Measuring homophily in the choice of upper secondary education

A new socio-metric instrument

Erlend Ingridsønn Nordrum

Master’s thesis, Sociology
Department of Sociology and Human Geography Faculty of Social Sciences

UNIVERSITY OF OSLO

June 2018
Measuring homophily in the choice of upper secondary education:

A new socio-metric instrument

Erlend Ingridsønn Nordrum
Master’s thesis, Sociology
Department of Sociology and Human Geography Faculty of Social Sciences
UNIVERSITY OF OSLO
June 2018
© Erlend Ingridsønn Nordrum

2018

Measuring homophily in the choice of upper secondary education: A new socio-metric instrument

Erlend Ingridsønn Nordrum

http://www.duo.uio.no/

Print: Reprosentralen, Universitetet i Oslo

IV
Summary

The aim of this thesis is to advance the methodology of homophily research, by providing a new socio-metric instrument that can be used by a novel methodology developed by Jansson et al. (under review) (Jansson 2017). A socio-metric instrument is a tool which provide information of individuals’ relationship with others. The choice so of socio-metric instrument implies specifying the relational content one is interest in and linking socio-metric instruments to actions is not always straight forward. The socio-metric instrument I propose in this thesis is based on lower secondary school graduates’ choice of upper secondary education. Hence, the socio-metric instrument is especially useful for researchers interested in segregation in lower and upper secondary education. It can be used to test what background dimensions are important for which classmates’ the lower secondary graduates gravitate towards, in their choice of upper secondary school. This have direct implication on the demographics of upper secondary schools and can have implication on the informal segregation in the destination upper secondary school. It can also be used to approximate self-selection bias for researchers interested in estimating peer effects in lower secondary school.

This choice has two components; track and school. I use findings from previous research to optimize the social considerations of the graduates in this choice, which I operationalize into the socio-metric instrument that measures friendly coordination of choices. In summary I expect graduate friends that coordinate their choices to apply more often to the same upper secondary school. Furthermore, I expect friends to do, to a far less degree than non-friends or non-coordinating graduates, is to apply to the same specialization track but at different schools.

The socio-metric instrument is used in empirical analyses of gender and ethnic/immigrant status homophily. Coupling the socio-metric with the methodology of Jansson et al. (under review) (Jansson 2017) shows promising ability to infer homophily from registry data. The analyses show homophily on the dimensions gender and immigrant status, and that homophily is more persistent from first to second-generation versus natives among female graduates than among male graduates. An analysis of immigrant country of origin and show trends of less homophily between natives and first-generation Somali immigrants than between natives and second-generation Somali immigrants.
The thesis also provides some notions on the application of the Pearson’s correlation coefficient in the methodology developed by Jansson et al. (under review) (Jansson 2017). This involve that the coefficients maximum positive value is constrained by the number possible choices is smaller than the number of independent categories. And that, coefficient’s the negative maximum is constrained by the sizes of the independent categories.
Acknowledgements

I will first and foremost extend my gratitude to my supervisors, Gunn Elisabeth Birkelund and Edvard Nergård Larsen. Both for great support, feedback, ideas and trust in me and the project. It has been a pleasure working with the both of you!

And an extra gratitude goes to Gunn, primary supervisor, for giving me the opportunity to work with the methodology, the opportunity to attend the conference abroad and including me in the research group. This has facilitated my academic growth greatly. Hence, I would also like to extend my gratitude to everyone at the “Ethnic segregation in schools and neighbourhoods: consequences and dynamics” research group for feedback a great environment. Here I can mention: Solveig Topstad Borgen who brought to my attention invaluable information to this thesis, Torkild Hovde Lyngstad for R support, and Are Skeie Hermansen for TSD support.

Thanks to my dear sister Solfrid Ingriddatter Nordrum for reading and correcting the thesis, along with constructive feedback. Finally, I am deeply grateful to Fredrik Jansson for developing the methodology my thesis builds on, and for taking the time to answer my many questions!

Erlend Ingridsønn Nordrum

Moss, June 2015
# Table of content

List of figures ................................................................. XII

1 Introduction ..................................................................................... 1
   1.1 Aim of this thesis ................................................................. 2
   1.2 The structure of the thesis ................................................... 3

2 Previous research on homophily .................................................. 5
   2.1 Homophily – definition ......................................................... 5
   2.2 Why and how homophily emerges ........................................ 7
   2.1 Schools as a social arena ...................................................... 8
   2.2 Empirical findings on gender and ethnic homophily ............. 9
      2.2.1 Gender ....................................................................... 10
      2.2.2 Ethnicity .................................................................... 10
   2.3 Analytical strategies ............................................................. 11
      2.3.1 Data sources .............................................................. 13

3 Methodology .................................................................................. 16
   3.1 Causal model ..................................................................... 17
   3.2 Data transformation ........................................................... 21
   3.3 Network correlation coefficient ........................................ 22
   3.4 Standard errors (The quadratic assignment procedure) ...... 24
   3.5 An exemplary walk trough ................................................. 25
   3.6 Meta-analysis ..................................................................... 30
      3.6.1 The random-effects model .......................................... 31
   3.7 Presentation of result ......................................................... 33

4 Choice of upper secondary education as a socio-metric instrument ........................................... 35
   4.1 Applying for upper secondary school ................................... 36
   4.2 Influencing the choice .......................................................... 37
   4.3 Combining choices of school and track into a network-metric ........................................... 41
   4.4 Validity ............................................................................. 45
      4.4.1 External validity ........................................................... 45
      4.4.2 Internal validity ........................................................... 45

5 Empirical results ........................................................................... 47
   5.1 Dataset ............................................................................. 47
5.2 Gender .................................................................................................................................................. 51
5.1 Ethnicity (immigrant status) .................................................................................................................. 52
  5.1.1 First and second-generation versus natives, separate analyses ......................................................... 53
  5.1.2 Different gender homophily .............................................................................................................. 55
  5.1.3 Nationalities .................................................................................................................................. 59
5.2 Discussion ............................................................................................................................................ 63
6 Concluding remarks ................................................................................................................................ 65
  6.1 Suggestions for further research ......................................................................................................... 66
Litteraturliste ................................................................................................................................................. 68
Appendix: .................................................................................................................................................... 71
List of figures

Figure 1: Homophily ............................................................................................................. 17
Figure 2: Intra class similaries of classmates ................................................................. 18
Figure 3: Interpersonal classmate homophily effect ...................................................... 19
Figure 4: Inter-class similaries of classmates ................................................................. 19
Figure 5: Example data: Visual representation of adjacencies ........................................... 26
Figure 6: Example data: Intra-class adjacency matrices ............................................... 26
Figure 7: Example data: Histogram of QAP coefficients at the intra-class level .......... 27
Figure 8: Example data: Inter-class Origin adjacency matrix ........................................ 29
Figure 9: Example data: Histogram of QAP coefficients at the inter-class level .......... 29
Figure 10: Example data: The probability distribution of a JHC if there is no effect ......... 30
Figure 11: Different possible similarities in combinations of upper secondary choice .... 41
Figure 12: Extended causal model ................................................................................... 43
Figure 13: Personal classmate homophily in the extended causal model ...................... 44
Figure 14: The socio-metric instrument .......................................................................... 44
Figure 15: Reduced dataset: Histogram of schools by their number of graduation cohorts. ... 49
Figure 16: Reduced dataset: Histogram of distinct choices of upper secondary education by number of applicants. ................................................................. 51
Figure 17: Immigrant status homophily by gender and different immigrant generations .... 58
List of tables

Table 1: Example data: Class 2015 ........................................................................................................... 25
Table 2: Example data: Classes 2015 and 2016 .................................................................................. 28
Table 3: Example of result layout ......................................................................................................... 33
Table 4: Students reported most important source of information when choosing upper secondary education ........................................................................................................... 38
Table 5: Students reported considerations on the choice of upper secondary education .............. 40
Table 6: Difference between friends and family on the importance of upper secondary education by track and school ........................................................................................................... 42
Table 7: Original dataset by graduation year .......................................................................................... 48
Table 8: Reduced dataset by application age .......................................................................................... 48
Table 9: Reduced dataset by gender ....................................................................................................... 49
Table 10: Reduced dataset: Choices of upper secondary schools ....................................................... 50
Table 11: Choices of upper secondary tracks ......................................................................................... 50
Table 12: Gender homophily .................................................................................................................. 52
Table 13: Immigration status homophily ............................................................................................... 53
Table 14: Immigration status homophily. First and second-generation separate .................................. 54
Table 15: Male Immigrants versus male natives ..................................................................................... 56
Table 16: Female Immigrants versus female natives ............................................................................. 57
Table 17: Female immigrants versus female natives versus male immigrants versus male natives ................................................................................................................................. 59
Table 18: Immigrants by origin and generation ...................................................................................... 59
Table 19: First-generation immigrants by nationality versus natives .................................................... 60
Table 20: Second-generation immigrants by nationality versus natives ............................................... 60
Table 21: Immigrant status homophily by immigrant country of origin and different immigrant generations ......................................................................................................................... 61
Table 22: First-generation immigrants by nationality ............................................................................. 62
Table 23: Second-generation immigrants by nationality. ................................................................. 63
1 Introduction

As a science about human society, sociology is essentially concerned with how people interact with each other. Humans may behave different in different social situations, and knowing the properties of the individuals in isolation is not enough to know how they behave together. Hence, sociologist need to be able to observe the behavior of humans in social situations, but perhaps even more important to know is to what degree individuals will interact in a situation. Interactions can take many forms, but the form does not matter if there are no interaction. Hence, one of the most important properties of a social system would be to know who will interact with who. Or the other way around, who will not likely interact.

Social interaction often involves social homophily, the tendency of people to associate with others similar to themselves. Knowing how characteristics of the individuals influence their inclination to interact is immensely useful information.

Homophily have implications to other areas of research such as one of the fundamental challenge for peer effect literature, the selection problem. In social settings, similar people tend to attract each other. This causes positive selection where similar people tend to join or be assigned the same group. When not properly dealt with when identifying peer effects, this causes an upwards bias in magnitude of effect of peers’ behavior and effect of peers’ average background. The selection problem is basically a problem of accounting for homophily (Sacerdote 2011, Manski 1993). To policymaker homophily is of crucial importance when understanding what happens informally when an arena has formally been desegregated, such as desegregating efforts in schools.

However, measuring homophily require data on relation, and observing interpersonal relations directly for large populations is difficult. Online social networks can serve as a data source for this purpose but comes with significant access and privacy issues. And conduction social network surveys is expensive. Thus, a manageable method for inferring interpersonal homophily in a society that does not massively invade the citizens’ personal life would be a very useful tool for sociologists and policymakers. A promising attempt to address this have been formulated by Jansson et al. (under review) (Jansson 2017). This is a method that measure the social converging effect on an educational choice by characteristic similarity of school classmates. Although the method is specialized at the particular institutional context, this particular institution is one I argue is of special importance because: 1) Every member of
society attends; 2) Which its very purpose is to shape its attendees for their future life; 3) Is related to residual segregation; 4) And is under regulatory control. In other words, a central institution with great significance to the broader society. This thesis extends on the methodology of Jansson et al. (under review) (Jansson 2017).

1.1 Aim of this thesis

The aim of this thesis is to advance the methodology of homophily research. I provide a socio-metric instrument that can be used the novel methodology developed by Jansson et al. (under review) (Jansson 2017) to infer homophily. I use lower secondary school graduates’ choice of upper secondary education as a basis for this. This choice has two components; track and school. I use findings from previous research to optimize the social considerations of the graduates in this choice, which I operationalize into one socio-metric instrument. The ability use the two components is something Jansson et al. (under review) (Jansson 2017) where not able to with their socio-metric instrument, upper secondary school students’ choice of optional subjects. And as I argue for in this thesis, the most socially dependent choice is school. The socio-metric instrument I provide thesis is especially useful for researchers interested in segregation in lower and upper secondary education. It can be used to test what background dimensions are important for which classmates’ the lower secondary graduates gravitate towards, in their choice of upper secondary school. This have direct implication on the demographics of upper secondary schools and can have implication on the informal segregation in the destination upper secondary school. It can also be used to approximate self-selection bias for researchers interested in estimating peer effects in lower secondary school. Coupling the socio-metric with the methodology of Jansson et al. (under review) (Jansson 2017) shows promising ability to infer homophily from registry data.

The thesis also provides some notions on the application of the Pearson’s correlation coefficient in the methodology developed by Jansson et al. (under review) (Jansson 2017). This involve that the coefficients maximum positive value is constrained by the number possible choices is smaller than the number of independent categories. And that, coefficient’s the negative maximum is constrained by the sizes of the independent categories. These constrainers are explained in section 3.3.
The thesis also conduct empirical analyses of gender and ethnic/immigrant status homophily. The analyses will show homophily on the dimensions gender and immigrant status. Further I will compare homophily between natives and first-generation immigrants than between second-generation and natives. I will show that homophily is more persistent from first to second-generation versus natives among female graduates than among male graduates. And I will analyse different immigrant countries and show that aggregated trends of less homophily between natives and first-generation Somali immigrants than between natives and second-generation Somali immigrants.

1.2 The structure of the thesis

This main contribution of this thesis is the socio-metric instrument that is presented in chapter 4. The instrument is developed on the basis of the methodology developed by Jansson et al. (under review) (Jansson 2017). Hence, I will outline this methodology first, and the socio-metric instrument second. Because the methodology is a very recent development, and perhaps involve some elements unfamiliar to the reader I will dedicate a large proportion of the thesis to explain the procedures involved.

In chapter 2 focus on what previous research on homophily. This is done to put the methodology and analyses in context with other research on homophily. First, I go through some key concepts, address some explanations on how homophily emerges, briefly outline schools as a social arena, and empirical findings of gender and ethnic homophily. Then I put this methodology of this thesis in context with previous developments in the scientific inquiry of homophily. In chapter 3 I outline the methodology used in this thesis as it was developed by Jansson et al. (under review) (Jansson 2017) in detail. I provide an extensive explanation of the rationale behind the methodology and the procedures involved. I also contribute with some observations on the conditions for implementing the methodology in section 3.3, and a way of representing the results in tables (section 3.7). In chapter 4 outline the main contribution of this thesis, a socio-metric instrument based on lower secondary school graduates’ choices of upper secondary education in Norway, that can be utilized with the methodology of Jansson et al. (under review) (Jansson 2017). The choice of upper secondary education is a two-component choice that by using survey information from students, I combine into one socio-metric instrument aimed at best capturing the social world of the students. In chapter 5 will apply the socio-metric instrument (section 4.3) to the background
dimensions: Gender, immigration status and combinations of these and country of origin. The purpose of this chapter is to demonstrate how the metric can provide insight on homophily. I have no theory’s I want to test per se, but will formulate some expectations along the way, based on what the previous research outlined in section 2.2, have found.
2 Previous research on homophily

This chapter focuses on what previous research on homophily. This is done to put the methodology and analyses in context with other research on homophily. First, I go through some key concepts, address some explanations on how homophily emerges, briefly outline schools as a social arena, and empirical findings of gender and ethnic homophily. Then I put this methodology of this thesis in context with previous developments in the scientific inquiry of homophily.

2.1 Homophily – definition

The idiom “birds of a feather flock together” is attributed to Robert Burton by Lazarsfeld & Merton who studied friendship process in Hilltown and Craftown, which seems to be the classical citation of the term homophily (McPherson et al. 2001, 417). Translated from ancient Greek meaning the love (philía) for same (homo) (Scott et al. 1968), the concept can be understood as a preference for others is like one self. Thus, the concept is opposed to heterophily, where one prefers someone who is different than one self. In the SAGE Encyclopedia of Social Networks, Marc-David L. Seidel (2011) describes it as “… a principle of social organizing defined as people sharing similarities tending to have more social interaction”. By this definition it is not necessarily rooted in a preference but may be a consequence of opportunity. In the absence of an actor preference\(^1\) for homophily, homogeneity\(^2\) alone will make similar actors more likely to interact. When we observe more homophily than expected from chance, given the opportunities, we can assume some mechanic of preference homophily to be involved. The distinction between what is expected by chance, and what exceeds this, is in the sociological literature referred to as the distinction between baseline- and inbreeding homophily (Seidel 2011, McPherson et al. 2001, 419):

Baseline homophily is the expected amount of homophily due to existing uneven distribution in the demography. That is, the homophily attributed to the homogeneity of the individuals’ selection pools.

---

\(^1\) Could also be organizational biases originating from external forces.
\(^2\) (as adjective; homogenous) is here understood as the uniformity of an actual composition. For instance, a population is completely gender homogenous if every member is male.
Inbreeding homophily is the homophily occurring beyond the baseline. This can be caused by individuals’ preference for self-similarity but is not always limited to that. For instance, segregation of subpopulations increase the homogeneity of individuals selection pools and is sometimes included in definitions of inbreeding homophily (McPherson et al. 2001, footnote 5).3

We can imagine a spectrum of organizational tendencies in between two poles denoted by the antonyms hetero- (another, with a sense of difference) (Scott et al. 1968) –phily and homophily. Any such tendency can be considered across a social dimension. The dimension denotes where the traits are placed relative to each other. In this thesis I use similarity in background characteristics as dimensions. Lazarsfeld & Merton splits homophily between two main categories of dimensions; status homophily and value homophily (McPherson et al. 2001, 419).

Status homophily is a tendency for similarity regarding some status, informal or formal, such as gender, age, ethnicity, religion, education, or behavior patterns. These dimensions are somewhat directly observable to the other actors.

Value homophily is a tendency for similar values, attitudes or beliefs. These dimensions are not directly observable by the actors.

Homophily can occur on any given social arena, informal as well as formal (McPherson et al. 2001, 416). An arena is here understood as a defined social setting, where a set of actors can influence each other. For instance, work places, schools or nightclubs.

One can also consider multiple types of relations in what is known as a multiplex (Carolan 2014, chapter 3). For instance, a hypothesis could be that friendship relations in combination with trade relations better predicts financial support relations, in comparison to friendship relations alone.

To summarize with an example; we observe that there are more work collaborations between people of the same ethnic background than between people of different ethnic backgrounds. The arena is work place, and the dimension is ethnicities (a status dimension). If one ethnic

3 One could argue that this just using wrong resolution to calculate the baseline, and it is the local environment which is the relevant selection pool for the individuals. But this begs the question of what segregated the local environments in the first place. It could be that the preference homophily is maximized and the segregated structures are stable results of the preference.
group is overrepresented, we can expect a level of baseline homophily. If the observed ratio of intra ethnic group collaboration to inter-ethnic group collaboration exceeds the baseline, we call this inbreed homophily. The inbreed homophily is usually assumed to be the consequence of preference. Why and how such preferences for interacting with self-similar people emerges is the topic of the next paragraph.

### 2.2 Why and how homophily emerges

There are different explanations of how the preference for interacting with self-similar people emerges. Different emergence processes can result in different social structures, something that will be relevant in empirical analysis (section 5.1.1). Hence, I will briefly discuss the two main explanations of homophily emergence.

McPherson et al. (2001, 436) argues that if people have similar knowledge then; A) Communications may flow easier, and B) They may share culturally based preferences. Both factors are beneficial for coordinating activities. Further, demographic similarity could yield shared knowledge, by an increased chance of similar experiences. This can produce homophily in two ways. 1) People obtain an attraction towards others like them self. 2) Relations between similar people last longer.

1) People may perceive relations with dissimilar others as potentially more problematic than with more similar others, thus gravitate towards the similar. Experiments in social psychology find that perceived similarity is an important attractor between people formulated the “law of attraction”, stating a linear relationship between attraction to a person and the proportion of attitudes shared with the person (McPherson et al. 2001, 428, 435). Seidel (2011) mentions social identity and self-categorization as possible mechanisms. Here, individuals seek to improve social identity. This leads them to categorize people based on similarities and maximize their out-group difference. Thus, people with similar demographic traits interact more.

2) On the other side, homophily may emerge through the dissolution of relations. It could be that people don’t choose similar friends per se, but that homogenous relationships are more stable. Longitudinal research has found evidence for this mechanism.
Heterogenous ties are particularly weak in intransitive friendships patterns\(^4\), and during times of crisis (McPherson et al. 2001, 435-7).

The attraction explanation (1) can only account for status homophily (section 2.1), because the similarities need to be communicated in order to have an attractive effect. The dissolution explanation (2) on the other hand may be closer related to that of value homophily, because different values, attitudes and beliefs can complicate interactions. However, status differences can also have a dissolution effect. For instance, intransitive friendships resolutions can work on status dimensions as well as value dimensions. Furthermore, I will emphasize that the two explanations are not necessarily in competition with each other. In fact, its found that attraction among fresh students initially followed proximity, while attitude similarity became more salient later (McPherson et al. 2001, 436).

Homophily can emerge through two main parallel processes, attraction and dissolution. In the next section will concentrate on the empirical findings regarding homophily. Where we find it, and across what dimensions.

### 2.1 Schools as a social arena

This thesis is centered on the methodology developed by Jansson et al. (under review) (Jansson 2017), which utilizes the structures of schools to make inference about homophily among classmates. Hence, I will briefly outline what previous research on homophily tell us about schools as social arenas.

Schools and universities are potent breeding grounds for relationships. Education is strongly linked to occupation, linking it to other meeting places like workplaces. Outside the family are most ties formed through workplace, schools, and voluntary organizations (Louch 2000, 53). Children’s pool of potential relations is manly from their school, and particularly their classmates. Among third graders’ friends, 88% are found within their own class (school cohort). However, age homophily tend to decrease in later grades (Shrum et al. 1988). And without an organizational support the friendships become unstable (Neckerman 1996). The friendships formed in childhood are found to make up 20% of adult (Detroit men) non-

---

\(^4\) Consider actors A, B and C. If A are friends with B and C, but C and B are not friends. A and B are similar where, but C is different. This is a structural imbalance centered at A. How this is resolved can disclose preferences. If friendship AC is dropped over AB, it is a indication of homophily.
kindship friends (Fischer 1977), making school an important breeding ground for close relations. The empirical literature has shown that schools are an important breeding ground for friendships. While it might decrease with age, a substantial amount of children’s friendships occurs within their school cohort. Although there might be exceptions, I argue that one of the general differences between students on the same school of the same cohort, versus students at the same school of different cohort is that the former includes tighter personal relations. By personal relations I mean well familiar with each other, regardless of whether the relation is characterized as positive, negative, or neutral. Thus, friendship homophily should assert itself significantly more within the cohorts than between the cohorts.

2.2 Empirical findings on gender and ethnic homophily

In this section I outline previous research on the dimensions that is used in the empirical analyses in chapter 5. The focus of the thesis is the methodology and not the particular dimensions I analyze. Thus, the content of this section is intended to formulate some expectations and contextualize the analyses. McPherson et al. (2001)’s “BIRDS OF A FEATHER: Homophily in Social Network” is one of the most extensive and recent summaries of empirical literature on homophily. This section relies heavily on their work. As becomes clear in their review of empirical findings, homophily is a consistent phenomenon in social networks. It is found across a variety of dimensions, and on many social arenas. Recent research has even shown that friendships can be predicted from similarities in natural responses (Parkinson et al. 2018).

But some of the most robust social background dimensions we find friendship homophily, is gender and ethnicity. These dimensions are formal status dimensions, which we can link to school records and utilize for statistical analysis. Hence, the social dimensions I have chosen to use in my analyses is gender and ethnic background.
2.2.1 Gender

McPherson et al. (2001, 422) points out the contrast that gender\(^5\) poses to race and ethnicity, regarding homophily. While race and ethnicity are highly segregated across many dimensions such as residual areas and socio-economic background, the genders are mostly evenly distributed. For individual’s then, same gender to opposite gender contact ratio are more even than same race or ethnicity to opposite race or ethnicity. Thus, a low baseline homophily, and consequently most gender and gender homophily is due to inbreed homophily.

At the age of entering school, gender homophily is observed in children’s play patterns. Also, boys tend to play in larger groups than girls provide other interesting differences between the sexes friendship structures. For instance, do they tend to solve transitive imbalance differently. That is if an individual has two friends who are not currently friends with each other. Among boys this usually resolves in the friends becoming friends with each other, while girls tend to drop one of their friends. Regarding cross-gender friendships, are children more likely to solve transitive imbalance by both dropping a cross-gender friendship or even a same-gender friendship, than to add a cross-gender friendship. This is most prominent among the youngest children, and less in adolescents. Adults show relatively weak homophily regarding gender compared to race. Also, after controlling for kindship, which otherwise masks considerable homophily in non-kin relations. That is, a lot of cross-gender friendship among adults is also kindship. However, less intimate and contend-bound relation tend to be more homogenous. For example, found strong gender homophily (especially among males) in political discussion networks. The gender homophily is lower in highly educated young adults McPherson et al. (2001, 422-4).

2.2.2 Ethnicity

McPherson et al. (2001, 420) describes race and ethnicity as “clearly the biggest divide in social networks in the United States today, and they play a major part in structuring the networks in other ethnically diverse societies as well”. Homophily on this dimension is found to be strong in relationships ranging from the most intimate to the most distant. While race and ethnicity are social dimensions where the baseline homophily is strong, substantial

\(^5\) The contributions in the literature uses sometimes the term sex and other times gender. In this thesis I do not differentiate these concepts and use gender as the term covering both. This means that in the referenced literature you might find the word sex used where I use gender.
inbreed homophily also occur. Ethnic homophily in schoolmate friendship is found to be increasing by grades, until middle school, where the development flattens out. From here throughout high school only 10% of the baseline cross-race relations is observed. There is findings that in classrooms, belonging to a numerically small ethnic group increase the likelihood of forming heterophile friendships. This tendency is coherent with finding that ethnic group size and homophily has a rank order correlation of -.821. The explanation for this is that small minorities have fewer homophile alternatives. Finding friends who satisfies other preferences within the group decreases, increasing the potential to look outside the group (McPherson et al. 2001, 421-2).

There is also found stronger racial homophily among girls than boys. This is explained by boys’ tendency to play in larger, less intimate groups. An explanation that fits neatly with the group size and effect size correlation. A preference for more friends demands a higher number of potential intragroup relations (McPherson et al. 2001, 420).

In the job search networks of more recently arrived groups in Toronto it is found stronger homophily than later arrived. This was most prominent in first-generation, low educated immigrants, suggesting inbreeding homophily to be driven by overlapping segregation with other domains (residential, occupation etc.) and hidden value homophily (McPherson et al. 2001, 422).

### 2.3 Analytical strategies

The main focus of this thesis is a novel approach to measure homophily. Hence, I will in this section put this new methodology in context with previous developments in the scientific inquiry of homophily. The development in methodology is closely linked to emergences of new data sources. New sources enabled new ways of analyzing the social world. The methodology used in this thesis enables a new way of utilizing an existing source, the registry data, to analyze the social world.

A historical review shows how the analytical strategies towards homophily has progressed through the years. McPherson et al. (2001, 417-18) summarize the overall development of research on homophily in the last century. They draw the main lines as follows: The early work was ethnographical observation studies, focused on small social groups. This provided extensive understanding within specific social arenas and revealed substantial homophily on
psychological and demographic traits. From these findings, research on homophily developed into two main branches. One focused on what happens informally when an arena has formally been desegregated, such as desegregating efforts in schools. The other focused on the influence of peer groups on individuals’ behavior and attitudes and viewed cross-sectional associations as evidence (McPherson et al. 2001). The problem with this is what Sacerdote (2011) and Manski (1993) claims the fundamental challenge for peer effect literature: identification. A challenge which has three problems; 1. The reflection problem. The effect of peers’ behavior is subject to endogeneity bias, because the students effects its peers just as they are effect by their peers. Thus, creating a problem of recursion. Depending on what the purpose of the investigation is, this might not be a problem. From a policy making point of view, simply knowing the effect of combining students with different levels of ability might be sufficient. 2. The Separate identification problem of the effect of peers’ behavior and the effect of peers’ average background characteristics. It is difficult to separate the endogenous effects of the peers from the exogenous effects emanating through the peers, because peer background itself affects peer outcome. Even exogenous variation in peer background does not imply that both coefficients are separately identified. Note that endogenous effects have the potential for being multiplied through the reflection problem. 3. The selection problem. In social settings, similar people tend to attract each other. This causes positive selection where similar people tend to join or be assigned the same group. When not properly dealt with when identifying peer effects, this causes an upwards bias in magnitude of effect of peers’ behavior and effect of peers’ average background. The selection problem is basically a problem of accounting for homophily.

Modern peer effect literature has resolved some of these identification problems (Sacerdote 2011), usually through some unrelated student redistribution-event (quasi experiments). Homophily however, continues as an interesting research topic in itself. Inclusions of networks items in technological advanced sample surveys such as the 1985 GSS enhanced the scale of evidence (McPherson et al. 2001, 418). Closer to the turn of the century (and the most recent in perspective of McPherson et al. (2001)) research focused on the organizational context, to analyze the effects networks has on individuals’ organizational trajectories. Availability of longitudinal data have allowed for sorting the effects of selection, socialization, and attrition (McPherson et a. 2001). After the turn of the millennium technological developments has yet again enabled new forms of homophily research. Network data has been obtained through: academic citations in journals (Hâncean & Perc 2016); trough
telecommunication meta-data (Ferrara et al. 2014); through online social networks (Binder et al. 2017); and synthetic data generation (Nettleton 2016).

2.3.1 Data sources

Although the analytical strategies to understand homophily has evolved through the years, the source has naturally relied on some kind of network data. And as with any analysis, the data source constrains the possibilities of inference. Thus, I will discuss the benefits, limitations and costs of different data sources for homophily, and position of the data sources used in this thesis.

Ethnographic observation studies are studies where the researcher is present, either interacting with or merely observing the social dynamics at play. This methodology provides deeper understandings of how homophily emerges in actual social situations. The causes for, the consequences of, and the dimensions of homophily can be illuminated by ethnographic studies (McPherson et al. 2001, 417-18). But while the method brings forth invaluable knowledge, it provides very little information on the extent (generalization) of the phenomena. It is also a time costly approach, as it requires the presence of the researcher. The ethnographic approach is, however insightful, methodologically distant to this thesis.

Survey network data (McPherson et al. 2001, 418, Carolan 2014) contains information on the respondents’ social or interpersonal context. This provides essentially the opposite to that of ethnographic studies; information on the extent of the phenomena, and other statistical information. Furthermore, representative samples can sometimes be made relatively inexpensive. On the other hand, being the ethnographic studies' counterpart, survey data lack the depth. Surveying interpersonal relations implies the problem of specifying relational content, known as the socio-metric or network instrument. A socio-metric instrument is tool provide information of respondents’ relationship with others. Such specification of the socio-metric instrument could for example be with whom do you (the respondent) discusses important matters? (Carolan 2014, chapter 4). Longitudinal versions of the survey data can enable causal inference. However, linking socio-metric instruments to actions. For instance, there is not clear what effect with whom you discuss important matters has on your behaviour.

Archive network data are networks reconstructed from some record, which is the type of data used in this thesis. The records could be anything from diaries to school records or Facebook
data or citation lists. The idea is to reconstruct a network from expressed behaviour in the archive. (Carolian 2016, 75). The advantages of this data source lie in bypassing the respondent. Depending on the storage and accessibility of the records, analysing them could be way cheaper than conducting a survey. Also, depending on the reliability of the records, these data can provide more accurate and relevant information of the social world. By analysing actions actually taken by the actors in the network, rather than asking them who their friends are, gains two advantages. First, respondent biases are non-existent\textsuperscript{6}. Second, actions propagate back into the network, and one do not need to worry about whether the socio-metric instrument has any real-world value. However, the advantages with archive network data relies heavily on the quality of the archive, access to it, and the use of it.

This thesis will use the Norwegian registry data of education choices, as source of archive network data. This archive has the benefit of recording every enrolled student’s actual educational choice. Hence, we know that the choice has a real-world value. They are however not records of actual social networks, per se. There is no information on who is someone’s friends, with whom they discuss important matters, or similar. What there are, is some choices with great social importance for the students. And with an innovative new method develop by Jansson et al. (under review) (Jansson 2017), we aim at distilling the social component of that choice. The general idea behind the method is to use similarities of choices as a socio-metric instrument that captures many of off the same social aspects as questions about who are their friends or discussing partner in important matters. The choice is the important matter, and who they discuss it with will manifest itself in the choice. This has both the benefit and the caveat of not being a self-reported socio-metric instrument. The caveat is that we do not know who students’ subjective important others are. But the benefit is that we know how background dimensions influence an objective choice. Hence, we do not need to justify the socio-metric instruments objective value on behaviour. However, we need to ensure that the choice is sufficiently social to capture as much of the social world of the students as possible. In this thesis I suggests a socio-metric instrument optimized to capture the social aspect of the choice of upper secondary education. This is a two-component choice that I, by using survey information from students, I combine into one socio-metric instrument aimed at best capturing the social world of the students. In the next chapters I will outline the rationale behind Jansson et al. (under review) (Jansson 2017) methodology, the procedure,\textsuperscript{6} Again, this relies on the reliability of the records.
and my further development of their rationale and methodology to a new socio-metric instrument, based on lower secondary graduates in Oslo and their choice of upper secondary education.
3 Methodology

In this chapter I outline the methodology used in this thesis as it was developed by Jansson et al. (under review) (Jansson 2017) in detail. I provide an extensive explanation of the rationale behind the methodology and the procedures involved. I also contribute with some observations on the conditions for implementing the methodology in section 3.3, and a way of representing the results in tables (section 3.7).

The methodology used in this thesis is a novel approach developed by Jansson et al. (under review) (Jansson 2017), which they label a meta-analytical QAP (quadratic assignment procedure) approach. It involves a three-stage analysis for every graduation class and a meta-analysis of all the classes. It is developed for a similar purpose as this thesis, namely to measure peer homophily effect in upper secondary students on their choice of specialization tracks. Thus, they used a slightly different choice as I use, but with the same kind of social components: The students’ motivation is related to a wish to spend time together, and peer effects on their curricular interests. However, the choice they used was not separable into components, in contrast to the choice measured in this thesis, which I separated into tracks and schools in section 4.3.

The methodology developed by Jansson et al. (under review) (Jansson 2017) is based on comparing similarities of educational choices among graduates at the same school, within the same cohort (intra-class), with the similarities of educational choices among graduates at the same school in different cohorts (inter-class). In short, the method takes difference between the product-moment correlations coefficients intra-class and inter-class, where the correlation is between similarity of choice and similarity of background characteristic. The general differences between intra-class and inter-class situations is that the former includes personal relations, where the latter does not, as outlined in section 2.1.

The stages of the procedure are as outlined in Jansson et al. (under review) (Jansson 2017): 1) Measuring the correlation between sharing an explanatory attribute and sharing a dependent choice within every school class. The statistical significance of the coefficient is calculated by

---

7 In this thesis the class is year specific cohort at individual schools. One can argue that this does not constitute a class, as these levels are often subdivided into different classrooms. That is the 2010 grade 10 cohort at a school can be divided into 10A and 10B. I however do not have information on this detail and will use the school-cohort as this unit.
the quadratic assignment procedure; 2) the next step is to synthesize the coefficients of all classes, into a mean effect size. This is done with a random effects meta-analytical approach; 3). Because we are interested in the effect of classmates, we should control for possible confounding variables. This is the general tendencies outside of ones’ class, for example immigrants from the same country of origin tendency of making a particular choice. To deal with this problem we also make a correlation coefficient in the same matter as in stages 1 and 2, for the rest of the school, also with a QAP measured statistical significance. Thus, we have an intra- and an inter-class correlation coefficient. Removing the inter-class coefficient from the intra-class coefficients leaves the Jansson homophily coefficient, as will be outlined in section 3.1. The stages will be explained in detail in the following sections. I will outline how we can convert regular registry data into the network data that we use. Then I will outline how we calculate the correlation. And following up with an example to show how the intra- and inter-classes coefficients are calculated. Then I will address the meta-analytical stage. At last I will outline how I represent empirical results from this methodology.

3.1 Causal model

In this section I present the rationale behind the methodology developed by Jansson et al. (under review) (Jansson 2017). That is, the rationale for making inference of informal relational homophily based on intra- to inter- class differences in similarity of graduates’ choices of further education in applications. It builds on the empirical backed assumption outlined in section 2.1, that personal relations are more likely to occur intra-cohort than inter-cohort. This results in two different causal models, where the main causal direction is straight forward. The aim is to infer how known background characteristics interfere with the personal relations between classmates made through lower secondary school, using the effect of social relations on their educational choices. The statistical relationship of interest is between a social dimension of different background characteristics and personal relations with classmates, visualized in Figure 1.

![Figure 1: Homophily](image-url)
The dependent variable in the model is classmate similarity of an educational choice with some social importance. The explanatory variable is classmate similarity of some background variable. In other words, the social dimension we are interested in measuring homophily on.

![Diagram: Intra class similarities of classmates](image)

Figure 2: Intra class similarities of classmates

The effects of homophily or heterophily should make characteristic-similar classmates' choices more or less similar, respectively (see Figure 2). Certain backgrounds can be associated with certain curricular interests and aspirations due to for example cultural norms. Social background could also affect the choice more directly, through family expectations or information flows in the broader network. However, the causal relation we are interested in here is where background characteristics influences the social relations with classmates. Homophily (or heterophily) on a background characteristic social dimension will propagate further down the causal chain. The interests of, and relationships with classmates, probably affects a student’s interests. Furthermore, graduates might coordinate their educational choice to spend future time together. In the model, there is also an arrow from curricular interests to personal relations, indicating that the curricular interests and aspirations plays a role in the relationships between classmates as well. This detail illustrates that the background can influence personal relations through curricular interests. What we are interested in is the part of the effect of background characteristics on educational choice which flows through personal relations (see Figure 3).

18
Because we do not have information on the social relations within the class, this is not separable in a single class. What we do instead is to calculate the relationship when we expect absents from this mechanism. We use student relations from the same school, but not the same graduation class. Here we expect a drastic decrease of social ties, but background effects unrelated to ties to be retained. Thus, the only causal chains making characteristic-similar graduates choose similarly, are the one not involving personal relations (see Figure 4). While there could be personal relations extending beyond the class, we expect them to be far less frequent. If any, these relations will result in an overly conservative estimate in the final social effect.

Figure 3: Interpersonal classmate homophily effect

Figure 4: Inter-class similarities of classmates

Thus, the correlation of similarity of background characteristic on the similarity of educational choice between classes will serve as a baseline, where personal relations within
the class would cause the correlation to shift. Arithmetically we can express the derived effect as Equation 1.

**Equation 1**

\[ t_{ij} = H^r X_{ij} + \varepsilon_{ij} \]

Where \( t_{ij} \) represents the actual social tie between a pair graduates: \( i \) and \( j \). \( X_{ij} \) is the similarity of background characteristic, and \( H^r \) is the background characteristic specific homophily effect (the weight) on classmate relations. \( \varepsilon_{ij} \) is the random errors influencing the tie, not related to the background characteristic. If graduate \( i \)'s choice of education, \( Y_i \) is somewhat determent of the ties to its classmates. \( i \)'s background characteristic \( X_i \), and idiosyncratic variation \( \varepsilon_i \), we can express this as Equation 2.

**Equation 2**

\[ Y_i = H^e t_{ij} + \beta X_i + \varepsilon_i \]

Where \( H^e \) is the influence effect of classmate relations (the weight), and \( \beta \) is the preferences for education related to the background characteristic \( X_i \). The similarity of choice is a product of the two choices, and can be expressed as Equation 3.

**Equation 3**

\[ Y_{ij} = H X_{ij} + \beta X_i X_j + \varepsilon_{ij} + \varepsilon_i + \varepsilon_j \mid H = H^r H^e \]

Where in addition to the shared attribute, unobserved factors constituting relevant social ties and idiosyncratic variation determine whether a pair of actors makes the same choice. It could for instance be that \( X_i \neq X_j \), but \( \varepsilon_i = \varepsilon_j \), thus the individuals make a similar choice unaffected of each other. Since these instances are by definition unrelated to \( X \), we can ignore the error terms. We account for \( \beta \) by taking a *between* (inter) classes correlation \( B[\Gamma_{XY}] \), by running separate regression for the school level. Here every graduate is paired with every other graduate from its school, who is not in their particular graduation class. Here we expect \( H=0 \), and the error terms to be the same. Thus, we express this as Equation 4.

**Equation 4**

\[ Y_{ij} = \beta X_i X_j \]  

*Where \( i \neq j \), and \( i \& j \) graduate from the same school but not the same class*
This correlation coefficient is then subtracted from the one made on only the students within the particular (intra) class $W[\Gamma_{XY}]$, expressed as Equation 5

\begin{equation}
Y_{ij} = HX_{ij} + \beta X_i X_j
\end{equation}

*Where i ≠ j, and i & j graduate from the same class*

Thus, we derive the classmate homophily effect of X on Y, expressed as Equation 6

\begin{equation}
H[\Gamma_{XY}] = W[\Gamma_{XY}] - B[\Gamma_{XY}]
\end{equation}

From now on I coin this difference as a quantity that we observe and refer to it as the Jansson homophily coefficient, or JHC for short. A substantial interpretation of the JHC is the observed attractional effect at poles of a social dimension that is exerted on a choice.

3.2 Data transformation

In this section I will explain how the registry data is transformed into the relational matrices that we use for the further analysis. For this thesis I will utilize registry data as an archive network data source with the aim of measuring homophily. However, registry data do not contain information on informal relationships, and are not structured as social networks. Initially the data is in the form of observations of every individual at their year of graduation from lower secondary school. These observations contain variables on: Graduation year, lower secondary school, and primary choice of upper secondary education (school as well as specialization track), and a set of background characteristics.

We then transform this data into adjacency matrices. The graduates are grouped on their lower secondary school (supergroup) and class (subgroup). The subgroup ties are removed from the supergroup calculation, and thus the supergroup contains relations inter-classes while the subgroup contains relations intra-classes. For every variable, a matrix is created with information on how similar the individuals are. For example, gender (X). Every graduate $i$ is paired with every other graduate $j$ in his/ her group and if they have the same gender ($X_i=X_j$) then $X_{ij}=1$, otherwise $X_{ij}=0$. Similar adjacencies are calculated for the other independent variables. For dependent variable, the choice of upper secondary education (Y). Every graduate $i$ is paired with every other graduate $j$ in his/ her group and if they have applied for
the same upper secondary school \((Y_i = Y_j)\) then \(Y_{ij} = 1\). If they have applied to the same track but at different schools I assign \(Y_{ij} = -1\), otherwise \(Y_{ij} = 0\). They are undirected networks, with everyone sharing a property forming their own clique subnetwork\(^8\). This is happening because of a phenomenon called complete transitivity. To illustrate this, consider the following triplet \(a, b\) and \(c\). If \(a\) and \(b\) make the same choice, and \(a\) and \(c\) also make the same choice, then \(b\) and \(c\) necessarily make the same choice. This have implications related to autocorrelation and will be discussed in section 3.4.

### 3.3 Network correlation coefficient

We calculate the network correlation \(\Gamma\) between a pair of matrices \(X_{ij}\) and \(Y_{ij}\), with the standard Pearson correlation coefficient, where the entries of the matrices are the observations as Equation 7.

**Equation 7**

\[
\Gamma_{XY} = \frac{\text{cov}(X_{ij}Y_{ij})}{\sigma_{X_{ij}}\sigma_{Y_{ij}}}
\]

The value of \(\Gamma\) range from -1 to 1. A positive value indicates that sharing \(X\) is associated with sharing \(Y\). Similarly, a negative value indicates that sharing \(X\) is associated with not sharing \(Y\). A value of 0 indicates that there is no linear association between sharing \(X\) and sharing \(Y\). The magnitude of the coefficient indicates the strength of the relationship, how much of the variance of \(Y\) that is explained by \(X\) (Carolan 2014, Pearson 1895, Fisher 1915).

When interpreting coefficients in this case there are some extra notions I will make that are related to how the matrices are constructed. Transitivity makes potential correlation range dependent on the structure of the class relative to the number of choices. A useful approach to understand this can be to explore the extremes.

- A maximum correlation \((\Gamma = 1)\) is obtained when all students who share the value on the explanatory variable also share a value on the dependent variable, and with no other. For example, if all the boys applied for the same school and all the girls applied for one other school. If for example some girls where to apply to a third school, it will decrease the coefficient. If they applied to the same school as the boys, the coefficient

---

\(^8\) All actors are connected to every other actor
will decrease additionally. This implies that if the possible choices are fewer than the explanatory categories, a coefficient of 1 is impossible. For example, if there were only two schools to choose from, and students from three different countries of origin. The configuration that maximizes the correlation is when the two smallest groups share one of the choices, and the large group exclusively choose the other. This configuration will have the property of \( \Gamma < 1 \), but the students are still maximally segregated with respect to country of origin. The data used in this analysis is from lower secondary schools in Oslo only, where there is a wide range of upper secondary schools and tracks to choose from.

- In the other end of the spectrum, the correlation range down to \( \Gamma = -1 \) as the theoretical minimum. In this case, no graduates make the same choice as any peer with the same explanatory value as itself, but the same as every peer with a different value. Because of transitivity again, this is impossible whenever there are more than one graduate that belong to an explanatory category. This becomes evident by the following paradox: Graduate A is male and thus shares choice with female graduate C, but not with male graduate B. Graduate B also shares choice with C. But this choice must be the same as A’s, with which B does not share a choice. This means the lower bound of the coefficient is theoretically constrained by the sizes of the explanatory categories as well as the number of dependent choices.

- Zero correlation is what to expect if the variable is socially irrelevant among the graduates. \( \Gamma = 0 \) means that the within explanatory category number of same choices are proportional to the between category number of same choices. This is our null hypothesis, and the p-value is the chance of observing a correlation of at least this size if there are no real effect. This is not the same as if the choice was made by chance, and thus we need to retain the network structure. How we test the null hypothesis is the subject of the section 3.4.

In this section I have presented the correlation measurement I will be using in the thesis and outlined observations on some potential problems with using this measurement. The validity of these observations were confirmed by Fredrik Jansson Senior Lecturer in Applied Mathematics, Mälardalen University and Postdoc in Cultural Evolution, Stockholm University (e-mail, 11. June 2018). The problems are identified with some particular

---

9 Except when there is only one category. Then it is only constrained by the number of choices.
circumstances, mostly related to the extremes of the coefficient spectrum. What effect these problems will have on more moderate parts of the spectrum is not clear. But if any, I suspect that they have a curbing effect and can cause us to underestimate homophily. The correlation coefficient is the same that is used by Jansson et al. (under review) and Jansson (2017). And the problems should not affect our estimation of standard errors. Thus, I precede with this correlation measurement. To counter some of the problems, I have restricted my data to graduates from lower secondary schools in Oslo only.

3.4 Standard errors (The quadratic assignment procedure)

Because there can be multiple relations going on between actors, social network observations are inherently not independent. And dyadic data from the same person is not independent of each other. Thus, relational data violates a basic assumption for using standard ordinary least square regression models. Ignoring this would lead to dramatic underestimated standard errors (Carolan 2014, 69). Jansson et al. (under review) (Jansson 2017) uses a solution to this has been put forth by Krackhardt (1987), called the Quadratic Assignment Procedure (QAP).

Here we assume \( H_0 \): that sharing a background characteristic of \( X \) is uncorrelated with sharing a choice of \( Y \). Then we test the probability of attaining a value of \( \Gamma \) or more extreme by chance, as expressed in Equation 8. This could be calculated by comparing the actual correlation \( \Gamma_{XY} \), to the distribution of all \( n! \) (Where \( n \) is the number of graduates) possible permutations. That is the other possible reordering’s of \( i \)’s and \( j \)’s in \( Y_{ij} \) (the rows and columns get the same reordering).

\[ \Gamma_{XY}(p) = \sum_{i,j} X_{p(i)p(j)} Y_{ij} \]

This means the different categories with their respective populations, as well as the number of ties are the same in every permutation. The proportion of permutations who exceeds \( \Gamma_{XY} \) is equal to the chance of attaining the value at random. We can set a significance level \( \alpha \), and know that the chance of drawing a value \( \Gamma \geq \Gamma_{XY} \) randomly is below the significance level. However, because \( n! \) grows exponentially with \( n \) and becomes computational impractical.

\[^{10}\text{In a sense the } H_0 \text{ has an expected value at the baseline homophily (described in section 1).}\]
even with moderate $n$, the significance testing is done on a Monte Carlo sampling of the permutations. This means that every permutation is equally probable. This thesis follows Jansson et al. (under review) (Jansson 2017), where 1000 permutations are used as the size of this sample. In short, the $p$-value provides the probability of the effect being larger than zero, if there are no relationship between $X_{ij}$ and $Y_{ij}$.

3.5 An exemplary walk trough

I will now go through the procedure with a small and fictive dataset to exemplify the process. This is an example data derived from Jansson's github page (Jansson 2016) but assigned some names for readability. In an imagined school the graduates of 2015 had the following properties: Country of origin. The graduates originated from Norway, Sweden or Pakistan. Choice of school. There were two different preferred schools in the class. Elvebakken VGS and Blindern VGS. This is illustrated in Table 1.

<table>
<thead>
<tr>
<th>Student</th>
<th>Origin</th>
<th>Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Norway</td>
<td>Elvebakken VGS</td>
</tr>
<tr>
<td>2</td>
<td>Sweden</td>
<td>Blindern VGS</td>
</tr>
<tr>
<td>3</td>
<td>Norway</td>
<td>Blindern VGS</td>
</tr>
<tr>
<td>4</td>
<td>Norway</td>
<td>Blindern VGS</td>
</tr>
<tr>
<td>5</td>
<td>Pakistan</td>
<td>Elvebakken VGS</td>
</tr>
<tr>
<td>6</td>
<td>Sweden</td>
<td>Elvebakken VGS</td>
</tr>
<tr>
<td>7</td>
<td>Pakistan</td>
<td>Elvebakken VGS</td>
</tr>
<tr>
<td>8</td>
<td>Norway</td>
<td>Blindern VGS</td>
</tr>
<tr>
<td>9</td>
<td>Pakistan</td>
<td>Elvebakken VGS</td>
</tr>
<tr>
<td>10</td>
<td>Norway</td>
<td>Blindern VGS</td>
</tr>
</tbody>
</table>

We generate two different networks on the basis on the different variables. First, we connect the graduates that apply to the same school in one network, and the second network where they are connected if they share country of origin. The connections is illustrated in Figure 5: and the intra-class adjacency matrices is illustrated in Figure 6.
Figure 5: Example data: Visual representation of adjacencies

\[
\mathbf{Y}_{ij} = \begin{bmatrix}
NA & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\
0 & NA & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 1 & NA & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 1 & 1 & NA & 0 & 0 & 0 & 1 & 0 & 1 \\
1 & 0 & 0 & 0 & NA & 1 & 1 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 & NA & 1 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 & 1 & NA & 0 & 1 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 & 0 & NA & 0 & 1 \\
1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & NA & 0 \\
0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & NA
\end{bmatrix}, \quad \mathbf{X}_{ij} = \begin{bmatrix}
NA & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & NA & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & NA & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\
1 & 0 & 1 & NA & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & NA & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & NA & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & NA & 0 & 1 & 0 & 1 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & NA & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & NA
\end{bmatrix}
\]

Figure 6: Example data: Intra-class adjacency matrices

The correlation between the two networks is \( W[\Gamma_{XY}] = 0.26834 \). To test for statistical significance we use the QAP method. We randomly reorder the order of rows and columns of one the \( \mathbf{Y}_{ij} \) and calculates a similar correlation, 1000 times. We get a distribution of coefficients, where \( p \) is the proportion of them is more extreme (in this case larger) than our coefficient.
Figure 7 shows that the distribution of permuted correlation values seems to be polymodal and skewed to the left. This is because of the size of the class and variance of dependent and independent attributes, which results in a finite number of possible permutation. However, we are interested in probability of the correlation strength 0.26834, which is only exceeded or matched by 8.8% of the permutated coefficients, thus it has a p-value of 0.088. The mean permutated coefficient is -0.00311, and the red line indicates the probability distribution of coefficient strength under null hypothesis.

When we do the procedure inter-class, every graduate is paired with every other graduate from the same school, except graduates from its own graduation glass. If we consider an extra class to the previous example as a school. In the 2016 class there were also a graduate with country of origin from India and the information about the classes is illustrated in Table 2.
Table 2: Example data: Classes 2015 and 2016

<table>
<thead>
<tr>
<th>Student</th>
<th>Origin</th>
<th>Choice</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Norway</td>
<td>Elvebakken VGS</td>
<td>2015</td>
</tr>
<tr>
<td>2</td>
<td>Sweden</td>
<td>Blindern VGS</td>
<td>2015</td>
</tr>
<tr>
<td>3</td>
<td>Norway</td>
<td>Blindern VGS</td>
<td>2015</td>
</tr>
<tr>
<td>4</td>
<td>Norway</td>
<td>Blindern VGS</td>
<td>2015</td>
</tr>
<tr>
<td>5</td>
<td>Pakistan</td>
<td>Elvebakken VGS</td>
<td>2015</td>
</tr>
<tr>
<td>6</td>
<td>Sweden</td>
<td>Elvebakken VGS</td>
<td>2015</td>
</tr>
<tr>
<td>7</td>
<td>Pakistan</td>
<td>Elvebakken VGS</td>
<td>2015</td>
</tr>
<tr>
<td>8</td>
<td>Norway</td>
<td>Blindern VGS</td>
<td>2015</td>
</tr>
<tr>
<td>9</td>
<td>Pakistan</td>
<td>Elvebakken VGS</td>
<td>2015</td>
</tr>
<tr>
<td>10</td>
<td>Norway</td>
<td>Blindern VGS</td>
<td>2015</td>
</tr>
<tr>
<td>11</td>
<td>Sweden</td>
<td>Blindern VGS</td>
<td>2016</td>
</tr>
<tr>
<td>12</td>
<td>Norway</td>
<td>Blindern VGS</td>
<td>2016</td>
</tr>
<tr>
<td>13</td>
<td>Norway</td>
<td>Elvebakken VGS</td>
<td>2016</td>
</tr>
<tr>
<td>14</td>
<td>Norway</td>
<td>Blindern VGS</td>
<td>2016</td>
</tr>
<tr>
<td>15</td>
<td>Pakistan</td>
<td>Elvebakken VGS</td>
<td>2016</td>
</tr>
<tr>
<td>16</td>
<td>Pakistan</td>
<td>Blindern VGS</td>
<td>2016</td>
</tr>
<tr>
<td>17</td>
<td>India</td>
<td>Elvebakken VGS</td>
<td>2016</td>
</tr>
<tr>
<td>18</td>
<td>Pakistan</td>
<td>Blindern VGS</td>
<td>2016</td>
</tr>
</tbody>
</table>

For example, the inter-class origin adjacency matrix of this school looks is illustrated in Figure 8.
The correlation between the sharing origin and sharing choice at the school level is $B[\Gamma_{XY}] = 0$. The QAP standard errors for this relationship (Figure 9) can be summarized in the same manner as for the from the intra-class relationship (Figure 7).

The first thing to notice from Figure 9 is that the distribution is no longer polymodal or skewed, but close to a normal distribution. This is because of the size and variance if the school is larger than for the class. There are still differences in the individual permutation strengths, but we are interested in the probabilities of magnitude. The correlation 0 is exceeded or matched by 53.5% of the permuted coefficients in the positive direction, and 54.3% in the negative direction. The mean permutated coefficient is -0.00033.
To get the JHC, we subtract the inter-class coefficient from the intra-class coefficient:\[ H(\Gamma_{XY}) = 0.26834 = 0.26834 - 0. \] To evaluate the new statistical significance of the new coefficient, we do the same to their QAP distributions. This yield a \( p = 0.116 \), which means that the probability of getting a Jansson homophily coefficient larger than 0.26834 is 0.116. This \( p \)-value is not low enough to conclude that the correlation is represent an actual homophily effect.

![Density distribution of controlled simulation coefficient](image)

*Figure 10: Example data: The probability distribution of a JHC if there is no effect*

In this section I have used a fictional example to illustrate how we transform data, calculate correlations and calculate standard errors with the Jansson et al. (under review) (Jansson 2017) methodology. The example showed how intra- and inter-class coefficients are different, and how we use the difference to calculate the homophily (JHC). However, they only included a single class at a single school. How we combine multiple coefficients into a single measure is the topic of section 3.6.

### 3.6 Meta-analysis

This study will look at students in several school classes in different schools, and we have to combine the results of different classes into a single measure. Jansson et al.’s (under review) and Jansson (2017) solution to this is to consider each class a separate experiment and do a meta-analysis of the results, where the coefficients are weighted (\( w_i \)) by dividing by the inverse of their variance (\( \sigma_i^2 \)). The variance is derived from the distribution of the permutation

---

When we have more classes and schools we first do the random effects meta-analysis and take the difference between the intra-class and inter-class meta-coefficients.
coefficients from the QAP for that class. Weighting the coefficients like this makes them comparable in respect to their statistical significance. \( w_i = \frac{1}{\sigma_i^2} \)

There are two main statistical approaches to meta-analysis, the fixed-effect and the random-effects model. Borenstein et al. (2010) provides an explanation for whether we should use one or the other. The fixed-effect model should be used if the individual studied are samples of the same population. In this case all studies estimate the same singular effect, and any variation is due to sampling error. On the other hand, if the individual studies are sampled on different populations, the true effects may differ from population to population. Thus, we have a problem of effect heterogeneity, and should be using the random-effects model. This model implements the between-study variance in the weighting of the coefficients. In our case the effect is something which emerges from classroom social dynamics. That is, a product of the individuals, their relations and their history together. This makes it reasonable to expect that the importance of a social dimension varies from class to class. Thus, the meta-analysis is conducted as a random effects model.

3.6.1 The random-effects model

In the random-effects model we weight the coefficients in addition to their individual variance, by the between study variance (\( \tau^2 \)). Because the individual coefficient might be an inaccurate estimate of the true individual effect, calculating \( \tau^2 \) is not as straight forward as squaring the distance from the coefficients to the mean of the coefficients and divide on the number of studies. We need to partition the within-study estimation error away from the observed variation, and into real heterogeneity of the effect size. This is done with the DerSimonian and Laird method (1986). The concept of this method is to estimate the expected variation of the observed effects if we assume the true effect was the same for all studies and remove this from the total amount of observed variation, expressed in Equation 9.

\[ \tau^2 = \max \left\{ 0, \frac{\sum_{i=1}^{m} u_i \tau_i^2 - \left( \frac{\sum_{i=1}^{m} u_i \tau_i}{\sum_{i=1}^{m} u_i} \right)^2 (m-1)}{\sum_{i=1}^{m} u_i} \right\} \]  

\[ u_i = \frac{1}{\sigma_i^2} \]

---

\[ \text{Equation 9} \]

12 The full equation as outlined in DerSimonian and Laird (1986): \( \tau^2 = \) \[
\max \left\{ 0, \frac{\sum_{i=1}^{m} u_i \tau_i^2 - \left( \frac{\sum_{i=1}^{m} u_i \tau_i}{\sum_{i=1}^{m} u_i} \right)^2 (m-1)}{\sum_{i=1}^{m} u_i} \right\} \]

\[ u_i = \frac{1}{\sigma_i^2} \]
Equation 9

\[ \tau^2 = \frac{Q - df}{C} \]

If \( \tau^2 < 0 \), it is replaced by 0

Where \( Q \) is the variance (\( \sigma_i^2 \)) weighted sum of squared distance from the effect size estimates to the mean of the effect size estimates, and is calculated according to Equation 10

Equation 10

\[ Q = \sum_{i=1}^{m} \frac{\Gamma_i - \Gamma}{\sigma_i^2} \]

We subtract the \( df \) which is equal to number of studies (\( m \)) minus 1, because if all studies share a common effect \( Q=df \). Thus, any remaining variance is the excess variation between studies. We divide by \( C \) put the index back into the same metric as the within-study variance, as expressed in Equation 11

Equation 11

\[ C = \sum_{i=1}^{m} w_i - \frac{\sum_{i=1}^{m} w_i^2}{\sum_{i=1}^{m} w_i} \]

With the DerSimonian and Laird estimate, statistically significant heterogeneity always yield a positive \( \tau^2 \). It however tends to overestimate \( \tau^2 \), especially for meta-analysis with small \( m \) (Borenstein et al. 2010). This will increase the confidence interval, and the risk of committing a type 2 error. This thesis will use large \( m \), so this should not be large problem.

After computing \( \tau^2 \) we simply implement it into the weighting of the coefficients as expressed in Equation 12

Equation 12

\[ w'_i = \frac{1}{\sigma_i^2 + \tau^2} \]

The variance and weighted mean is computed over their correlation coefficients Fisher’s z transformed values to better approximate the sampling variance (Fisher 1915). The Fisher’s z transformation is expressed in Equation 13
Then run the random-effects model with this weighting as in Equation 14.

\[
z(\bar{\phi}) = \frac{\sum_{i=1}^{m} w'_{i} z(\Gamma_{i})}{\sum_{i=1}^{m} w'_{i}}
\]

The \(z(\bar{\phi})\) is transformed back to original scale and we get the estimated mean \(\bar{\phi}\) of the population of effects.

In our case, the variance used in the random-effects model is the variance of the distribution of QAP-coefficients. We also combine a set of the QAP-coefficients into a weighted mean, weighted by the same weights \((w'_{i})\) as the coefficients. We are left with a meta-coefficient and a meta QAP-distribution, and calculate the p-value on the proportion of meta QAP-coefficients that exceeds the observed meta-coefficient.

### 3.7 Presentation of result

Because this methodology is novel and fairly different from what many sociologists are used to work with, there exists no obvious way of presenting results derived from it. Hence, I will outline and explain how the empirical results will be presented in this thesis here. Table 3 is an example of the layout.

**Table 3: Example of result layout**

<table>
<thead>
<tr>
<th>Independent dimension</th>
<th>JHC (H[\Gamma_{XY}]) 0.2***</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intra (W[\Gamma_{XY}])</td>
<td>0.3***</td>
</tr>
<tr>
<td>Inter (B[\Gamma_{XY}])</td>
<td>0.1***</td>
</tr>
<tr>
<td>N</td>
<td>22102</td>
</tr>
<tr>
<td>Classes</td>
<td>344</td>
</tr>
<tr>
<td>Schools</td>
<td>65</td>
</tr>
<tr>
<td>T2 Intra</td>
<td>0.0008</td>
</tr>
<tr>
<td>T2 Inter</td>
<td>0.0002</td>
</tr>
</tbody>
</table>

Note: * = \(p<0.05\); ** = \(p<0.01\); *** = \(p<0.001\)
I will use a table where the explanatory variable is referenced in the line on the top of the table. In the first row within the table I have placed (in bold characters) the Jansson homophily coefficient that we are interested in estimating. The JHC is the difference between the intra-class and inter-class coefficients in the two following lines. All numbers in these three lines should be interpreted as a correlation coefficient and are accompanied by asterisks indicating their statistical significance level. The asterisks increments are given in the line at the bottom of the table. N indicates the number of graduates that is included in the model. The Classes– and Schools–lines indicate the number of classes and schools respectively that were included in the model. The numbers in these three lines (N, Classes and Schools) can vary from model to model because some matrices lack variation. For example, if every graduate in a class makes the same choice. Without variation we cannot estimate standard errors and therefore exclude this class from the meta-analysis. At the bottom of the table we have the T2-lines for the Intra- and Inter-class meta-analysis respectively. The T2 number is the part of the variation that is attributed to effect variation, and not sampling errors.
4 Choice of upper secondary education as a socio-metric instrument

In this chapter I outline the main contribution of this thesis, a socio-metric instrument based on lower secondary school graduates’ choices of upper secondary education in Norway, that can be utilized with the methodology of Jansson et al. (under review) (Jansson 2017). The choice of upper secondary education is a two-component choice that by using survey information from students, I combine into one socio-metric instrument aimed at best capturing the social world of the students.

Jansson et al. (under review) (Jansson 2017) measured homophily with an innovative new method, based on the revealed preferences in the actual choices students make. More specifically, they attempt to measure homophily in upper secondary school by looking at students’ choice of specialization tracks. This is a choice which in addition to curricular interests, involve who they will spend time with. Thus, a choice likely to reflect the social context in which they were made, at least to some extent (under review) (Jansson 2017). It is this social component of the choice which is distilled through the new method developed by Jansson et al. (under review) (Jansson 2017). Using the revealed preferences has two major advantages as opposed to social network surveys. Firstly, the registries already exist, on the whole population. This is an inexpensive option in comparison to conducting surveys. Secondly, the choice is a social action itself. This means we avoid interpreting the strength and meaningfulness of the social tie in question and potential misreporting. That is, the method focusses on a choice influenced by social ties and infer the relational information from the choices. Thus, we know by definition that the underlying ties are meaningful to the extent that they influence the choice (under review) (Jansson 2017).

As Jansson et al. (under review) (Jansson 2017) points out, their method can be expected to be relevant for other situations, such as higher educational tracks later in life. They argue that this gives an extra dimension to the choice, as the applicants must consider geographical distance to different schools as well as their curricular interests and social environment. The argument is that geographical distance to schools is important for maintaining social contact with friends. In this thesis however, I suggest using educational decisions earlier in life, namely the transition from lower secondary to higher secondary education in Norway. The choice of upper secondary school also contains the geographical dimension of school
location, but also have a much wider attendance. And as I will show, there is different considerations that lower secondary school graduates make on the choice of upper secondary school and the choice of upper secondary specialization tracks. This difference I utilize to optimize a socio-metric instrument that best reflects the social considerations of the lower secondary graduates on their choice.

4.1 Applying for upper secondary school

In Norway, adolescents are generally at the age of 15/16 when they graduate from lower secondary school, which is the last year of mandatory schooling. They then have the right to attend upper secondary school education. Although this is optional education, almost everyone attends. They make a priority list of their three top choices of upper secondary education and are legally entitled to one of their choices (Opplæringslova § 3-1). This choice contains what kind specialization track (general studies, mechanics, health and social studies, etc.) they want to attend as well at which school they want to attend.

From these three top three choices one can derive many different measures. I have no data on the second and third options and use only data about the choice with highest priority. The potential caveats of this is that including the second and third choices might give other insight. However, Lødding and Helland (2007) conducted a rapport on regional based affirmative action in the upper secondary educational choices in Oslo. In a survey roughly three fourths responded that they were accepted at their preferred school, with some regional differences. In registry data about applicants for general studies, the acceptance rate for first choice of schools where 70.9%. The students at general studies track where asked if their primary choice of school reflected their actual preference. Two thirds claimed their primal choice reflected their actual preference, and that they expected to be admitted. Less than one fifth claimed their primal choice reflected their actual preference, but they did not expect to get in. 14% claimed their primal choice did not reflect their actual preference, and they

---

13 92.2% of all 16 to 18-year old’s in 2015 (Statistics Norway, 2017)
14 For example, using all three choices as one. Containing most information, it is the combination graduates are the least likely share without communicating. This could however, be to conservative when considering that the social component is not the only factor working. One might desire to attend the same school as a friend, but the friend might not have identical curricular interest. Also, a completely similar application combination is not necessarily the optimal chance of being granted the same choice, whenever the graduates differ in their grade scores. Considering this the simpler first choice, may even be more suited. Perhaps some ranking of similarity, on how well the combinations fits each other could be optimal.
applied strategically. This also had some regional differences, and depended on grades. If taken at face value\textsuperscript{15} the students’ first choice is a good indication of the desired choice of the graduates. The minority who applied strategically at a different choice than their preference would constitute a problem if they did so to attend with friends who would expect to get in. If this kind of behavior is dominantly associated with higher grades or with strategic applications we have a source of error in our data. This is not possible to address without extended data on the applications and is a possible weakness.

From 2004 to 2008 the application system in Oslo where subjected to regional based affirmative action, where half of the available slots at any school where reserved for region locals. The regions where west, centrum, south-east and north-east. This arrangement made a difference primarily for the centrum region, where locals suffer the hardest competition from other regions. Inter-regional applications are more uncommon in the periphery (Lødding and Helland 2007). There were more east to west applications, than west to east. Most of the east to west applications had high grades and high education parents. Low grades lower secondary graduates from west would rather attend private upper secondary schools than public alternatives on the east (Hansen 2005). The arrangement did not change the ethnic segregation levels as originally intended, but there were small changes in some schools' proportions of low grade students (Andersen 2014). It is unclear what, if any, influence the affirmative action had on the generalization from 2007 applicants, regarding their applications relative to their graduation peers. However, because applications before and after are made under different circumstances, it can make choices before and after the changes less similar. Thus, I will not calculate the inter-class coefficients between graduates before and after 2008.

4.2 Influencing the choice

Like Jansson et al. (under review) (Jansson 2017) I expect applicants to choose “[..] according to their capabilities and exogenous preferences, which are likely to reflect their upbringing.” Exogenous preferences are in this context influences stemming from outside the individuals' graduation class, in their lower secondary school. If there are systematic relationships between background characteristics and preferences in education, we can expect applicants

\textsuperscript{15} Because the survey was conducted post acceptance, the can be some uncertainty whether or not it reflected thoughts pre application. For instance, the respondents might remember how much they wanted the current school, partly influenced by how much they liked it when questioned.
from similar backgrounds to choose somewhat similarly, invariant of their personal relations at the school. In addition to exogenous preferences we can expect *endogenous* effects, as the graduates may have made influential relationships within their class at lower secondary school. This can affect their choice in two ways: First, their curricular preferences and aspirations might have been altered by their interactions with their friends, such as gaining or losing interest in specialization tracks. Second, the graduates might be motivated to keeping their lower secondary school friendships, and therefore apply to the same school and specialization as their friends to maximize time spent together. Both mechanisms can potentially express homophily. Either the overrepresented influence from self-similar peers, or the direct preference for their self-similar peers in the coming year(s). While Jansson et al.’s (under review) (Jansson 2017) data were not able to separate the two endogenous effects, the choice of upper secondary schooling might allow for some insight to the distinction, because it contains the distinction between school and track. This will be elaborated in section

<table>
<thead>
<tr>
<th>Choice of track</th>
<th>Choice of school</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primal source</td>
<td>%</td>
</tr>
<tr>
<td>Counselor</td>
<td>23.8</td>
</tr>
<tr>
<td>Parents</td>
<td>18.5</td>
</tr>
<tr>
<td>Friends</td>
<td>17</td>
</tr>
<tr>
<td>Siblings</td>
<td>7.6</td>
</tr>
<tr>
<td>Internet</td>
<td>6.8</td>
</tr>
</tbody>
</table>

From Lødding & Helland 2007, Table 4. N=1733.

A survey asking post-applicants in Oslo about their most important source of information regarding the choice of upper secondary school revealed that friends where most reported response for the choice of school (Table 4). For the choice of track, friends was the third most reported response, surpassed by student counselor and parents, sources that might appear more competent on long term outcomes (Lødding & Helland 2007). An explanation to this finding could be that choosing track regards ones' professional life while choosing school more regards ones’ temporary whereabouts. When questioned about important factors when making the choice, about equal proportions of students agreed with the statement “To get away from previous classmates” as with the statement “Parents opinion of school”. Thus, negative social relations alone weigh about as heavy as parents’ opinions. Positive social relations was even more important. More than twice as many fully agreed with the statement “Important to know other students at the school”, than the previous two statements. Together
this suggests that the choice of upper secondary school is a choice that is more about social environment than education. But as noted, the choice of track is a different kind of choice. The upper secondary track is the first and for many, one of the most important career choice in their life. They choose if they want to become a mechanic, plumber, social care worker, pursue an academic preparation track, among other choice. Hence, the students are influenced on this choice more so from adults. Many students emphasize their career when they make their choice of education. Only 9.8% reported fully agreed with the statement “School was more important to me than track”, whereas 19.7% fully agreed with “Track was more important to me than school”. And 75.7% partially or fully agreed with the statement that “To attend a school which offer my preferred Vg2 Track” was important. I interpret this as the choice of track was the primal choice, and the choice of school a secondary choice. It can also be noted that more students reported that they were less sure about what school to prioritize than what track to prioritize. Thus, the choice of school will be more susceptible to spontaneous influences.

In sum I interpret these responses as following (Table 5): The choice of school is more susceptible to graduation class endogenous influences, than the choice of track. The choice of track is the most important choice, and the choice of school is a secondary choice. Thus, the effect of social relations is prominent in the choice of school but is constrained by the choice of track.

16 Vg2 is the second year of upper secondary education and involves additional specialization of track.
Table 5: Students reported considerations on the choice of upper secondary education

<table>
<thead>
<tr>
<th>Questions</th>
<th>Fully disagree</th>
<th>Partially disagree</th>
<th>Partially agree</th>
<th>Fully agree</th>
<th>No answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>School was more important to me than track</td>
<td>29,8</td>
<td>34,3</td>
<td>23,5</td>
<td>9,8</td>
<td>2,6</td>
</tr>
<tr>
<td>Track was more important to me than school</td>
<td>16,3</td>
<td>31,9</td>
<td>28,5</td>
<td>19,7</td>
<td>2,6</td>
</tr>
<tr>
<td>Important to know other students at the school</td>
<td>19,8</td>
<td>21,2</td>
<td>36,1</td>
<td>21,9</td>
<td>1,1</td>
</tr>
<tr>
<td>Parents opinion of school was important</td>
<td>27,6</td>
<td>26,8</td>
<td>33,6</td>
<td>10,8</td>
<td>1,3</td>
</tr>
<tr>
<td>To get away from previous classmates</td>
<td>28,5</td>
<td>28,5</td>
<td>31,3</td>
<td>10,5</td>
<td>1,2</td>
</tr>
<tr>
<td>To attend a school which offer my preferred Vg2 Track</td>
<td>8,4</td>
<td>14,3</td>
<td>35,6</td>
<td>40,1</td>
<td>1,6</td>
</tr>
<tr>
<td>I was unsure what track to prioritize</td>
<td>46,6</td>
<td>15,7</td>
<td>22,9</td>
<td>13,2</td>
<td>1,6</td>
</tr>
<tr>
<td>I was unsure what school to prioritize</td>
<td>33,8</td>
<td>16,2</td>
<td>29,7</td>
<td>18</td>
<td>2,4</td>
</tr>
</tbody>
</table>

From Lødding & Helland 2007, Appendix v.1 & v.2

Lødding and Helland (2007) conducted a factor analysis of surveyed motives for applying to upper secondary school among 2006-07 first year upper secondary students and revealed three main factors which they labeled the learning motivated, the diversity motivated, and the supporters of region based affirmative action. The learning motivated factor can have some implications on our measurement of homophily\textsuperscript{17}.

\textsuperscript{17} The supporters of region based affirmative action factor consisted of judgments on the fairness of the affirmative action, and valuing short commute influenced position on this factor. This factor is not relevant to this thesis, and will not be discussed further.

The diversity oriented valued an environment with students of a different language and cultural background than their own. Low diversity orientation indicated that this was not important, not that they were negative to diversity. The factor also includes a positive attitude towards minority language students, and finding pure native schools boring. There were gender differences in this factor as well. Females where slightly positive, while males where indifferent. Non-Western immigrant descendents where more diversity oriented than natives; Diversity is a social orientation, and can be viewed as an emphasis on personal relations. The valuing a different cultural background than your own item makes this factor a question of cultural heterophily, for all respondents. But the positive attitude towards minority language students and negative attitude (finding it boring) towards pure native schools means that this factor have different meanings for minority language students and natives. For minority students this is a matter of homophily, and for natives this is a matter of heterophily. Thus, the fact that non-western immigrants descendents where more diversity oriented than natives could mean that 1) The dependents prefer diversity more diversity than natives 2) The dissentents and natives are expressing homophily. This factor is not relevant to this thesis, and will not be discussed further.
The learning motivated, reported a more positive attitude towards school, and that competent teachers where important when picking school. The learning motivation where significantly more prominent among females then male, and non-Western immigrant descenders more motivated by this than natives;

This motivation emphasis preferences that are exogenous stable, making personal relations relatively less important. And together with less importance of friends, homophily will be less visible with our measurement. Thus, we can expect females and non-western immigrants to express less homophily than on other socio-metric instruments of homophily.

4.3 Combining choices of school and track into a network-metric

The choice of upper secondary education in Norway has two components. The applicants simultaneously chose their prioritized educational track and their prioritized school. In this section I will discuss the possibilities of utilizing this twofoldness to make inferences on how homophily is expressed through the choices of upper secondary education. Given the two elements of the choice of upper secondary education, the choice of school and the choice of track, there are six different combinations of how we can calculate similarities between graduates’ choices (Figure 11). Any pair of graduates can apply to the: same school, same track, same school and track, same track at different schools, same school but different tracks, or they can apply to different tracks at different schools.

![Figure 11: Different possible similarities in combinations of upper secondary choice](image)

Based on the empirical findings of Lødding & Helland (2007) outlined in section 4.2 I argue that the two parts of the choice of upper secondary education contain somewhat different considerations. Friends were reported to be a more important source of information when considering choice of school than choice of track. And family (parents and siblings) where more important when considering choice of track than choice of school (Table 6). This holds
for the sources relative importance, such that friends are more important than family when considering school and family more important than friends when considering track.

Table 6: Difference between friends and family on the importance of upper secondary education by track and school

<table>
<thead>
<tr>
<th>Primal source of information</th>
<th>Choice of track</th>
<th>Choice of school</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friends</td>
<td>17%</td>
<td>28.7%</td>
</tr>
<tr>
<td>Parents and siblings</td>
<td>26.1%</td>
<td>21.8%</td>
</tr>
</tbody>
</table>

Based on Lødding & Helland 2007, Table 4: Respondents marking of most important source of information when choosing upper secondary education. N=1733.

A plausible explanation of this pattern can be: While the choice of track is a choice of the applicants' professional life, the choice of school is a choice of where the applicants spend their next 2-3 years. In this sense, the track is a long term more important choice. Parents and student counselor can be perceived by the graduates as more competent and relevant sources of information than peer graduates on matters regarding the choice of track. And choice regarding ones' professional life is arguably more influenced by life aspirations and consequently, parents or other sources exogenous to the classroom are likely to influence their choice through expectations and as sources of inspiration. Curricular interests and life aspirations can also be transmitted endogenously (between classmates), making classroom associates more likely to choose same track. But the choice of school should be less subjected to exogenous influences, and thus less similar at the inter-class level. Using this as a base I expand the causal model and form a choice dimension that should best reflect the friendship dynamics. The expanded causal model shows the effect of classmate relations on the choice of upper secondary school, and the effect of the background characteristics directly on the choice of track (visualized in Figure 12).
If graduates consider this choice of track as more important than the choice of school, we can expect the choice of track to have priority over choice of school. Thus, the choice of school is conditioned on the choice of track. If the track is set exogenously, then the graduates might consult their classmate friends about what school they shall attend to. Friends that have the same track-ambitions will likely try to attend the same school. There might be differences in grades that could separate friends in terms of what schools are achievable. But this is also the case for non-friends, and graduates would arguably be more prone to apply to a school with entrance requirements above their grades if their friends with higher academic achievements where applying to the school. Also, friends that have different track ambitions would also be more inclined to attend the same school, if the school offers both tracks. Although not all schools have this opportunity, we can expect this trend more among friends than among non-friends. In sum we can expect graduate friends that coordinate their choices to apply more often to the same upper secondary school (Figure 13). Furthermore, something we can expect friends to do, to a far less degree than non-friends or non-coordinating graduates, is to apply to the same specialization track but at different schools. With these expectations I will now outline how I calculate the adjacency matrices that reflects these expectations from the possible combinations outlined in the beginning of this section. The categorization of choices in the adjacency matrices is visualized in Figure 14.

Figure 12: Extended causal model
I consider the graduates that apply to the same school as more likely to have coordinated choices and I will assign their choice adjacency 1. The graduates that apply to the same tracks but at different schools are considered more likely to be uncoordinated or non-friends and I will assign their choice adjacency -1. This will be my dependent variable. It has the properties of emphasizing the most important part of the choice of upper secondary education; the school. It penalizes extra, dissimilar schools when choice collaborations are extra likely to result in a similar choice of school; when the graduates apply to the same track. This is done without compromising similarities when graduates do not apply to the same track, which friends not necessarily do.

**Figure 13**: Personal classmate homophily in the extended causal model

**Figure 14**: The socio-metric instrument
4.4 Validity

4.4.1 External validity

Because the approach uses educational choice as an indicator of social relations, a caution on the external validity should be noted. There might be friendship homophily that does not affect the educational choice but is nonetheless important for other inquiries. I will only be able to infer on the homophily that affects the educational choices, among the lower secondary graduates in Oslo region. Any theorizing based on the empirical findings in this thesis should take this into account.

4.4.2 Internal validity

The potential for a reversed causal relationship are not likely present, as background characteristics are something the individuals are born with and not a consequence of choice later in life.

A possible objection to this claimed inference of relational attributed homophily is related to what Manski (1993) defines as correlated effect, a problem for peer effect measures in general. Correlated effects are defined as when similar behavior emanates from the individuals’ similar characteristics or similar institutional environment. Similar characteristics is in this case taken care of by the inter-class coefficient (see section 4). The inter-class coefficient also takes care of similarities in institutional environments at school level. However, similarities in the institutional class level environment is not necessarily addressed by the inter-class coefficient. An example of such a possible correlated effect can be the teacher of the class. Teachers can affect this measure of homophily, if the teacher would treat students differently depending on their characteristics. For example, it is plausible that some teachers will consciously or unconsciously separate the genders to different tasks, thus facilitating similar interests and/or intra gender relations. However, I argue that this does not pose a problem. First, any systematic gender specific interest would have to be local to the graduation class, in order to not be controlled away by the school level effect. That is, the teacher would separate the boys and girls to different tasks without the task being gender specific across classes. Thus, the problem would only arise if different teachers conduct different gender separations. Further, any separation is likely to include increased intra group
contact, and in length homophily. This begs the question of what we aim to measure. The homophily effect is something that emerges between the individuals\textsuperscript{18}. It is in a sense an aggregated phenomenon, existing at the class level. If teachers or other institutional environmental factors facilitates homophily behavior, it should not be excluded from the analysis.

\textsuperscript{18} While it can partly be a product of preferences at the individual level, it is likewise a product of the peers’ response to that.
5 Empirical results

In this chapter I will apply the socio-metric instrument (section 4.3) to the background dimensions: Gender, immigration status and combinations of these and country of origin. The purpose of this chapter is to demonstrate how the metric can provide insight on homophily. I have no theory’s I want to test per se, but will formulate some expectations along the way, based on what the previous research outlined in section 2.2, have found. At the end of the chapter I will discuss the analyses.

I will start the chapter with some data trimming and descriptive statistics on the dataset I use in the later analyses. Then I begin to analyze gender as a social dimension. Next, I construct and analyze a set of immigration status dimensions. First, I use the cruder natives versus immigrants (first and second-generation combined) distinction. Then I follow up with a somewhat finer detail analysis, looking at natives versus first-generation immigrants and natives versus second-generation immigrants separate. At last I also separate the immigration analyses into different genders as well as immigrant generations. The analyses will reveal if these groups make choices that correspond to there being an interpersonal homophily effect between classmates across the dimensions I distinguish the groups.

5.1 Dataset

The dataset originally contained the population who graduated from a Lower secondary school, registered to Oslo municipality, the year the individuals first graduated from lower secondary education. Access to the data was acquired through the "Ethnic segregation in schools and neighborhoods: consequences and dynamics“ research project at the Department of Sociology and Human Geography at the University of Oslo. The data used are individual anonymized observations, with information on gender, immigrant status, country of origin, lower secondary graduation year, lower secondary school, first priority choice of upper secondary school and first priority choice of upper secondary track.

The total number of graduates in the original dataset was 56083. Who graduated between 2005 and 2015 (Table 7).
Table 7: Original dataset by graduation year

<table>
<thead>
<tr>
<th>Graduation Year</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>11608</td>
</tr>
<tr>
<td>2006</td>
<td>4856</td>
</tr>
<tr>
<td>2007</td>
<td>4611</td>
</tr>
<tr>
<td>2008</td>
<td>4907</td>
</tr>
<tr>
<td>2009</td>
<td>4738</td>
</tr>
<tr>
<td>2010</td>
<td>4692</td>
</tr>
<tr>
<td>2011</td>
<td>5070</td>
</tr>
<tr>
<td>2012</td>
<td>5122</td>
</tr>
<tr>
<td>2013</td>
<td>5133</td>
</tr>
<tr>
<td>2014</td>
<td>5248</td>
</tr>
<tr>
<td>2015</td>
<td>98</td>
</tr>
</tbody>
</table>

Because of the changes in upper secondary application systems in 2008 (see section 4.1), I remove applications before 2009. Apparently, the year 2015 only contained 98 graduates and is not complete. Observations from 2015 have therefore been removed. There are applications to 117 distinct Upper secondary schools. Not all schools and tracks have received applications every year, and it is likely that some of the choices have been unavailable some years.

Because the schools are anonymized I cannot know for which schools or tracks this applies.

Therefore, to ensure equal opportunities across cohorts, only applications to schools and tracks with applications in all years were included. The remaining dataset contains a total of 22102 graduates.

Table 8: Reduced dataset by application age

<table>
<thead>
<tr>
<th>Application Age</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>174</td>
</tr>
<tr>
<td>16</td>
<td>21546</td>
</tr>
<tr>
<td>17</td>
<td>351</td>
</tr>
<tr>
<td>18</td>
<td>26</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
</tr>
</tbody>
</table>

The graduates are between ages 14 to 20, with almost everyone graduating at age 16 (Table 8). The gender distribution is balanced with 11516 male graduates and 10586 female graduates (Table 9).
Table 9: Reduced dataset by gender

<table>
<thead>
<tr>
<th>Gender</th>
<th>n</th>
<th>proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>11516</td>
<td>0.5210388</td>
</tr>
<tr>
<td>Female</td>
<td>10586</td>
<td>0.4789612</td>
</tr>
</tbody>
</table>

There are 65 distinct Lower Secondary Schools. Most, but not all schools have graduates every year (Figure 15). In this thesis the graduation classes is defined as year specific cohort at individual schools. In total there are 344 distinct school graduation cohorts:

![Histogram of schools by their number of graduation cohorts.](image)

Figure 15: Reduced dataset: Histogram of schools by their number of graduation cohorts.

There are 17 distinct upper secondary schools with applications from all the years (Table 10).
Table 10: Reduced dataset: Choices of upper secondary schools

<table>
<thead>
<tr>
<th>USSchool</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>O01101657</td>
<td>288</td>
<td>249</td>
<td>278</td>
<td>227</td>
<td>204</td>
<td>126</td>
</tr>
<tr>
<td>O03375409</td>
<td>275</td>
<td>226</td>
<td>194</td>
<td>184</td>
<td>103</td>
<td>69</td>
</tr>
<tr>
<td>O03375425</td>
<td>158</td>
<td>148</td>
<td>149</td>
<td>108</td>
<td>98</td>
<td>85</td>
</tr>
<tr>
<td>O03375433</td>
<td>249</td>
<td>190</td>
<td>183</td>
<td>224</td>
<td>354</td>
<td>313</td>
</tr>
<tr>
<td>O03375437</td>
<td>929</td>
<td>880</td>
<td>666</td>
<td>881</td>
<td>737</td>
<td>678</td>
</tr>
<tr>
<td>O03375441</td>
<td>54</td>
<td>60</td>
<td>79</td>
<td>75</td>
<td>35</td>
<td>51</td>
</tr>
<tr>
<td>O03375445</td>
<td>201</td>
<td>135</td>
<td>121</td>
<td>91</td>
<td>105</td>
<td>113</td>
</tr>
<tr>
<td>O03375449</td>
<td>190</td>
<td>182</td>
<td>348</td>
<td>261</td>
<td>257</td>
<td>323</td>
</tr>
<tr>
<td>O03375453</td>
<td>75</td>
<td>54</td>
<td>49</td>
<td>60</td>
<td>42</td>
<td>48</td>
</tr>
<tr>
<td>O03375457</td>
<td>181</td>
<td>123</td>
<td>144</td>
<td>103</td>
<td>67</td>
<td>65</td>
</tr>
<tr>
<td>O03375465</td>
<td>388</td>
<td>336</td>
<td>457</td>
<td>394</td>
<td>366</td>
<td>372</td>
</tr>
<tr>
<td>O03375469</td>
<td>48</td>
<td>36</td>
<td>35</td>
<td>36</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>O03375473</td>
<td>188</td>
<td>154</td>
<td>157</td>
<td>162</td>
<td>156</td>
<td>137</td>
</tr>
<tr>
<td>O03375477</td>
<td>368</td>
<td>457</td>
<td>388</td>
<td>408</td>
<td>470</td>
<td>381</td>
</tr>
<tr>
<td>O03375485</td>
<td>236</td>
<td>238</td>
<td>252</td>
<td>307</td>
<td>270</td>
<td>270</td>
</tr>
<tr>
<td>O03376509</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>O05156561</td>
<td>187</td>
<td>214</td>
<td>261</td>
<td>281</td>
<td>194</td>
<td>293</td>
</tr>
</tbody>
</table>

Reduced frequencies of applications for Upper secondary schools

And 15 distinct upper secondary tracks with applications from all the years (Table 11).

Table 11: Choices of upper secondary tracks

<table>
<thead>
<tr>
<th>Track</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOLOV0</td>
<td>26</td>
<td>25</td>
<td>23</td>
<td>22</td>
<td>24</td>
<td>13</td>
</tr>
<tr>
<td>BABAT1</td>
<td>112</td>
<td>103</td>
<td>97</td>
<td>77</td>
<td>50</td>
<td>26</td>
</tr>
<tr>
<td>DHDHV1</td>
<td>69</td>
<td>50</td>
<td>56</td>
<td>51</td>
<td>36</td>
<td>26</td>
</tr>
<tr>
<td>ELELE1</td>
<td>253</td>
<td>257</td>
<td>257</td>
<td>254</td>
<td>169</td>
<td>190</td>
</tr>
<tr>
<td>HSHSF1</td>
<td>163</td>
<td>161</td>
<td>119</td>
<td>128</td>
<td>90</td>
<td>84</td>
</tr>
<tr>
<td>IDRET1</td>
<td>330</td>
<td>230</td>
<td>257</td>
<td>239</td>
<td>250</td>
<td>263</td>
</tr>
<tr>
<td>MDMDD1</td>
<td>232</td>
<td>185</td>
<td>223</td>
<td>206</td>
<td>191</td>
<td>209</td>
</tr>
<tr>
<td>MKMED1</td>
<td>402</td>
<td>321</td>
<td>332</td>
<td>321</td>
<td>259</td>
<td>232</td>
</tr>
<tr>
<td>RMRMF1</td>
<td>41</td>
<td>32</td>
<td>46</td>
<td>45</td>
<td>32</td>
<td>33</td>
</tr>
<tr>
<td>SSSSA1</td>
<td>84</td>
<td>59</td>
<td>45</td>
<td>26</td>
<td>28</td>
<td>25</td>
</tr>
<tr>
<td>STFOR1</td>
<td>80</td>
<td>55</td>
<td>91</td>
<td>58</td>
<td>41</td>
<td>86</td>
</tr>
<tr>
<td>STREA2</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>STSSA2</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>8</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>STUSP1</td>
<td>2145</td>
<td>2137</td>
<td>2137</td>
<td>2311</td>
<td>2268</td>
<td>2126</td>
</tr>
<tr>
<td>TPTIP1</td>
<td>76</td>
<td>60</td>
<td>80</td>
<td>57</td>
<td>36</td>
<td>29</td>
</tr>
</tbody>
</table>

Reduced frequencies of applications for Upper secondary tracks
Combined this amount to 76 distinct interactions of applications to tracks and schools (Figure 16).

![Figure 16: Reduced dataset: Histogram of distinct choices of upper secondary education by number of applicants.](image)

This trimmed dataset is the basis for the analyses in the rest of this chapter. In specific analyses the data used will be a subset of this dataset, as I use background dimensions that does not all the graduates fall into. There will also be some attrition due to lack variation in some of the classes and schools, which makes it impossible to estimate their standard errors.

## 5.2 Gender

As outlined in section 2.2.1, gender is a social dimension where we can expect to find social homophily. I noted that gender homophily is observed in children’s play patterns. That boys tend to play in larger groups than girls. And that boys build different friendship network structures, as a result of different strategies to solving friendship unbalances. Between the genders, children are more likely to solve transitive imbalance by either dropping a cross-gender friendship or even a same-gender friendship than to add a cross-gender friendship
Thus, we can expect cross-gender friendships to be less common than same-gender friendships. Shrum et al. (1988) found that the gender homophily is most prominent among the youngest children and weakens into adolescence. Our data is constructed on the choices of mostly 16-year-old adolescents. Hence, we can expect to find gender homophily.

Table 12: Gender homophily

<table>
<thead>
<tr>
<th></th>
<th>Males versus Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>JHC</td>
<td>$H_\Gamma_{XY}$ 0.0153***</td>
</tr>
<tr>
<td>Intra</td>
<td>$W_\Gamma_{XY}$ 0.0.1793***</td>
</tr>
<tr>
<td>Inter</td>
<td>$B_\Gamma_{XY}$ 0.002634***</td>
</tr>
<tr>
<td>N</td>
<td>22057</td>
</tr>
<tr>
<td>Classes</td>
<td>323</td>
</tr>
<tr>
<td>Schools</td>
<td>57</td>
</tr>
<tr>
<td>T2 Intra</td>
<td>0.000738</td>
</tr>
<tr>
<td>T2 Inter</td>
<td>0.000154</td>
</tr>
</tbody>
</table>

Note: * = p<0.05; ** = p<0.01; *** = p<0.001

I find a statistically significant JHC when I apply the network-metric to the gender dimension. If a pair of classmates have the same gender, they are more likely to have applied as if they coordinated their choices to attend the same school, than if they have different genders. The interpretation of this is that there is gender homophily among lower secondary graduation classmates. The inter-class coefficient is also significant, which means that the graduates are making somewhat similar choices on the basis of their gender alone. But almost all the similarity is found intra-class. The T2 variance is the amount of variance in effect sizes that is to effect variance, and not to sampling errors. Compared to the JHC, the T2 variance is very low. This means that the gender homophily is a stable phenomenon across the classes.

5.1 Ethnicity (immigrant status)

In section 2.2.2 I outlined the broader research on ethnic homophily. This social dimension is characterized by substantial homophily in ethnically diverse societies, in relationship ranging from the most intimate to the most distant. The homophily is attributed to residential and institutional structures as well as to substantial inbreed homophily (McPherson et al. 2001, 420-2). Opposed to gender, the homophily among schoolmates tends to increase by age (Shrum et al. 1988) until middle school, where it stabilizes. My dataset contains information
on the immigration background of the graduates, which imbeds ethnicity differences. Thus, I will start this section with analyzing the differences between immigrants and natives. Hence, the explanatory social dimension will be whether or not the graduate or its parents emigrated from another country to Norway. That is those registered with immigrant status A (Born in Norway to Norwegian-born parents) versus those registered with immigrant status B (Immigrants) or C (Norwegian-born to immigrant parents) by the standard of the national statistical institute of Norway (Statistics Norway, 2017).

Table 13: Immigration status homophily

<table>
<thead>
<tr>
<th></th>
<th>Natives versus Immigrants</th>
</tr>
</thead>
<tbody>
<tr>
<td>JHC [H[Gamma XY]]</td>
<td>0.01038***</td>
</tr>
<tr>
<td>Intra [W[Gamma XY]]</td>
<td>0.1089***</td>
</tr>
<tr>
<td>Inter [B[Gamma XY]]</td>
<td>0.0005095</td>
</tr>
<tr>
<td>N</td>
<td>18660</td>
</tr>
<tr>
<td>Classes</td>
<td>303</td>
</tr>
<tr>
<td>Schools</td>
<td>57</td>
</tr>
<tr>
<td>T2 Intra</td>
<td>0.002663</td>
</tr>
<tr>
<td>T2 Inter</td>
<td>0.001037</td>
</tr>
</tbody>
</table>

Note: * = p<0.05; ** = p<0.01; *** = p<0.001

I find a statistical significant JHC for this background dimension as well, and conclude that there is immigrant status homophily. Natives and immigrants tend to associate less with each other than within their own status group. Here, the inter class coefficient is very small and not significant. The choice similarities are almost exclusively attributed to them being classmates. The JHC is lower than what I found for the gender dimension and indicates that the gender of classmates better predicts who the graduates associates with, when applying to upper secondary school. Immigrant status have a higher intra-class T2 than gender. Hence the immigration status homophily appear less stable than gender homophily. That is, in some classes the ‘distance’ between natives and immigrants is longer than other classes. However, it could be reasonable to view first-generation-immigrants and second-generation immigrants as separate groups. As I will show in the next section combining immigrants of group B and group C masks a different pattern that can be unraveled by analyzing them separately.

5.1.1 First and second-generation versus natives, separate analyses

It has been found that immigration homophily was more prominent for first-generation immigrants than second (McPherson et al. 2001, 422), something one might expect from the
theories of causes of homophily (2.2). A) That people perceive dissimilar others as potentially more problematic relations than the more similar, thus gravitate towards the similar. B) Or that homophily may emerge through the dissolution of relations, where homogenous relationships are more stable. Either way one can expect the descendants of immigrants to have a better understanding with the natives, compared to those who immigrated themselves. This understanding can facilitate personal relations with natives through more successful communication and interaction, in accordance with explanation B. However, problematic relations between first-generation immigrants and the native population can be transferred to the second-generations’ relations with the native population through explanation A. Negative attitudes towards the natives can be inherited by the second-generation from the first, and native negative attitudes towards the first-generation can be generalized to the second-generation. Thus, curbing the interactions between the groups and reducing the effect of being second-generation immigrant versus first. Thus, I will see if the relationship with natives is closer for second-generation immigrants than for first-generation immigrants, by running separate models for the two groups of immigrants.

Table 14: Immigration status homophily. First and second-generation separate

<table>
<thead>
<tr>
<th></th>
<th>First-generation Immigrants</th>
<th>Second-generation Immigrants</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>JHC</td>
<td>0.01502***</td>
<td>0.008804*</td>
<td>0.006216</td>
</tr>
<tr>
<td>Intra</td>
<td>0.03836***</td>
<td>0.01118***</td>
<td>0.02718***</td>
</tr>
<tr>
<td>Inter</td>
<td>0.02334***</td>
<td>0.002375</td>
<td>0.02097***</td>
</tr>
<tr>
<td>N</td>
<td>13458</td>
<td>15773</td>
<td></td>
</tr>
<tr>
<td>Classes</td>
<td>279</td>
<td>295</td>
<td></td>
</tr>
<tr>
<td>Schools</td>
<td>56</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>T2 Intra</td>
<td>0.004279</td>
<td>0.002013</td>
<td>0.002266</td>
</tr>
<tr>
<td>T2 Inter</td>
<td>0.00155</td>
<td>0.00181</td>
<td>0.000396</td>
</tr>
</tbody>
</table>

Note: * = p<0.05; ** = p<0.01; *** = p<0.001

As expected I find stronger homophily between natives and first-generation immigrants than between natives and second-generation immigrants. All parameter as lower for the second-generation model than compared to the first-generation model. The JHC between natives and first-generation immigrants is about the same strength as what I found between males and females (5.2). But in the natives versus first-generation model, we also observe a stronger inter-class coefficient, and a much stronger intra-class coefficient. Natives and first-generation immigrants makes more within group similar choices even if they are not classmates. This indicate that there is difference between the groups on what schools are
preferable, unrelated to social relations within their graduation glass. On top of that, we observe homophily that interfere with the choices of upper secondary education within the classes. If we look at the natives versus second-generation immigrants model, the picture chances. Here, it is a much smaller and statistically insignificant inter-class coefficient. The intra-class coefficient is statistically significant, but much weaker than what we observe in the natives versus first-generation model. When we control for the inter-class coefficient, we get a somewhat weak JHC that is stronger than what we would expect in 95% trails without any classmate homophily between natives and second-generation immigrants. The overall trend compared to the models with both the first and second-generation immigrants, is that there is stronger homophily in the models with only first-generation immigrants. But if I try to compare the models in a statically matter, I find that the JHC in the natives versus first-generation -model is not statically larger than the JHC in the natives versus second-generation immigrants -model.

Analyzing the two groups of immigrants as separately revealed the stronger homophily between natives and first-generation immigrants. This leads me to try another analysis, where I separate the models also by gender. This follows in the next section.

5.1.2 Different gender homophily

As shown in the last section, it can be fruitful to separate heterogeneous subgroups in analyses. Overgeneralization will fail to recognize potential homophily between the subgroups and heterogeneity in the homophily with respect to the reference group. For example, homophily between first-generation immigrants and the reference group natives, was revealed when we separated them from second-generation immigrants. This begs the question of whether the groups can be separated into even more meaningful subgroups. In the previous research (see 2.2.2) on homophily we found stronger racial homophily among girls than boys. Boys tended to play in larger, less intimate groups than girls. And with larger groups, more diversity usually followed (McPherson et al. 2001, 421). Based on this, we could expect that we will find larger immigrant versus native homophily among girls than among boys. However, there is also possible that large play groups do not translate into homophily in choices regarding upper secondary education. To agree on such choices within a group becomes increasingly difficult with group size. Thus, the boys can be faced with the choice of who to attend with in the large play group and consolidate with only a smaller
closer group of friends where there could be extra homophily. Even if there is more
homophily among girls than boys when they choose playmates, this does not necessarily
translate to a choice of upper secondary education. As noted in section 4.2 Lødding &
Helland (2007)’s used a factor analysis to derive different motivations behind applications
into upper secondary school. One factor where “The learning motivated”, that emphasized
competent teachers when picking upper secondary schools. And I argued that this is a
preference that are exogenously stable, and that leaves personal relations relatively less
important. If personal relations are less important, then homophily will be less visible with
our measurement. The learning motivation where significantly more prominent among
females then male. Thus, we can expect females to express less homophily than on other
socio-metric instruments of homophily.

<table>
<thead>
<tr>
<th></th>
<th>Natives versus First-generation Immigrants</th>
<th>Natives versus Second-generation Immigrants</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>JHC $H[\Gamma_{XY}]$</td>
<td>0.01952**</td>
<td>0.008414</td>
<td>0.01111</td>
</tr>
<tr>
<td>Intra $W[\Gamma_{XY}]$</td>
<td>0.04312***</td>
<td>0.006279</td>
<td>0.03684***</td>
</tr>
<tr>
<td>Inter $B[\Gamma_{XY}]$</td>
<td>0.0235***</td>
<td>-0.002135</td>
<td>0.02573***</td>
</tr>
<tr>
<td>N</td>
<td>6566</td>
<td>7546</td>
<td></td>
</tr>
<tr>
<td>Classes</td>
<td>257</td>
<td>267</td>
<td></td>
</tr>
<tr>
<td>Schools</td>
<td>55</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>T2 Intra</td>
<td>0.005438</td>
<td>0.002825</td>
<td>0.002613</td>
</tr>
<tr>
<td>T2 Inter</td>
<td>0.00125</td>
<td>0.000646</td>
<td>0.000604</td>
</tr>
</tbody>
</table>

Note:* = p<0.05; ** = p<0.01; *** = p<0.001

When I analyze only male graduates’ immigrant status homophily, I find a stronger JHC
coefficient between natives and first-generation immigrants, compared to the model with both
genders. However, the statistical significance level is lower. The coefficient is larger than
99% that of random permutations, while the natives versus first-generation model with both
genders had a chance of occurring less than 0.1%, if there was no true effect. The fact that we
see a higher coefficient but lower statistical certainty, might indicate that I am pushing the
boundaries when it comes to specifying the groups. Because this reduces the classes. The
male second-generation immigrant versus natives’ model have about the same JHC as the
 corresponds to the model with both genders but are not statically significant even at the 0.5%
level. Hence, we see a large drop in homophily between the first and second-generation
models among the male graduates, but this difference is not statistically significant.
Table 16: Female Immigrants versus female natives

<table>
<thead>
<tr>
<th>JHC</th>
<th>$H[\Gamma_{XY}]$</th>
<th>$W[\Gamma_{XY}]$</th>
<th>$B[\Gamma_{XY}]$</th>
<th>Natives versus First-generation Immigrants</th>
<th>Natives versus Second-generation Immigrants</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intra</td>
<td>0.01905**</td>
<td>0.01736**</td>
<td>0.01689</td>
<td>0.04789***</td>
<td>0.03031***</td>
<td>0.01758*</td>
</tr>
<tr>
<td>Inter</td>
<td>0.02884***</td>
<td>0.001295***</td>
<td>0.01589*</td>
<td>5939</td>
<td>7059</td>
<td>0.01758*</td>
</tr>
<tr>
<td>N</td>
<td>249</td>
<td>262</td>
<td>0.006336</td>
<td>0.005993</td>
<td>0.000343</td>
<td></td>
</tr>
<tr>
<td>Schools</td>
<td>51</td>
<td>54</td>
<td>0.00238</td>
<td>0.002749</td>
<td>-0.000369</td>
<td></td>
</tr>
</tbody>
</table>

Note: * = $p<0.05$; ** = $p<0.01$; *** = $p<0.001$

The female immigrant homophily appears more stable from the first- to the second-generation models. The JHC for the female first-generation immigrants versus female natives’ model is about the same as the JHC for the male first-generation immigrants versus male natives model, with the same level of statistical significance. But contrary to the males, female second-generation versus natives JHC is still significant at the 99% confidence level, and almost as high as the female first-generation versus natives JHC. My expectations from previous research was that ethnic homophily was stronger among females than males. This analysis suggests that immigrant status homophily is more persistent from first-generation to second-generation immigrants among females than among males. However, it should be mentioned that the first and second-generation immigrants are perhaps different in demographic composition as well.
Figure 17: Immigrant status homophily by gender and different immigrant generations

I saw that the two genders were different with respect to immigrant status homophily. Analyzing separate models for the two genders found larger homophily compared to models that did not account for gender. However, there are another way we could include genders in the models. We can create groups based on both immigrant status and gender. A group for male natives, a group for male immigrants, a group for female natives and a group for female immigrants. Then we calculate the background adjacency matrix ($X_{ij}$) on whether the graduates belong to the same group. This approach will include any homophily between the male natives and female natives, or between male immigrants and female immigrants. Hence, we cannot know if it is gender or immigrant status that drives these results, from this model alone. And if I use this approach, I find no difference between the first- and second-generation models’ JHC. Perhaps the gender homophily is stronger for second generation immigrants than for the first generation immigrants.
Table 17: Female immigrants versus female natives versus male immigrants versus male natives

<table>
<thead>
<tr>
<th></th>
<th>First-generation Immigrants</th>
<th>Second-generation Immigrants</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>JHC $H[\Gamma_{XY}]$</td>
<td>0.01788**</td>
<td>0.01792***</td>
<td>-4.104e-05</td>
</tr>
<tr>
<td>Intra $W[\Gamma_{XY}]$</td>
<td>0.03174***</td>
<td>0.01996***</td>
<td>0.01179***</td>
</tr>
<tr>
<td>Inter $B[\Gamma_{XY}]$</td>
<td>0.01386***</td>
<td>0.002031***</td>
<td>0.01183***</td>
</tr>
<tr>
<td>N</td>
<td>14380</td>
<td>19906</td>
<td>0.01179***</td>
</tr>
<tr>
<td>Classes</td>
<td>309</td>
<td>313</td>
<td></td>
</tr>
<tr>
<td>Schools</td>
<td>56</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>T2 Intra</td>
<td>0.001222</td>
<td>0.000953</td>
<td>0.000269</td>
</tr>
<tr>
<td>T2 Inter</td>
<td>0.000304</td>
<td>0.000481</td>
<td>-0.000177</td>
</tr>
</tbody>
</table>

Note: * = p<0.05; ** = p<0.01; *** = p<0.001

5.1.3 Nationalities

Again, I saw that the two genders were different with respect to immigrant status homophily and treating the two groups of immigrants as one homogenous group underplayed homophily. I found by specifying first- and second-generation models. Immigrants are a heterogeneous in another way. Graduates classified as immigrants or parents of immigrants have emigrated from different places and can have a very different cultural backgrounds that can result in different relationships with the natives. Hence, I also include some models where I estimates homophily between immigrants from a selection of origin countries, and the natives.

Table 18: Immigrants by origin and generation

<table>
<thead>
<tr>
<th>Origin</th>
<th>First-generation</th>
<th>Second-generation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pakistan</td>
<td>462</td>
<td>2541</td>
<td>3191</td>
</tr>
<tr>
<td>Somalia</td>
<td>772</td>
<td>511</td>
<td>1301</td>
</tr>
<tr>
<td>Sweden</td>
<td>113</td>
<td>32</td>
<td>1090</td>
</tr>
</tbody>
</table>

In my data, the three most common countries of origin, where Pakistan, Somalia and Sweden. However, most of the swedes where neither first- nor second-generation, but third generation-immigrants, and categorized as natives. But because Sweden is probably the culturally the closest country to Norway, I include them as a reference for the Pakistanis and Somalis.
Table 19: First-generation immigrants by nationality versus natives.

<table>
<thead>
<tr>
<th></th>
<th>Natives versus First-generation Pakistanis</th>
<th>Natives versus First-generation Somalis</th>
<th>Natives versus First-generation Swedes</th>
</tr>
</thead>
<tbody>
<tr>
<td>JHC $H[\Gamma_{XY}]$</td>
<td>0.02017*</td>
<td>0.01232</td>
<td>0.01344</td>
</tr>
<tr>
<td>Intra $W[\Gamma_{XY}]$</td>
<td>0.05315***</td>
<td>0.04799***</td>
<td>0.03641***</td>
</tr>
<tr>
<td>Inter $B[\Gamma_{XY}]$</td>
<td>0.03298***</td>
<td>0.03566***</td>
<td>0.02298***</td>
</tr>
<tr>
<td>N</td>
<td>3135</td>
<td>5136</td>
<td>2320</td>
</tr>
<tr>
<td>Classes</td>
<td>105</td>
<td>145</td>
<td>52</td>
</tr>
<tr>
<td>Schools</td>
<td>33</td>
<td>42</td>
<td>32</td>
</tr>
<tr>
<td>T2 Intra</td>
<td>0.005742</td>
<td>0.000691</td>
<td>0.002039</td>
</tr>
<tr>
<td>T2 Inter</td>
<td>0.000718</td>
<td>0.000307</td>
<td>0.000604</td>
</tr>
</tbody>
</table>

Note: * = $p<0.05$; ** = $p<0.01$; *** = $p<0.001$

Between first-generation immigrants and natives, the JHC is strongest among the Pakistani immigrants and natives, with a $p$-value of 0.05. For Swedish and Somali immigrants, the JHC is not significant, and around the same magnitude. This is perhaps surprising as one might expect Swedes to be closer than Somalis to native Norwegians. However, this measurement indicates no statistically significant homophily for either dimension.

Table 20: Second-generation immigrants by nationality versus natives.

<table>
<thead>
<tr>
<th></th>
<th>Natives versus Second-generation Pakistanis</th>
<th>Natives versus Second-generation Somalis</th>
<th>Natives versus Second-generation Swedes</th>
</tr>
</thead>
<tbody>
<tr>
<td>JHC $H[\Gamma_{XY}]$</td>
<td>0.01904***</td>
<td>0.01991*</td>
<td>-0.01245</td>
</tr>
<tr>
<td>Intra $W[\Gamma_{XY}]$</td>
<td>0.06338***</td>
<td>0.06405***</td>
<td>-0.008837</td>
</tr>
<tr>
<td>Inter $B[\Gamma_{XY}]$</td>
<td>0.04434***</td>
<td>0.04413***</td>
<td>0.003614</td>
</tr>
<tr>
<td>N</td>
<td>8159</td>
<td>4706</td>
<td>1068</td>
</tr>
<tr>
<td>Classes</td>
<td>210</td>
<td>129</td>
<td>18</td>
</tr>
<tr>
<td>Schools</td>
<td>50</td>
<td>38</td>
<td>13</td>
</tr>
<tr>
<td>T2 Intra</td>
<td>0.003298</td>
<td>0.001708</td>
<td>0.000569</td>
</tr>
<tr>
<td>T2 Inter</td>
<td>0.000701</td>
<td>0.000341</td>
<td>0.000522</td>
</tr>
</tbody>
</table>

Note: * = $p<0.05$; ** = $p<0.01$; *** = $p<0.001$

For the second-generation immigrants models the picture is somewhat different. In the JHC between second-generation Pakistani immigrants the coefficient magnitude is about the same as for first-generation Pakistani immigrants, but with a higher statistical significance level ($p < 0.001$). There are more second-generation Pakistani immigrant than first-generation Pakistani immigrants, and this is perhaps contributing to the higher statistically significance. The second-generation Somali immigrants’ model have a higher JHC than the first-generation Somali immigrants’ model and have a $p$-value of 0.05. There are not more second-generation Somali immigrants than first-generation Somali immigrants. Hence, it indicates that in fact
first-generation Somali immigrants is closer to their native classmates than second-generation Somali immigrants are close to their native classmates. The Swedish second-generation immigrant model has in fact a negative but not statistically significant JHC.

Table 21: Immigrant status homophily by immigrant country of origin and different immigrant generations

In the models with first-generation immigrants, I found statistically significant results only between Pakistanis and natives, but not between natives and Swedes or Somalis. However, in the models with second-generation immigrants, I also found statistically significant results between natives and Somalis. This is contrary to what has been found in other research that showed that immigration homophily was more prominent for first-generation immigrants than second (McPherson et al. 2001, 421). Therefore I also include models where I measure homophily between Pakistanis and Somalis, without natives, to see how this changes from first to second-generation.
The first-generation model where I measure homophily between Somalis and Pakistanis yield a relative large JHC, which have p-value of 0.05. However, this model has few graduates included. From the 462 Pakistanis and 772 Somalis, only 231 graduates where included in the analysis. The once excluded where in classes with too little variation in either the choices or background to compute the QAP-standard errors. Hence, it could be that the first-generation immigrants Somali attends different schools than the first-generation Pakistani immigrants or second-generation Somali immigrants, that could influence the homophily. But in the classes with enough first-generation Somali and Pakistani immigrants there is a relative large JHC between these groups. But in the second-generation immigrant models, I find a slightly negative and not significant JHC between Somalis and Pakistanis. Thus, the first-generation Somali immigrants is closer to the natives than first-generation Pakistani immigrants, and second-generation Somalis are closer to second-generation Pakistani immigrants than to natives. However, a more thorough analysis of the classes should be in place before these results are taken as something more than aggregated patterns.

### Table 22: First-generation immigrants by nationality.

<table>
<thead>
<tr>
<th></th>
<th>First-generation Pakistanis versus Natives</th>
<th>First-generation Somalis versus Natives</th>
<th>First-generation Somalis versus First-generation Pakistanis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>JHC</strong></td>
<td>0.02017*</td>
<td>0.01232</td>
<td>0.09007*</td>
</tr>
<tr>
<td>Intra</td>
<td>0.05315***</td>
<td>0.04799***</td>
<td>0.08146</td>
</tr>
<tr>
<td>Inter</td>
<td>0.03298***</td>
<td>0.03566***</td>
<td>-0.008604</td>
</tr>
<tr>
<td>N</td>
<td>3135</td>
<td>5136</td>
<td>231</td>
</tr>
<tr>
<td>Classes</td>
<td>105</td>
<td>145</td>
<td>44</td>
</tr>
<tr>
<td>Schools</td>
<td>33</td>
<td>42</td>
<td>28</td>
</tr>
<tr>
<td>T2 Intra</td>
<td>0.005742</td>
<td>0.000691</td>
<td>0.012071</td>
</tr>
<tr>
<td>T2 Inter</td>
<td>0.000718</td>
<td>0.000307</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: * = p<0.05; ** = p<0.01; *** = p<0.001
Table 23: Second-generation immigrants by nationality.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>JHC</td>
<td>0.01904***</td>
<td>0.01991*</td>
<td>-0.003327</td>
</tr>
<tr>
<td>Intra</td>
<td>0.06338***</td>
<td>0.06405***</td>
<td>0.01948</td>
</tr>
<tr>
<td>Inter</td>
<td>0.04434***</td>
<td>0.04413***</td>
<td>0.022181*</td>
</tr>
<tr>
<td>N</td>
<td>8159</td>
<td>4706</td>
<td>1026</td>
</tr>
<tr>
<td>Classes</td>
<td>210</td>
<td>129</td>
<td>101</td>
</tr>
<tr>
<td>Schools</td>
<td>50</td>
<td>38</td>
<td>34</td>
</tr>
<tr>
<td>T2 Intra</td>
<td>0.003298</td>
<td>0.001708</td>
<td>0.000731</td>
</tr>
<tr>
<td>T2 Inter</td>
<td>0.000701</td>
<td>0.000341</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: * = p<0.05; ** = p<0.01; *** = p<0.001

5.2 Discussion

In this chapter I have applied the Jansson et al. (under review) (Jansson 2017) with my suggested socio-metric instrument on a variety of combinations gender and immigrant status dimensions. The aim has been to show that this measurement can provide useful information. I have found gender homophily and immigrant status homophily. I have showed that we can separate first-generation immigrants from second-generation immigrants. There is lower homophily between the natives and second-generation immigrants than between natives and first-generation immigrants in general. But this generational decline is weaker among females than among males. And I have found less homophily between natives and first-generation Somali immigrants than between natives and second-generation Somali immigrants, something that is opposite to the general trend. However, it could be that the first-generation immigrants Somali attends different schools than the first-generation Pakistani immigrants or second-generation Somali immigrants, that could influence the homophily. That is, they could be in class with natives with different attitudes towards immigrants, for instance. Hence, attributing such trend to the country of origin is premature. This could be the case for any other separation into different models like the difference in immigration homophily for the two genders as well, but it’s more likely that genders are equally distributed across the classes. However, this problem raises an important limitation with the methodology as it stands. In order to make more accurate distinctions we need to take into account what classes are similar and what classes are different in the separate models. As of now, we can only infer about homophily at the aggregate level. Knowing properties about what classes’ exhibit more
or less homophily can be very useful for answering theoretical questions. Similarly, a possibility to include more exploratory variables into the same manner where one controls for the relationship between $X_{ij}$ and $Y_{ij}$ by $Z_{ij}$ would be useful for some more detailed theoretical questions.
6 Concluding remarks

The aim of this thesis was to provide a socio-metric instrument that can be used in the novel methodology developed by Jansson et al. (under review) (Jansson 2017) to infer homophily. I used lower secondary school graduates’ choice of upper secondary education as a basis for this. This choice has two components; track and school. I have used findings from previous research to optimize the social considerations of the graduates in this choice, which I operationalized into one socio-metric instrument. And I have applied this instrument to a set of background dimensions. I have showed how we can split and combine the background dimensions to get more nuanced pictures of the important dimensions of homophily. Coupling the socio-metric with the methodology of Jansson et al. (under review) (Jansson 2017) shows promising ability to infer homophily from registry data. However, both the methodology and the socio-metric instrument suggested by this thesis is novel advancements, and more research should be conducted for further developments. Hence, I will outline some suggestions for further research based on what I have learned. I suggest four main areas of improvement.

This thesis provides a socio-metric instrument to measure homophily that is especially useful for researchers interested in segregation in lower and upper secondary education. It can be used to test what background dimensions are important for which classmates’ the lower secondary graduates gravitate towards, in their choice of upper secondary school. This have direct implication on the demographics of upper secondary schools and can have implication on the informal segregation in the destination upper secondary school. It can also be used to approximate self-selection bias for researchers interested in estimating peer effects in lower secondary school.

The thesis also provides some notions on the application of the Pearson’s correlation coefficient in the methodology developed by Jansson et al. (under review) (Jansson 2017). This involve that the coefficients maximum positive value is constrained by the number possible choices is smaller than the number of independent categories. And that, coefficient’s the negative maximum is constrained by the sizes of the independent categories.

The thesis also conducts empirical analyses of gender and ethnic/ immigrant status homophily. The empirical analyses show homophily on the dimensions gender and immigrant
Further I have found stronger homophily between natives and first-generation immigrants than between second-generation and natives. The homophily is more persistent from first to second-generation versus natives among female graduates than among male graduates. And I have found aggregated trends of less homophily between natives and first-generation Somali immigrants than between natives and second-generation Somali immigrants.

6.1 Suggestions for further research

To advance this methodology further I suggest that future research conduct more traditional social network surveys on the lower secondary school graduates, with socio-metric instruments as reported friends and who the graduates discuss the choice of upper secondary school with and correlate these metrics with the socio-metric instrument suggested in this thesis. Research like this can provide us with more clarity in what part of the social world are able to measure with this socio-metric instrument, and what we are missing. Perhaps new information can lead to more sophisticated socio-metric instrument to measure homophily with the registry data.

In section 3.3 I outlined a potential problem with applying the methodology to situations with few choices. Because of this problem, I have constrained the analysis to the school district of Oslo, which have many available choices. However, there can be situations where one would be interested in homophily in rural areas with few choices. Hence, if one could build a model that takes into account the number of choices, it could open a wider empirical terrain for the methodology.

Some theoretical questions might require us to distinguish between different schools and different classes based on properties of the schools and classes. For instance, if one is interested in the relationship between the share of immigrant students at schools and the natives versus immigrant homophily in the classes at the schools. This require a different meta-analysis approach, that does not aggregate the intra-class coefficients and inter-class coefficients into two separate measures with the meta-analysis, before comparing them. It would perhaps be more useful to construct an intra-class to inter-class difference at every school first. This could then be correlated with other properties that schools can exhibit.
Another possible improvement to the methodology is a possibility to include control variables. As of now, the only way to account for different dimensions in a model is to interact the backgrounds to specify different groups or to exclude graduates from the analysis. It could be useful to include confounding dimensions directly in the models. For example, academic achievements are a potential confounding dimension that should be accounted for. But there might also be demographic differences across the dimensions that might influence the


DerSimonian, Rebecca and Nan Laird. 1986. "Meta-analysis in Clinical Trials." Controlled Clinical Trials 7, no. 3: 177-88. DOI: 10.1016/0197-2456(86)90046-2

Ferrara, Emilio, Pasquale De Meo, Salvatore Catanese, and Giacomo Fiumara. 2014. "Detecting Criminal Organizations in Mobile Phone Networks." Expert Systems With Applications 41, no. 13: 5733. ISSN: 0957-4174


Hâncean, Marian-Gabriel and Matjaž Perc. 2016. "Homophily in Coauthorship Networks of East European Sociologists." Scientific Reports 6: 36152. DOI: 10.1038/srep36152

Hansen, Marianne Nordli. 2005. "Ulikhet I Osloskolen ; Rekruttering Og Segregering." Tidsskrift for Ungdomsforskning 5, no. 1: 3-26. ISSN: 1502-7759

Jansson, Fredrik, Gunn Birkeland, Mats Lillehagen M. "Segregation within School Classes: Deriving Influential Social Ties from Register Data". under review.


Krackardt, David. 1987. "QAP Partiialling as a Test of Spuriousness." Social Networks 9, no. 2: 171-86. DOI: 10.1016/0378-8733(87)90012-8

Louch, Hugh. 2000. "Personal Network Integration: Transitivity and Homophily in Strong-tie Relations." Social Networks 22, no. 1: 45-64. DOI: 10.1016/S0378-8733(00)00015-0

Lødding, Berit and Håvard Helland. 2007. "Alle Får, Men Hvem Får Hva? Evaluering Av Oslo Kommunes Inntaksordning Til Videregående Opplæring I Studieretning for Allmennfag/utdanningsprogram for Studiespesialisering". NIFU STEP. ISSN: 1504-1824


Pearson, Karl. 1895. "Note on Regression and Inheritance in the Case of Two Parents." Proceedings of the Royal Society of London 58: 240-42. ISSN: 03701662


https://www.ssb.no/en/utdanning/statistikker/vgu

Statistics Norway 2017 "Classification of immigration categories" Read June 2017  
https://www.ssb.no/en/klass/klassifikasjoner/82

All sources used in this assignment are provided

Number of words in this thesis is 19208
Appendix:

Appendix A: R-code

```r
### This code is based on the code developed by Fredrik Jansson, that can be found at

# QAP according to the socio-metric instrument
# Combines elements adjacency matrices 1 and 2 such as adjacencies is: 1 = sosdim 1 is shared and -1 if sosdim
# 2 is shared but not 1.
# Then runs qaps and stores the results

# Takes:
# data.m - list of intra-class adjacency matrices
# data.m.b - list of inter-class adjacency matrices
# groupVSgroup - string indicating the name of the background dimension
# X - Independent variable number, acording to g$sosdim

# Saves inter and intra lists of QAP- objects, containing:
# gs - list of two adjacency matrices per group
# info - list of data frames divided into groups
# q - list of qap-results
# sosdim - list of soscial dimentions
# classifiers - super- and subclassifiers

US_School_VS_Track_and_not_School <- function(data.m, data.m.b, groupVSgroup, X){
  comp <- function(data, Y, X, id){
    file <- qap(data, Y, X, id)
    saveRDS(file, file = print(filename))
    file
  }
  choice <- function(data){
    g <- data
    g$gs <- lapply(data$gs, function(x) {
      x[1,] <- ((x[1,]*2-1) * x[2,]) + (x[1,]*)((x[2,]-1)^2)
      x
    })
    g$sosdim[1] <- "USSchool VS USTrack at different schools"
    g
  }
  filename <- paste(groupVSgroup, "_intra", sep="")
  m <- choice(data.m)
  m <- comp(m, 1, X, id="LS_School_Cohort")
  print(m$compTime / 3600)
  filename <- paste(groupVSgroup, "_inter", sep="")
  mb <- choice(data.m.b)
  mb <- comp(mb, 1, X, id="LS_School")
  print(mb$compTime / 3600)

  out <- list()
  out$sintra <- m
  out$sinter <- mb
  out
}

# Construct inter-class adjacency matrices from data between subgroups
```

# Constructs adjacency matrices based on having the same value(s) for a set of social dimensions, for each
group in the data. Elements are matched to other
# elements in the same group, excluding elements in the same subgroup.
#
# data - a data frame sorted on superclass
# superclass - a classifier defining group membership
# subclass - a classifier defining subgroup to be excluded
# sosdim(s) - vector with names of columns in data containing values of sosdim(s)
#
# Returns a data.m.b list containing:
# gs - list of adjacency(one for every sosdim) matrices per group
# info - list of data frames divided into groups
# sosdim - list of social dimensions
# classifiers - super- and subclassifiers

pairwise.m.b <- function(data, superclass, subclass, sosdim) {
  # data sorted on superclass
  ds <- length(sosdim)
  n <- nrow(data)
  changing <- c(1, which(data[2:n, superclass] != data[1:(n-1), superclass]) + 1, n+1)
  classes <- length(changing) - 1
  gs <- list()
  infos <- list()
  out <- list()
  for (k in 1:classes) {
    m <- changing[k+1] - changing[k]
    g <- array(NA, c(ds, m, m))
    info <- data[changing[k]:(changing[k+1]-1),]
    if (m > 1) {
      start <- changing[k] - 1
      for (i in 1:(m-1)) {
        for (j in (i+1):m) {
          if (data[start+i, subclass] != data[start+j, subclass]) {
            for (D in 1:ds) {
              g[D, i, j] <- 1 * (data[start+j, sosdim[D]] == data[start+i, sosdim[D]])
              g[D, j, i] <- 1 * (data[start+j, sosdim[D]] == data[start+i, sosdim[D]])
            }
          }
        }
      }
    }
    gs[[k]] <- g
    infos[[k]] <- info
  }
  out$gs <- gs
  out$info <- infos
  out$sosdim <- as.list(sosdim)
  out$classifiers <- list(paste("Superclass:" , superclass), paste("Subclass:" , subclass))
  out
}

# Construct intra-class adjacency matrices from data
#
# Constructs adjacency matrices based on having the same value(s) for a set of social dimensions, for each
group in the data. Elements are matched to other
# elements in the same subgroup.
#
# data - a data frame sorted on superclass
# superclass - a classifier defining group membership
# subclass - a classifier defining subgroup membership
# sosdim(s) - vector with names of columns in data containing values of sosdim(s)
#
# Returns a data.m list containing:
# gs - list of two adjacency matrices per group
# info - list of data frames divided into groups
# sosdim - list of social dimensions
# classifiers - super- and subclassifiers

pairwise.m <- function(data, superclass, subclass, sosdim) {
  # data sorted on superclass
  ds <- length(sosdim)
  n <- nrow(data)
  changing <- c(1, which((data[2:n, subclass] != data[1:(n-1), superclass]) & (data[2:n, subclass] != data[1:(n-1), subclass]) + 1, n+1))
  classes <- length(changing) - 1
  gs <- list()
  infos <- list()
  out <- list()
  for (k in 1:classes) {
    m <- changing[k+1] - changing[k]
    g <- array(NA, c(ds, m, m))
    info <- data[changing[k]:changing[k+1]-1,]
    if (m > 1) {
      start <- changing[k] - 1
      for (i in 1:(m-1)) {
        for (j in (i+1):m) {
          for (D in 1:ds) {
            g[D, i, j] <- 1 * (data[start+i, sosdim[D]] == data[start+j, sosdim[D]])
            g[D, j, i] <- 1 * (data[start+i, sosdim[D]] == data[start+j, sosdim[D]])
          }
        }
      }
    }
    gs[[k]] <- g
    infos[[k]] <- info
  }
  out$gs <- gs
  out$info <- infos
  out$sosdim <- as.list(sosdim)
  out$classifiers <- list(paste("Superclass:", superclass), paste("Subclass:", subclass))
  out
}

# Reduces adjacency matrices s according to filtering-conditions

# filtering - string with filtering conditions
# data - List with:
#  gs - list of two adjacency matrices per group
#  info - list of data frames divided into groups
#  sosdim - list of social dimensions
#  classifiers - super- and subclassifiers

subsetting <- function(data, filtering) {
  out <- data
  pos <- list()
  pos <- lapply(data$info, function(df)(as.logical(eval(parse(text=filtering)))))
  out
groups <- length(data$info)
for (k in 1:groups){
    out$gs[[k]] <- data$gs[[k]][,pos[[k]],pos[[k]]]
    out$info[[k]] <- data$info[[k]][pos[[k]],]
}
pos <- lapply(out$info, nrow) > 1
out$gs <- out$gs[pos]
out$info <- out$info[pos]
out
}

# Metanalysis over QAP tests
# Combines effect sizes and distributions from QAP tests into one measure and mean distribution
# using metanalytic methods.
#
# Input parameters are generated by metaqap:
# q - lists of QAP test results
# info - the input data (optional)
#
# Returns a metaqap.stats object containing:
# sel - the groups that were included in the test (based on sufficient amount of data)
# Y - effect sizes for each group
# dist - QAP distribution for each group
# V - QAP variance from each group
# k - number of groups
# T2 - tau^2 for random-effects metaanalysis
# m - combined measure
# p - combined p-value
# mean.dist - a mean distribution from the QAP tests
# m.fixed - combined measure using fixed-effect metaanalysis
# p.fixed - combined p-value using fixed-effect metaanalysis
# mean.dist.fixed - a mean distribution from the QAP tests using fixed-effect metaanalysis
# N - number of subjects included

metaqap.stats <- function (q, info = NULL, id = NULL) {
    Y <- sapply(q, [[, "testval")
    dist <- sapply(q, [[, "dist")
    V <- apply(dist, 2, var, na.rm = T)
    W <- 1 / V
    sel <- !is.na(W) & !is.na(Y) & !is.infinite(W) & colSums(is.na(dist)) < 50
    dist <- dist[sel]; Y <- Y[sel]; V <- V[sel]; W <- W[sel]
    k <- length(Y)
    # Estimate tau^2 to perform a random-effects meta-analysis
    T2 <- (sum(W * Y^2) - (sum(W) * sum(Y)^2 / sum(W) - k + 1)) / (sum(W) - sum(W^2) / sum(W))
    T2 <- max(0, T2)
    m.fixed <- weighted.mean(Y, 1 / V)
    m <- weighted.mean(Y, 1 / (V + T2))
    mean.dist.fixed <- apply(dist, 1, function(x) weighted.mean(x, 1 / V, na.rm = T))
    mean.dist <- apply(dist, 1, function(x) weighted.mean(x, 1 / (V + T2), na.rm = T))
    p.fixed <- sum(m.fixed < mean.dist.fixed) / length(mean.dist.fixed)
    p <- sum(m < mean.dist) / length(mean.dist)
    out <- list()
    out$sel <- sel; out$Y <- Y; out$dist <- dist; out$V <- V; out$k <- k; out$ST2 <- T2
    out$m <- m; out$p <- p; out$mean.dist <- mean.dist
    out$m.fixed <- m.fixed; out$p.fixed <- p.fixed; out$mean.dist.fixed <- mean.dist.fixed
    if (!is.null(info))
        out$N <- sum(sapply(info[sel], nrow))
        out$id <- sapply(info, function(x) levels(droplevels(x[, id])))
    class(out) <- c("metaqap.stats", "metaqap")
    out
}
# Metaanalysis over QAP tests using Fisher transformed values

# A wrapper function to metaqap.stats first Fisher transforming measures before carrying out a
# metaanalysis and then transforming them back.

# Input and output variables are the same as for metaqap.stats. Output is a metaqap.stats.fisher
# object.
metaqap.stats.fisher <- function (q,info=NULL, id=NULL) {
  Y <- sapply(q,[[,"testval"
  Y[Y==1] <- 0.9999; Y[Y==1] <- -0.9999 # arbitrarily chosen to avoid division by 0
  dist <- sapply(q,[[","dist"
  dist[dist==1] <- 0.9999; dist[dist==1] <- -0.9999
  Y.z <- 0.5 * log((1+Y)/(1-Y))
  dist.z <- 0.5 * log((1+dist)/(1-dist))
  q.z <- q
  for (i in 1:length(q.z)) { q.z[[i]]$testval <- Y.z[i]; q.z[[i]]$dist <- dist.z[i] }
  out <- metaqap.stats(q.z,info, id)
  out$Y <- Y[out$sel]; out$dist <- dist
  out$m <- (exp(2*out$m) - 1) / (exp(2*out$m)+1)
  out$mean.dist <- (exp(2*out$mean.dist) - 1) / (exp(2*out$mean.dist)+1)
  out$m.fixed <- (exp(2*out$m.fixed) - 1) / (exp(2*out$m.fixed)+1)
  out$mean.dist.fixed <- (exp(2*out$mean.dist.fixed) - 1) / (exp(2*out$mean.dist.fixed)+1)
  class(out) <- c("metaqap.stats.fisher","metaqap.stats")
  out
}

# Difference between two metaanalytic QAP measures

# Gives the difference between two effect sizes and between their two associated metaanalytic
# QAP distributions.

# q1, q2 - two metaqap.stats objects

# Returns a metaqap.diff object containing:
# m - difference between measures
# p - proportion of values in dist greater than m
# dist - difference between metaQAP distribution values (paired in stored order)
# m.fixed - difference between fixed-effect measures
# p.fixed - proportion of values in dist.fixed greater than m.fixed
# dist.fixed - difference between metaQAP distribution values using fixed-effect measures
metaqap.diff <- function (q1,q2) {
  m <- q1$m - q2$m
  dist <- q1$mean.dist - q2$mean.dist
  p <- if(m<0) {
    sum(m>dist) / length(dist)
  } else{
    sum(m<dist) / length(dist)
  }
  m.fixed <- q1$m.fixed - q2$m.fixed
  dist.fixed <- q1$mean.dist.fixed - q2$mean.dist.fixed
  p.fixed <- if(m.fixed<0){
    sum(m.fixed>dist.fixed) / length(dist.fixed)
  } else{
    sum(m.fixed<dist.fixed) / length(dist.fixed)
  }
  out <- list(m=m,p=p,mean.dist=dist,m.fixed=m.fixed,p.fixed=p.fixed,dist.fixed=dist.fixed)
  class(out) <- c("metaqap.diff","metaqap.stats")
  out
}


# Runs QAP over the matrices an gives out null hypothesis distribution
# Yij - Dependent variable number, according to g$sosdim
# Xij - Independent variable number, according to g$sosdim
# g - List with:
# gs - list of two adjacency matrices per group
# info - list of data frames divided into groups
# sosdim - list of social dimensions
# classifiers - super- and subclassifies

# Returns a list containing:
# gs - list of two adjacency matrices per group
# info - list of data frames divided into groups
# q - list of qap-results
# sosdim - list of social dimensions
# classifiers - super- and subclassifies

qap <- function (g, Yij, Xij, id=NULL){
  time <- proc.time()
  require(sna)
  out <- list()
  pos <- sapply(g$gs,function(x) isTRUE(var(as.vector(x[,Yij,]),na.rm=T)>0))
  g$gs <- lapply(g$gs[pos], '[', c(Yij, Xij),,)
  g$info <- g$info[pos]
  q <- lapply(g$gs,function(x) qaptest(x,gcor,g1=1,g2=2,mode="graph"))
  out$g <- g$gs; out$info <- g$info; out$q <- q; out$Yij <- g$sosdim[Yij]; out$Xij <- g$sosdim[Xij];
  out$classifiers <- g$classifiers
  out$N <- sapply(out$info, nrow)
  out$id <- sapply(out$info, function(x) levels(droplevels(x[,id])))
  out$compTime <- proc.time() - time
  class(out) <- "qap"
  out
}