summarized in Column 4 of Table 4. The mean and variance of the number of friends does not change drastically, but there is a great reduction in the same-ethnicity biases. My

results are similar to those of Hajdu et al. (2015). They used simulations to estimate the effect of equalizing the share of Roma students in all classes. Their simulations were based on estimating the expected number of friends based on class, gender and ethnicity. They also found that desegregation would have positive effects on friendship ties between Roma and non-Roma students.

3.3.3. Sensitivity of the model

I produced some comparative statics to understand better the workings of my model. I used the model in Column 2 of Table 4. In each case I only changed the value of one parameter while leaving the other three unchanged at their calibrated level. Please see Figure 1-4 for the results.

Figure 1

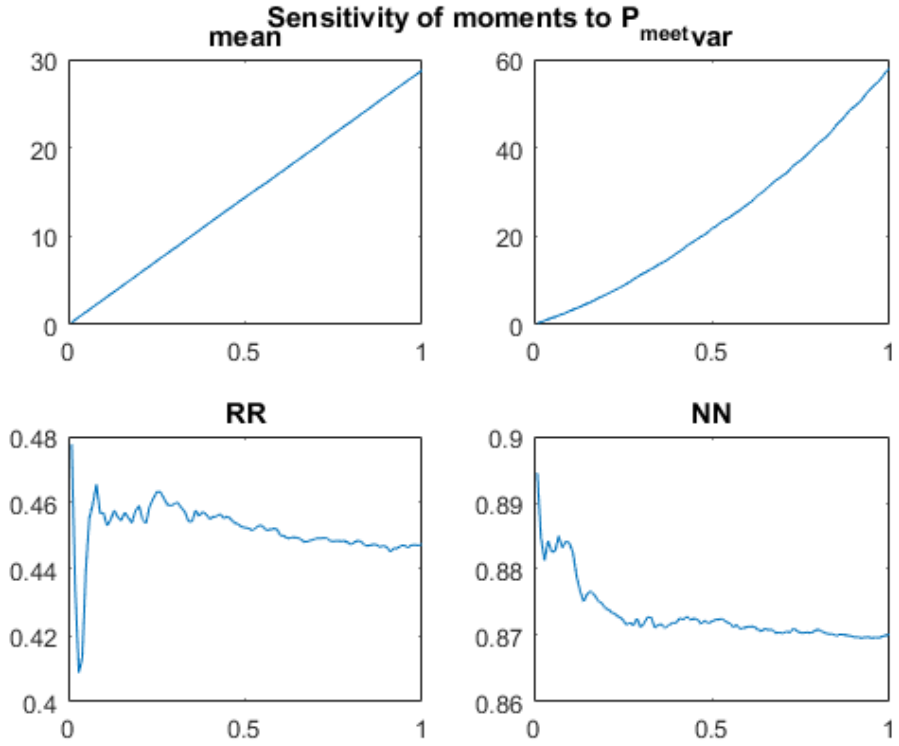

Figure 2

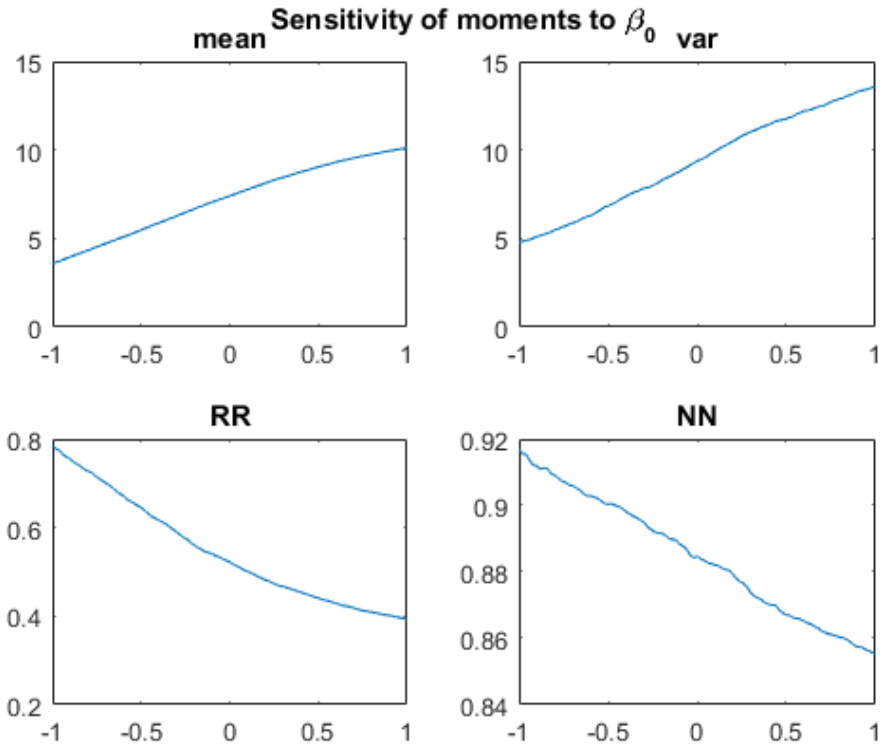

Figure 3

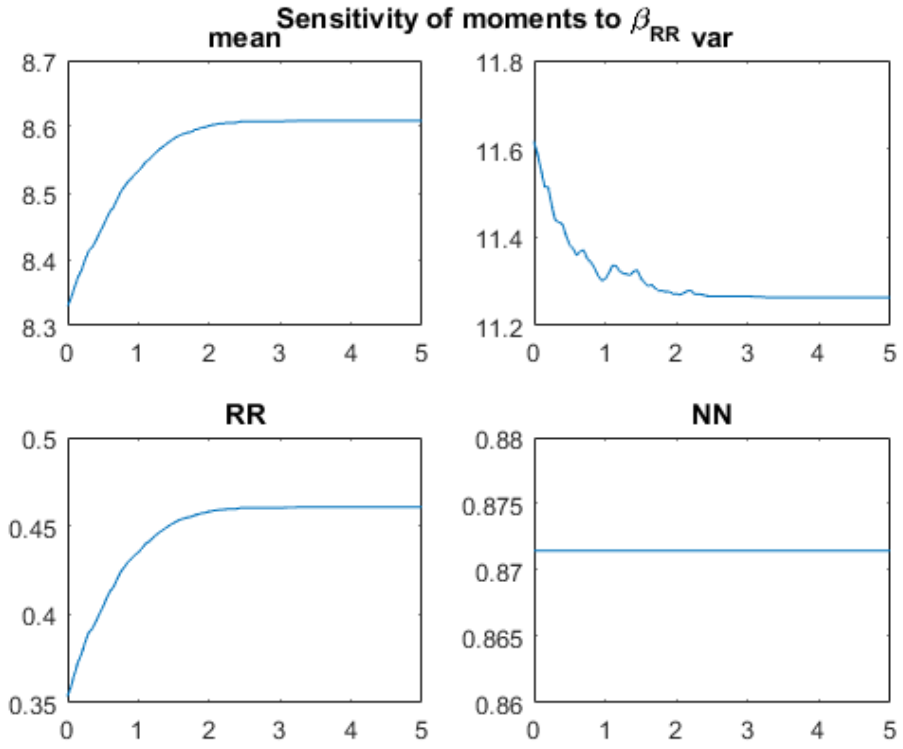

Figure 4

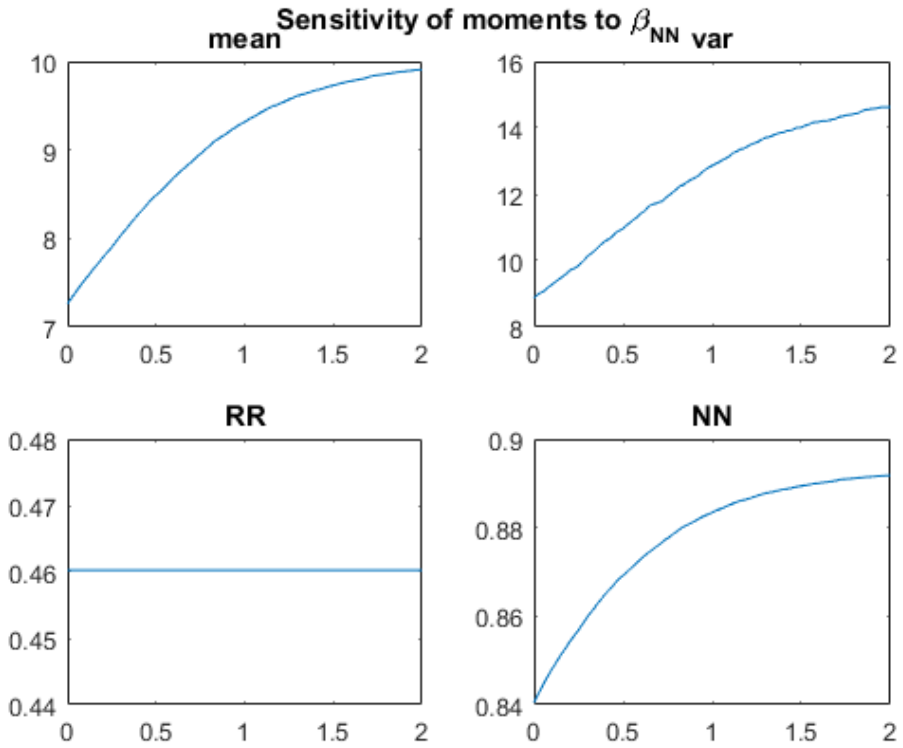

Figure 1 shows that increasing the probability of meeting leads to higher mean and variance of the number of friends. The same-ethnicity biases decrease.

Figure 2 shows that increasing the intercept leads to students forming more friendships. As the intercept increases, the importance of the same-ethnicity preference coefficients decreases. This leads to a decrease in the same-ethnicity biases.

Figure 3 shows that increasing the coefficient of Roma same-ethnicity preferences leads to Roma students becoming friends more often. After a certain value, it does not matter if the coefficient is increased. I suspect that at that point all Roma students become friends upon meeting. All Roma students have more friends, so the variance of the number of friends decreases. As the coefficient increases, the same-ethnicity bias of Roma becomes stronger. It does not affect the bias of non-Roma students.

Figure 4 shows similar results to those of Figure 3. One difference is that increasing the coefficient does not decrease the variance of the number of friends, but increases it.

3.3.4. Comments and future research

I think that this model may be further expanded in other directions. It could include more moments (for example clustering, skewness of the number of friends, etc.). The reasons why I did not include more moments and more parameters in the model are the following: Firstly, my main point of interest was same-ethnicity bias. This aspect is reproduced well by the model. Secondly, another version of the model which included a process of meeting friends of friends could not be calibrated due to computational issues.

I think that in future research I could improve on the number of moments (supposing that the computation issues can be solved) or on the data being used (e.g. Facebook data or other social media).

4. Conclusion

In my thesis I analyzed a dataset of Hungarian students in 8th grade. The dataset allowed me to reconstruct class-level friendship networks. I set out to examine ethnic homophily with respect to Roma and non-Roma students.

I found positive inbreeding homophily for both subgroups. I estimated linear probability models to find the strongest predictors of friendship links. I found that ethnicity, gender and parental education are good predictors. I calibrated a model based on the model by Mayer and Puller (2008). My model could reproduce the same-ethnicity friendship biases found in the data. Same-ethnicity preferences explain part of this bias. I conducted a counterfactual experiment in which I simulated the effect of desegregation. The simulation showed a decrease in same-ethnicity friendship biases.

References

Baerveldt, Chris, Marijtje A.J Van Duijn, Lotte Vermeij, and Dianne A Van Hemert. 2004. "Ethnic Boundaries and Personal Choice. Assessing the Influence of Individual Inclinations to Choose Intra-Ethnic Relationships on Pupils' Networks." *Social Networks* 26 (1): 55–74. doi:10.1016/j.socnet.2004.01.003.

Coleman, James. 1958. "Relational Analysis: The Study of Social Organizations with Survey Methods." *Human Organization* 17 (4): 28–36.

Currarini, Sergio, Matthew O. Jackson, and Paolo Pin. 2009. "An Economic Model of Friendship: Homophily, Minorities, and Segregation." *Econometrica* 77 (4): 1003–45. doi:10.3982/ECTA7528.

———. 2010. "Identifying the Roles of Race-Based Choice and Chance in High School Friendship Network Formation." *Proceedings of the National Academy of Sciences* 107 (11): 4857–61. doi:10.1073/pnas.0911793107.

Foster, Gigi. 2005. "Making Friends: A Nonexperimental Analysis of Social Pair Formation." *Human Relations* 58 (11): 1443–65. doi:10.1177/0018726705061313.

Hajdu, Tamás, Gábor Kertesi, and Gábor Kézdi. 2015. "High-Achieving Minority Students Can Have More Friends and Fewer Adversaries." *Budapest Working Papers on the Labor Market* 2015/7.

Lewis, Kevin, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, and Nicholas Christakis. 2008. "Tastes, Ties, and Time: A New Social Network Dataset Using Facebook.com." *Social Networks* 30 (4): 330–42. doi:10.1016/j.socnet.2008.07.002.

Mayer, Adalbert, and Steven L. Puller. 2008. "The Old Boy (and Girl) Network: Social Network Formation on University Campuses." *Journal of Public Economics* 92 (1–2): 329–47. doi:10.1016/j.jpubeco.2007.09.001.

Wimmer, Andreas, and Kevin Lewis. 2010. "Beyond and Below Racial Homophily: ERG Models of a Friendship Network Documented on Facebook." *American Journal of Sociology* 116 (2): 583–642. doi:10.1086/653658.