UNIVERSITY OF AMSTERDAM

MASTER THESIS

Understanding Segregation Within Schools Using A Social Network Model

Author:
Luyao Xu

Supervisor:
Dr. Michael Lees

Examiner:
Dr. Willem Boterman

A thesis submitted in fulfillment of the requirements for the degree of MSc Computational Science

in the

Institute for Advanced Study

as part of

Computational Modeling of Primary School Segregation

July 21, 2021

Declaration of Authorship

I, Luyao XU, declare that this thesis titled, "Understanding Segregation Within Schools Using A Social Network Model" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"Research is what I'm doing when I don't know what I'm doing"

Wernher von Braun

UNIVERSITY OF AMSTERDAM

Abstract

Faculty of Science

Institute for Advanced Study

MSc Computational Science

Understanding Segregation Within Schools Using A Social Network Model

by Luyao XU

School segregation, defined as the difference in the distribution of a particular population across schools, has been found in many multi-ethnic societies. And it is feared to have negative effects on education equality and social cohesion. Various attempts have been taken in order to combat segregation across schools and promote the integration of ethnicities. However, it is also found that mixing of students with different backgrounds may aggravate segregation within schools, particularly in terms of the friendship networks of students. In order to make effective strategies to integrate students of various ethnic groups within schools, it is essential to understand how school ethnic composition relates to friendship segregation and interethnic friendships. This thesis proposes an agent-based model to simulate the formation of segregation in students' friendship networks. We find that the proposed model is able to replicate the patterns observed in real friendship network data collected from schools. Through further experimentation, we show that friendship segregation peaks when mixed schools have two dominant ethnic groups, and it declines when there are three groups. Friendship segregation can be reduced even further when the third group can act as a "bridge" between the other two groups.

Acknowledgements

This is to acknowledge all people who have given me inspiration and support. Without them I would not have been able to complete my thesis, especially during this tough time of the corona pandemic.

First of all I would like to truly thank my supervisor Dr. Mike Lees for his valuable guidance. He has given me many new perspectives whenever I was seeking help. Also on a personal level he has given me invaluable support. Next I would like to sincerely thank Dr. Willem Boterman for sharing his precious knowledge of social science and reminding me of the social implications of my project. I also would like to express a great appreciation towards Eric Dignum for his generous support. He has always offered help and given me timely feedback. A special thank you is given to Dr. Debraj Roy for his advice concerning sensitivity analysis. Also, I am very grateful for Nigel van Herwijnen and Thomas van der Veen. They have been an inspiration, and it was a pleasure to have weekly meetings with them to share interesting findings and experiences.

Finally, I would like to thank my friends and family for their unconditional support. Particularly, I want to express my deepest gratitude to my partner Matthijs de Wit for his immense care and support for me during this special time period. He has always accompanied me, motivated me, and given me endless patience while comforting me when times were tough. I would not have been able to spend this time in the Netherlands so happily and complete my thesis without them.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Background

1.1.1 Segregation across schools in the Neterlands

Over the past decades, school segregation has become an increasingly important social issue in the Netherlands. The Dutch primary education system consists of pupils with various ethnic backgrounds, especially in the large cities. According to Statistics Netherlands (CBS[1]), in 2019-2020, 26.9% of all primary school students have an immigrant background, with 68.9% of those immigrant backgrounds being non-western, such as Moroccan and Turkish. In large cities, the proportion of students with an immigrant background are much higher. In Amsterdam, 59.9% of all students have an immigrant background, with 75.5% of those being non-western. This had led to many schools having a majority of students with a non-Dutch background. Instead of all the schools being equal in ethnic compositions, they have become very segregated. Some schools have a high concentration of native Dutch students, while many other schools have a majority of non-western students such as Turkish, Moroccan, Surinamese and Antillean. [8, 30, 54].

Segregation across schools is driven by multiple factors, such as residential segregation across neighborhoods, parental school choice, and the distribution of different types of schools [8]. Residential segregation and school segregation are closely intertwined in many contexts. In the United States, it is, depending on the state, not always allowed for parents to register their children to schools outside of their designated school [60]. In this context, residential segregation plays a large part in the segregation across schools. On the contrary, in the Dutch context, parents have freedom of school choice as a constitutional right [9]. Therefore there are no catchment areas for schools, where children are assigned a school based on their address. However, residential patterns are still found to explain much of the school segregation in urban regions [9]. This is mainly because parents give a high priority to the distance from home to school when they make school choices [26]. Ethnic minority parents tend to give a higher priority to distance than Dutch parents, and Dutch parents deem the "match" between home and school more important than ethnic minority parents [26, 8]. These differences in the parental preferences for schools enhances segregation across schools.

[1] https://opendata.cbs.nl/#/CBS/nl/dataset/83295NED/table?dl=8C08

1.1.2 Interventions combating school segregation

Segregation across schools is feared to have negative effects on education equality and social cohesion. According to the annual report on integration in the Netherlands in 2018, native Dutch students have a higher chance to be recommended to enroll in senior general secondary (HAVO) or pre-university education (VWO) in the final year of primary school than students with an immigrant background [53]. Moreover, students have fewer opportunities to contact peers of different backgrounds and form interethnic friendships in a segregated school. Schools tend to be the first social system that shapes children's attitudes towards other ethnicities and prepares students for integrating into a multi-ethnic society. The segregation across schools leads to a lack of contact between native Dutch students and immigrant students, which can result in more ethnic prejudice [40]. It is also shown that having fewer interethnic contacts with natives is not beneficial to the socioeconomic positions of immigrants [24, 31].

In order to eliminate the negative effects of school segregation on educational equality and ethnic integration, a series of policies are made aiming at mixing students of various ethnic groups into schools. In some municipalities, measures are taken in order to influence parental preferences for mixed schools. Parents are provided with well-structured information on neighborhood schools through websites, brochures, and information markets. Cities like Rotterdam also organize school tours for parents to visit schools in a given neighborhood. Some municipalities in the Netherlands have taken further steps to control the application and acceptance of students by schools. In Nijmegen, a central registration system was introduced to allocate students to primary schools. One goal of the system is to approach a ratio of 30% disadvantaged students and 70% advantaged students within schools [58].

1.1.3 "Hidden segregation" within schools

While those interventions can limit the uneven distribution of immigrant and native Dutch students across schools, they can not prevent the formation of "hidden segregation" within schools and classrooms. Previous studies suggest that the mixing of students with different backgrounds may aggravate segregation within schools, particularly in terms of friendship networks of students [2, 38, 57]. This means that students in schools that are integrated at the population level might have more friendships within their own ethnic groups. In this thesis we will focus on understanding how friendship segregation emerges in mixed schools, and provide a model that can be applied to study segregation in social networks.

Scholars explain the formation of higher friendship segregation in more integrated schools as a result of different mechanisms. The two most popular of these are *contact theory* and *homophily theory*. On the one hand, mixed schools expose students to peers of different ethnic backgrounds, providing students with more opportunities for interethnic contacts [27]. These contacts under the right conditions can effectively contribute to positive interethnic relations and reduction of prejudice [27, 52]. On the other hand, students prefer to connect with others similar to themselves. The preference for similarity in social relations, also known as homophily, widely exists in various social networks, and on a wide range of dimensions, such as race, gender, religion, education, etc [36]. The contact opportunity structure in schools, combined with students' homophily on multiple dimensions, plays an important role in structuring friendship networks in schools. Apart from that, the tendency of students to reciprocate friendships and to connect friends of friends, to some extent shapes friendship networks.

Students' ethnic homophily, however, can be affected by the ethnic composition of schools. According to the ethnic competition theory [57, 49], as the population of ethnic minorities grows to a large size, ethnic majorities are likely to feel threatened by minorities. And this can intensify students' preferences for same-ethnic friends, as well as their negative attitudes towards other ethnic groups. Moody's research on United State schools [38] shows that students' preference for same-ethnic friendships reaches a maximum in moderately mixed schools. Smith et al. [49] find that the ethnic homophily of native students increases when the group of immigrant students becomes large. These findings suggest that the relationship between the diversity of schools and friendship integration is not necessarily positive.

1.2 Research questions

Mixed schools can, in principle, offer a way for students to create interethnic friendships that can mitigate the potential for ethnic prejudice in later life. This can in turn lead to integrated multi-cultural societies in the long-term. However, the empirical evidence demonstrates that simply having a mixed school population does not necessarily lead to increased interethnic friendships. In order to make effective strategies to integrate students from different ethnic groups within schools, it is essential to understand how school ethnic composition affects friendship segregation and interethnic friendships. Therefore, this study aims to answer the following questions:

1. *Is there a relationship between school ethnic composition and friendship segregation?*

2. *What is the optimal ethnic composition for mixed schools to have the lowest level of friendship segregation?*

To concretize the above questions, we set up scenarios where we need to allocate students from two or three ethnic groups into schools. Our goal is to mix students from different ethnic backgrounds into schools while minimizing the level of segregation in their friendship networks.

1.3 Research approach

As mentioned in section 1.1.3, the formation of friendship segregation within schools could be the result of multiple mechanisms and school ethnic composition has impact on friendship segregation at least by changing the contact structure and students' ethnic homophily. In an attempt to study how school ethnic composition affects friendship segregation, it is critical to disentangle the impacts of school ethnic composition on contact structure and on students' ethnic homophily.

This will be achieved by the agent-based model we develop in this thesis. In the model, students' homophily is determined by a homophily function, we can keep students' homophily unchanged to see how school ethnic composition influences friendship segregation through contact structure. Moreover, we can also correlate students' ethnic homophily with school ethnic composition to show how school ethnic composition influences friendship segregation through students' ethnic homophily. In addition, the model uses a utility function with components including utility from direct, mutual, and indirect links, which allows us to investigate how mechanisms such as reciprocity and transitivity drive friendship segregation.

The homophily function in the model is defined over social distance. The concept of social distance is used to measure how dissimilar two people are with respect to their characteristics. Because homophily is characterized by multiple dimensions, it is important to also take characteristics other than ethnicity into consideration when we study friendship segregation along ethnic lines. Therefore, we define social distance over multiple attributes, such as ethnicity, gender, and social class. These characteristics are weighted in the function of social distance, and their weights determine students' homophily with respect to each attribute.

We will first study the behavior of the model using artificial data. And to further evaluate and validate the model we will use empirical data. The original aim of this thesis was to use friendship network data collected from Dutch schools. However, due to the COVID-19 [55] pandemic, it is difficult to obtain data from Dutch schools. Therefore we use data from the US schools instead, and discuss how the results from the US schools relate to the Dutch context in Chapter 7. The US school data used in this thesis is from Wave I of Add Health[19], a study of adolescents in junior high or high schools in the United States. It contains information on both the friendship networks of students and their individual attributes including race, gender, and grade. The data is described in further detail in Chapter 5.

1.4 Research contribution

The model proposed in this thesis is a continuation of the work by Angelo Mele [37]. We made several changes to the utility function. First, we included an explicit cost component to get a better insight into the effects of cost on friendship choice. Second, we normalized the indirect utility component, otherwise it might dominate other components. Finally we implemented a homophily function which controls the probability of students making friends based on social distance.

On top of that, we did more computational analysis for studying the model behavior and applying the model into our research. Both global sensitivity analysis and local sensitivity analysis were implemented to give insight into the relative importance of model parameters. Also, we validated the model against real data, to show how well the model performs on real friendship networks. Finally we conducted some theoretical experiments using the model to demonstrate the effects of school ethnic composition on friendship segregation. From the experiments, we find that friendship segregation is correlated to school composition. Friendship segregation peaks when schools have two dominant racial groups, and it declines when schools have three racial groups. It can be reduced even further when the third group can act as a "bridge" between the other two groups.

1.5 Thesis overview

The thesis is organized in the following way. In chapter 2, a literature study about segregation across and within schools in the Netherlands will be presented. This is followed by a review of modeling social networks, where Angelo Mele's model will be discussed in detail. Finally, we will introduce the measure of segregation used in the thesis. In chapter 3, the agent-based model of social network formation will be described. In chapter 4, both global sensitivity analysis and local sensitivity analysis will be performed. The performance of the model against real social network data will be evaluated in chapter 5. In chapter 6, experiments will be conducted to answer the research questions. The main findings and conclusions will be discussed in chapter 7.

Chapter 2

Literature Review

In this section, we first give a review of segregation across schools in the Netherlands, as this thesis is a part of the COMPASS project [1]. Then we focus on segregation within schools, as it is the main focus of the thesis. Next, we introduce several models for generating friendship networks. This is followed by a review of measures for quantifying segregation in friendship networks.

2.1 Segregation across schools

2.1.1 Overview of school segregation in the Netherlands

The degree of school segregation along ethnic lines in Dutch cities is quite high, and it is not only an issue in large cities like Amsterdam and Rotterdam, but also in medium-sized cities like Breda, Haarlem, and Zwolle [8]. Wolfgram measured the degree of school segregation by comparing the school population to the neighborhood population [61]. He finds that among the schools in the 38 largest cities in the Netherlands, 63% of them reflect the population of their neighborhoods decently. 17% out of the schools are "too white" compared to their neighborhoods, and 20% of the schools are "too black". Besides, the degree of school segregation for the four largest Dutch cities (Amsterdam, Rotterdam, The Hague, and Utrecht) is at least 40%, which means less than 60% of the schools in these cities can reflect the population of their neighborhoods. Moreover, if a neighborhood is "whiter", school populations reflect the composition of the neighborhood better.

2.1.2 Factors affecting school segregation

Segregation across primary schools in the Netherlands is driven by a number of factors, which include but not limited to residential segregation across neighborhoods, parental school choice, and distribution of different types of schools, e.g. schools that are based on religious backgrounds [8, 26, 17]. And the main considerations behind school choice for native Dutch and ethnic minority parents are different [26, 8].

School segregation is closely intertwined with residential patterns of ethnic groups [17]. During the immigration waves in the 1960s and 1970s, a large number of immigrants such as Turkish, Moroccan, and Surinamese entered the Netherlands. Because these people tended to have lower incomes, they were concentrated in lower-priced housing, particularly spatially and socially separate areas of large cities [12]. The influx of immigrants and the spatial separation of ethnic groups, to

[1] https://www.compass-project.nl/

a large extent, changed the composition of schools and triggered a rapid increase in the number of "black" schools [8, 29]. Schools became more segregated at the end of the 1980s than they were before the immigration waves [12].

Research on 22 Dutch cities shows that the levels of school segregation along ethnic lines are generally higher than residential segregation [8]. The discrepancy between them can mainly be explained by parents' school choices. Parents make choices based on many factors such as distance, school profile, and school quality. The travel distance to schools seems to be a key consideration for parents' school choice, based on the fact that most children attend schools in their residential neighborhoods. It is also supported by a survey conducted by Karsten et al, which shows that almost 70% of 931 parents in Amsterdam did not consider any schools outside their postcode districts [26]. Moreover, they find a correlation between the likelihood of considering schools further away and parents' ethnic and socioeconomic backgrounds. For instance, native Dutch parents are more likely to choose schools outside their local districts than ethnic minority parents.

For some parents, the profile of schools is a more important consideration than distance. These parents usually have preferences for specific curriculum, religious denomination, or pedagogical principles [33, 15]. If parental preferences for a specific school profile are correlated with a specific ethnic or socioeconomic group, schools that match their preferences tend to attract students from that group [33]. It thereby has an influence on segregation across schools.

The preference for school quality differs between native Dutch and ethnic minority parents. Studies by Karsten et al indicate that native Dutch parents take the "match" between home and school as the most important factor for school choice, whereas ethnic minority parents put more emphasis on the degree of differentiation and the academic standard of schools [26]. These differences in the parental school choices enhances segregation across schools.

2.1.3 Interventions combating school segregation

Advocates of school desegregation argue that segregated schools have negative effects on students' interethnic attitudes and social cohesion, because the lack of interethnic contact in segregated schools can result in more ethnic prejudice. In order to combat school segregation, attempts are taken by local governments and school boards. Most attempts focus on three aspects: student enrollment procedures, providing parents better information, and facilitating parent initiatives to realize mixed schools [59].

"Double waiting lists" had been used in Rotterdam in 2004, which allowed oversubscribed schools to give priority to students who would contribute to the ethnic diversity of the school. In addition, Rotterdam has organized school tours for parents to visit (specific) schools in a given neighborhood. During the tour, parents can share information and exchange views about schools [29].

In Nijmegen, a central registration system has been used to allocate students into primary schools, including religious schools. The system still takes efforts to accommodate parental school choice, while makes priority rules for oversubscribed schools. The top priority is given to children with siblings at the preferred school, and then to children who live in the neighborhood. The subsequent priority is given to children who would contribute to a ratio of 30% disadvantaged and 70% advantaged students at the school [58]. Apart from that, segregated schools with a high proportion of disadvantaged students have received additional funding from the municipality of Nijmegen [29]. Various measures have been taken by other municipalities, such as providing better information on schools to disadvantaged parents, and encouraging high-educated parents to enroll their kids at segregated but well-performed schools [29, 59].

2.2 Friendship segregation within schools

So far, we have looked at segregation across schools, however the underlying motivation is to reduce segregation on a social level. School as the first social system children experience, is expected to expose students to a variety of backgrounds and allow students to have friendly, day-to-day contact with peers of other backgrounds. This can, in principle, reduce students' prejudice towards other ethnic groups and promote interethnic relationships in the long run.

Mixed schools are believed to create a diverse environment where students have opportunities to make interethnic friendships and thereby promote ethnic integration. However, substantive segregation emerges within mixed schools when students restrict their friendships to their own ethnic groups [38]. The correspondence between ethnic or other socioeconomic characteristics and friendship choice is defined as friendship segregation [38]. Empirical research has found friendship segregation along ethnic lines within ethnically diverse schools [38, 39, 49, 32, 22]. Based on an analysis of friendship network data from US schools, Moody finds that there is a strong and generally positive relationship between the ethnic diversity of schools and friendship segregation. And this relationship is nonlinear: friendship segregation reaches a maximum in moderately diverse schools and it decreases in highly diverse schools [38]. Moody's findings suggest that simply exposing students to ethnically diverse settings does not promote ethnic integration.

Scholars have come up with three mechanisms (homophily theory, contact theory, and ethnic competition theory) to explain why friendship segregation emerges in mixed schools. These mechanisms can also explain the nonlinear relationship between friendship segregation and school diversity in Moody's research.

2.2.1 Homophily theory

People prefer friends who are similar to themselves in multiple dimensions, such as age, sex, race, and education. The preference for similarity in social relations is defined as the homophily principle [36, 18, 35]. Homophily is considered as the key mechanism that drives friendship segregation in mixed schools.

Racial and ethnic homophily play a major part in structuring networks in ethnically diverse settings [36]. Students have preferences for same-race friends, leading to friendships concentrating within races. Previous research finds that school ethnic composition has an impact on students' ethnic homophily [38, 49, 23, 57]. Furthermore, Smith et al examine 529 friendship networks in English, German, Dutch, and Swedish schools, and they conclude that the effects of school ethnic composition on the ethnic homophily are different between immigrant students and native students [49]. The ethnic homophily of immigrant students disproportionately increases when there are more peers of their own ethnic groups. And the ethnic homophily of native students is relatively low until the population of immigrant students becomes dense.

Because homophily is characterized by multiple dimensions, it is important to also take other characteristics into account in the studies of friendship segregation along ethnic lines. For example, segregation with regard to gender is a powerful phenomenon of childhood [34]. Homophily for gender has been observed in many children's friendship networks [18, 34, 48, 56]. Students' homophily for race might be affected by their homophily for gender: if students have stronger preferences for same-gender friends than same-race friends, they might make many interracial friendships of same-gender. In fact, there is evidence showing that same-gender is an important factor for the stability of interracial friendships. Hallinan et al examine longitudinal data from 375 fourth- to seventh-grade students in 16 desegregated classrooms, and they find that interracial friendships of the same gender are more likely to endure more than interracial friendships between boys and girls [18]. In this thesis, homophily is defined as a function of multiple characteristics. And by weighting these characteristics in the homophily function, we can investigate how students' homophily for one characteristic affects segregation along other lines.

2.2.2 Contact theory

The contact theory proposed by Allport et al states that under appropriate conditions, intergroup contact could efficiently reduce prejudice between members of majority and minority groups [16]. Previous research indicates that contacts between people of different racial groups allow them to gain more knowledge about each other, resulting in less prejudice [45]. There is also research showing that contact across racial groups is associated with less out-group anxiety and more self-exposure [41].

Allport et al specify four key conditions of contact to reduce prejudice. First, equal status is essential for reducing prejudice through contact. Friendships are unlikely to form between members of different racial groups if they do not engage equally in a relationship [20]. Second, common goals that can not be attained individually, such as team sports, can positively affect interracial contacts. Moreover, interracial cooperation is needed to achieve these common goals. Finally, interracial contact has more positive effects with support of authorities, law or customs. These conditions suggest that simply exposing students to other racial groups does not always cause reduced prejudice and positive racial relations. Also, if minority students have preferences for same-race friendships, concentrating them in a school would allow them to find their desired number of friends within their own racial groups, which will aggravate friendship segregation [38].

2.2.3 Racial competition theory

Racial competition theory argues that racial majorities are more likely to feel threatened by minorities when they see a dense population of minorities [5]. According to this theory, increasing interracial contacts will result in more negative attitudes towards racial out-groups. This has been observed in a research by Scheepers et al [46], they investigate ethnic exclusionism in European countries and find that the larger the proportion of non-EU citizens in a country, the stronger the support for ethnic exclusionism. They conclude that people who live in individual competitive conditions feel threatened by ethnic out-groups, which in turn exacerbates ethnic exclusionism.

Although this theory is mainly used for adults, adolescents who are developing their ethnicity-based social identity are also found to perceive competition when the presence of ethnic out-groups becomes salient [62, 57]. Vervoort et al [57] examine 2386 students from 117 school classes in the Netherlands, and they find that in school classes with more than 50% of ethnic minority students, both ethnic majority and minority students show more negative out-group attitudes. Their explanation for this finding is that when the group of ethnic minority students becomes large, majority students feel threatened by minority students, therefore they show more negative attitudes towards them. At the same time, the perception of prejudice is likely to cause more negative attitudes from ethnic minorities towards the majority students.

2.3 Models of social network formation

In order to study segregation within friendship networks, we need a model to generate social networks while allowing us to apply the previously mentioned theories. Here we give an overview of several commonly used models for generating social networks and introduce the framework we will build upon.

One of the most important network formation models is the random network model developed by Paul Erdős and Alfred Rényi, named the Erdős–Rényi model [42]. The model constructs networks that are truly random: it starts with N isolated nodes, and then assigns a certain number of links among them randomly. A variant of the Erdős–Rényi model connects the isolated nodes with a fixed probability, instead of a fixed number of links. The networks generated by the random network model, however, fail to capture many features of real networks. For example, real networks have more highly connected nodes than the random network model could account for [4]. It implies that real networks are not truly random.

In contrast to random network models, the Barabási-Albert model [1] assumes that networks grow through the addition of new nodes, and new nodes are more likely to connect to nodes with more links. These two assumptions: growth and preferential attachment, allow the Barabási-Albert model generate scale-free networks. The Barabási-Albert model starts with m_0 nodes and randomly assigned links. At each time step, a newly arriving node j is connected to m ($m \leq m_0$) pre-existing nodes in the network. The probability that the new node j is connected to node i is proportional to the degree of node i:

$$P_i = \frac{k_i}{\sum_l k_l} \tag{2.1}$$

where k_i is the degree of node i, and the sum is made over all the existing nodes l. The degree distribution of the networks generated by the Barabási-Albert model follows a power law with a degree exponent of 3:

$$P(k) \sim k^{-3} \tag{2.2}$$

As a result, most nodes in the networks generated by the Barabási-Albert model have only a few links, and these nodes are held together by a few highly connected nodes.

The Barabási-Albert model is widely used in studies of social networks related to racial topics. In a study on how homophily influences the ranking of minorities in social networks, Karimi et al [25] propose a variant of the Barabási-Albert model that incorporates the homophily principle. In their model, each node has an attribute with a value of a or b, representing the ethnic minority and majority respectively. And homophily is defined as a parameter h in the range from 0 to 1. The value of homophily between two nodes is determined by their attributes. Same as the Barabási-Albert model, their model also begins with an initial network. At each time step, a newly arriving node j is connected to m pre-existing nodes. Unlike the Barabási-Albert model, the probability of node j to connect to i depends on both the homophily between them and the degree of i:

$$P_i = \frac{h_{ij}k_i}{\sum_l h_{ij}k_l} \tag{2.3}$$

Therefore, the formation of links in the model is controlled by the interplay between preferential attachment, via the degree of nodes, and homophily, via the node attributes. This model is a good example of incorporating homophily into other factors driving network structures. The addition of new nodes in the network, however, may not be suitable for school-based social networks, since students enroll and attend schools at the same time in each academic year.

Strategic network models make use of concepts from game theory. They assume that the configuration of a network is the equilibrium of a strategic game. In strategic network models, nodes are usually called players, and links are formed due to the strategic behaviors of players rather than based on probabilities. Players form links that will bring them benefits more than costs, otherwise they cut off links. Utility functions are defined to calculate the net benefits players can receive from networks. The definition of network stability varies for different network formation procedures. For example, Nash equilibrium is commonly used in non-cooperative games. A network is Nash stable if no one in the network has an incentive to change its links. The concept of pairwise stability is applicable to social networks in which forming a link between two players requires mutual consent, while severing a link only involves the consent of one player. A network is pairwise stable if no player has an incentive to delete a link, and no pair of players have an incentive to form a new link. However, one challenge in strategic network models is the presence of multiple equilibria.

2.3.1 Angelo Mele's model

Angelo Mele [37] combines components from both strategic and random network models and proposes a network formation model that converges to a unique stationary equilibrium. The model implemented in this thesis is based on his model, which is modified by incorporating homophily and reconstructing the components of the utility function.

In Angelo Mele's model there are n players, each player i is identified by a vector of A characteristics $\boldsymbol{X}_i = \{X_i^1, X_i^2, ..., X_i^A\}$. These characteristics can be race, gender, age, etc. Matrix $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n\}$ is a collection of vectors of the characteristics for the population. The social network is represented by a $n \times n$ binary matrix $\boldsymbol{g}$ with entry $g_{ij} = 1$ if player i considers j as a friend, and $g_{ij} = 0$ otherwise. Note that network $\boldsymbol{g}$ is directed, which means links are not necessarily mutual. And the diagonal entries g_{ii} are set to 0 by convention. The network configuration at time t is denoted as $\boldsymbol{g}^t$, and the network including all the current links except for g_{ij}^t is denoted by g_{-ij}^t.

Players receive utility from their connections in the network, directly or indirectly. And they always aim to increase the utility they receive from the network. The total utility players receive from all of their connections is given by a utility function consisting of four components:

$$\begin{aligned} U_i(\boldsymbol{g}, \boldsymbol{X}; \boldsymbol{\theta}) = & \underbrace{\sum_{j=1}^{n} g_{ij} u_{ij}(\theta_u)}_{direct\ friends} + \underbrace{\sum_{j=1}^{n} g_{ij} g_{ji} m_{ij}(\theta_m)}_{mutual\ friends} \\ & + \underbrace{\sum_{j=1}^{n} g_{ij} \sum_{\substack{k=1 \\ k \neq i,j}}^{n} g_{jk} v_{ik}(\theta_v)}_{friends\ of\ friends} + \underbrace{\sum_{j=1}^{n} g_{ij} \sum_{\substack{k=1 \\ k \neq i,j}}^{n} g_{ki} w_{kj}(\theta_w)}_{popularity} \end{aligned} \tag{2.4}$$

where $u_{ij}(\theta_u) = u(\boldsymbol{X}_i, \boldsymbol{X}_j; \theta_u)$, $m_{ij}(\theta_m) = m(\boldsymbol{X}_i, \boldsymbol{X}_j; \theta_m)$, $v_{ij}(\theta_v) = v(\boldsymbol{X}_i, \boldsymbol{X}_j; \theta_v)$, and $w_{ij}(\theta_w) = w(\boldsymbol{X}_i, \boldsymbol{X}_j; \theta_w)$ are functions of characteristics. And $\boldsymbol{\theta}$ is a vector of parameters with entries θ_u, θ_m, θ_v, and θ_w. In Mele's model, these functions have no explicit forms, which allows us to define our own functions which are based on the homophily principle. Besides, it is possible to include other components into the utility function.

The utility provided by each link comes from four sources. When player i connects to player j, he receives utility $u_{ij}(\theta_u)$ from this direct link. Mele assumes that the direct utility is a net value including both benefits and costs. Players receive negative direct utility if the benefit received from the link is less than the cost paid to maintain it. Player i will receive additional utility from the connection between i and j is reciprocate. In addition, once player i connects to player j, he is indirectly connected to the friends of player j, which will give him indirect utility. Finally, connecting to player j has effects on i's popularity, because i's friends are indirectly connected to j through the link.

Two random processes are included in Mele's model. The first one is random meetings: at each time step, two players i and j are randomly selected, and player i will revise its connection to j based on the utility function. And another one is random noises: players' preferences are subject to an idiosyncratic noise. Mele has proven that under two assumptions, the network generated by the model will converge to a a unique stationary equilibrium. The assumptions are as follows:

- Any player in the population has a positive probability of meeting any other players.
- The noise follows a *Type I extreme value* distribution, i.i.d. among links and across time.

In the next chapter, we will modify Mele's model, and the modified model will still satisfy the assumptions for convergence.

2.4 Measuring segregation in social networks

Various measures have been proposed for measuring the degree of segregation with respect to a certain attribute in social networks [6]. Because in this thesis networks are directed, we only focus on segregation measures applicable for directed networks. Assume the attribute we are interested in has K different groups. Let d_{khy} denote the number of dyads between nodes of group k and nodes of group h, $y = 1$ if dyads are connected and $y = 0$ otherwise.

The E-I (external - internal) index is proposed by Krackhardt and Stern [28]. It measures segregation by the difference in the number of links within groups and between groups. Specifically, it takes the number of out-group links, subtracts the number of in-group links, and divides by the total number of links (see Equation 2.5).

$$EI = \frac{\sum_{k=1}^{K} \sum_{h \neq k} d_{kg1} - \sum_{k=1}^{K} d_{kk1}}{\sum_{k=1}^{K} \sum_{h=1}^{K} d_{kh1}} \tag{2.5}$$

By this definition, the E-I index is 1 when all links in the network are out-group, and it is -1 when all links are in-group. When the index is equal to 0, the number of in-group links is equal to the number of out-group links. However, it does not mean that the network is not segregated, because the E-I index ignores the difference in the opportunities for forming in-group links and out-group links.

The odds ratio for within-group ties (ORWG), by contrast, takes the opportunities to form in/out-group links into consideration. It takes the odds ratio for link existence versus non-existence for in-group dyads and out-group dyads (see Equation 2.6) [6]. Simply put, the ORWG measures how much more likely two nodes will be connected if they are from the same group.

$$ORWG = \frac{\sum_{k=1}^{K} d_{kk1} \Big/ \sum_{k=1}^{K} d_{kk0}}{\sum_{k=1}^{K} \sum_{h \neq k} d_{kh1} \Big/ \sum_{k=1}^{K} \sum_{h \neq k} d_{kh0}} \tag{2.6}$$

When the ORWG is 1, in-group and out-group links are equally likely to form in the network, which means the network is not segregated. If the ORWG is larger than 1, in-group links are more likely to form than out-group links, which implies that the network is segregated with respect to the attribute. And the larger the ORWG, the higher level of segregation the network has. The maximum of the ORWG is positive infinity. Conversely, if out-group links are more likely to form than in-group links, the ORWG is less than 1, which means the network is integrated. And the more likely it is to form out-group links, the closer the ORWG approaches 0. One property of the ORWG is that it is sensitive to isolated nodes. This is because the addition of isolated nodes into the network will change the opportunities for forming links. Therefore, this measure is applicable to networks in which information on isolated nodes is also taken into account.

The concept of the assortativity coefficient is proposed by Newman [**Newman2003**].

It compares the fraction of links connecting nodes of the same group with the expected value if the links are randomly distributed in the network. Let f_{kh} denote the fraction of links connecting nodes of group k to nodes of group h. And f_{k+} denotes the fraction of links that come from nodes of group k, and f_{+k} denotes the fraction of links that end at nodes of group k:

$$f_{kh} = \frac{d_{kh1}}{d_{kh1} + d_{kh0}} \tag{2.7}$$

$$f_{k+} = \sum_h f_{kh} \quad \text{and} \quad f_{+k} = \sum_h f_{hk} \tag{2.8}$$

The assortativity coefficient is then defined as:

$$Assortativity = \frac{\sum_{k=1}^{K} f_{kk} - \sum_{k=1}^{K} f_{k+} f_{+k}}{1 - \sum_{k=1}^{K} f_{k+} f_{+k}} \tag{2.9}$$

The index is normalized such that $Assortativity = 1$ when there is perfect assortative mixing in a network: all links only connect nodes of the same group. $Assortativity = 0$ when there is no assortative mixing: the number of links connecting two nodes of the same group is equal to the expected value when links are randomly assigned. For a network that is perfectly disassortative, the index reaches its minimum value:

$$\min(Assortativity) = -\frac{\sum_{k=1}^{K} f_{k+} f_{+k}}{1 - \sum_{k=1}^{K} f_{k+} f_{+k}} \tag{2.10}$$

The value of the minimum varies in the range of $[-1, 0)$, depending on the relative number of links in each group.

Using the assortativity coefficient has several advantages in our case. First of all it clearly specifies the conditions under which the index would return value 0, identifying no segregation. On top of that it is insensitive to isolated agents added to the network. This property ensures that the measure is robust while handling incomplete real-life network data, which we will use for model evaluation. Therefore in this thesis we will apply the assortativity coefficient to measure segregation in students' friendship networks.

Chapter 3

Methods

3.1 An agent-based model for social network formation

In this section an agent-based model for social network formation is described. The model aims to generate social networks that resemble those found in schools given their student population. The model is based on the framework proposed by Angelo Mele[37], which converges to a unique equilibrium under certain assumptions. On top of his model, we incorporate a homophily function into the utility function so that students have preferences for similar friends. And the concept of social distance is used to measure the similarity between two students: the shorter the social distance between them, the more similar they are. In addition, we modify and give weights to the components of the utility function so that we can compare the contribution of each utility source to the total utility. Mele has proven that the model will converge to a unique stationary equilibrium under two conditions: nonzero meeting probabilities and agents' decisions are subject to noise that follows a *Type I extreme value* distribution. The two conditions are still satisfied in the modified model. And in our simulations, we do observe the convergence behavior of the model.

3.1.1 Agents' attributes and preferences

In Angelo Mele's model, a social network forms as a result of a game in which players make connections that are beneficial to them. These players are called agents in the agent-based model. For modeling a school with n students, we have n agents, and each agent represents a student. Agent i is identified by a vector of A attributes $\boldsymbol{X}_i = \{X_i^1, X_i^2, ..., X_i^A\}$. These attributes can be race, gender, age, or other socioeconomic characteristics. The Matrix $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n\}$ is a collection of agents' individual attributes. Agents form friendships with others, which are represented by the links in the social network. Let $\boldsymbol{g}$ denote a $n \times n$ adjacency matrix, with entry $g_{ij} = 1$ if agent i considers j as a friend, and $g_{ij} = 0$ otherwise. We assume that network $\boldsymbol{g}$ is directed, which means friendships are not necessarily mutual. And the diagonal entries of the adjacency matrix g_{ii} are set to 0. The network configuration at time t is denoted as $\boldsymbol{g}^t$, and the network including all the current links except for g_{ij}^t is denoted by g_{-ij}^t. The general notations used in the model are listed in Table 3.1.

The links in the social network are defined to contribute a utility to agents. Through these links, agents receive benefits such as information or help from others to whom they are directly or indirectly connected, and maintaining these links also

TABLE 3.1: General notations

t	$\triangleq$	Time
i	$\triangleq$	Index of an agent
$\boldsymbol{X}$	$\triangleq$	Attribute matrix of the population
$\boldsymbol{X}_i$	$\triangleq$	Attribute vector of student i
X_i^a	$\triangleq$	The attribute a of agent i
$\boldsymbol{g}^t$	$\triangleq$	Adjacency matrix at time t
g_{ij}^t	$\triangleq$	The connection between i and j at time t
$\boldsymbol{g}_{-ij}^t$	$\triangleq$	Adjacency matrix at time t excluding g_{ij}^t

requires a cost, such as time and energy. We assume that agents form a link according to the net utility it would provide, which consists of four components: direct utility from the relationship, additional utility if the link is mutual, utility from being indirectly linked to friends of a friend, and the cost of maintaining the link. The net utility one receives from a link is the sum of the three utility components minus the cost. An example is demonstrated in Figure 3.1.

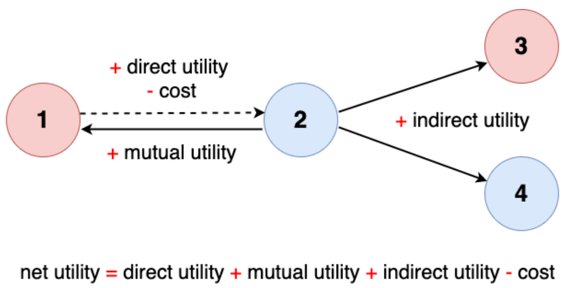


FIGURE 3.1: An example friendship network to demonstrate the components of the utility function. In this example, agent 1 is considering agent 2 as a friend based on the utility function. If agent 1 connects to agent 2, it will receive utility from three sources: the direct link, the mutual link, and the indirect links. Meanwhile, it will pay a cost to maintain the link. Utility from the indirect links is the average utility from all friends of agent 2: agent 3 and 4.

The total net utility U of agent i from a network $\boldsymbol{g}$ and population attributes $\boldsymbol{X}$ is the sum of the net utilities i receives from all links it has:

$$U_i(\boldsymbol{g}, \boldsymbol{X}) = \underbrace{\alpha \sum_{j=1}^{n} g_{ij} h_{ij}}_{direct friends} + \underbrace{\beta \sum_{j=1}^{n} g_{ij} g_{ji} h_{ij}}_{mutual friends} + \underbrace{\gamma \sum_{j=1}^{n} g_{ij} \frac{\sum_{k=1, k\neq i,j}^{n} g_{jk} h_{ik}}{\sum_{k=1, k\neq i,j}^{n} g_{jk}}}_{indirect friends} - \underbrace{c \sum_{j=1}^{n} g_{ij}}_{cost} \tag{3.1}$$

where α, β, γ and c are weights for each component, and h_{ij} is determined by a homophily function (see Equation 3.2). By assuming that $\alpha + \beta + \gamma = 1$, we normalize the benefit agents receive from their links to other agents.

There are several differences between the utility function used in Angelo Mele's model and the one used in our model. Firstly, we give each component a weight so that we can quantify the relative importance of each utility source. This will also provide insight into the important factors that drive students' friendship choices.

Secondly, there is no explicit component of friendship cost in Angelo Mele's utility function. Angelo Mele assumes that the direct utility is a payoff including both benefit and cost. In contrast, we separate benefit from cost so that we can investigate the influence of benefit and cost on students' friendship choices individually. Mele assumes that the indirect utility is the sum of the utilities received from all indirect links, whereas we assume that the indirect utility is the average of the sum. Under Mele's assumption, an agent i is more likely to connect to other agents who have more friends similar to i, even though these agents might also have many friends dissimilar to i. By taking the average of the utilities from all indirect links, we assume that it's not the absolute number of indirect similar friends, but the relative number of indirect similar friends, that is important to the indirect utility. Therefore, agent i will receive more indirect utility if it connects to other agents who have a higher proportion of friends similar to i.

In Mele's model, the direct, mutual and indirect benefits agent i receives by connecting to agent j are determined by function u_{ij}, m_{ij}, and v_{ij} respectively. Mele applies the homophily principle by relating these functions to the similarity between the attributes of agent i and j. However, these function have no explicit forms, which allows us to define our own homophily functions. We assume that homophily is a decreasing sigmoidal function of the similarity between two agents. And we define the similarity between two agents as the distance between them in social space, which is inspired by the work of Talaga et al [7]. Specifically, the more similar two agents are, the smaller the social distance is between them. The homophily between agent i and agent j, denoted as h_{ij}, takes the form of:

$$h_{ij} = \frac{1}{1 + (p^{-1}d_{ij})^q} \tag{3.2}$$

where $d_{ij} = d(\boldsymbol{X}_i, \boldsymbol{X}_j)$ is the social distance between i and j (see Equation 3.5), p and q determine the shift and slope of the homophily function respectively.

The homophily function is demonstrated in Figure 3.2. It shows that the homophily between i and j is 1 when their social distance d_{ij} is 0, and the homophily decreases to almost 0 as the social distance increases to 1. Also, we can see that parameters p and q control the sigmoidal transition between high and low homophily. p is the characteristic distance where homophily is 0.5. The larger the p is, the more tolerant agents are to social distance. q decides the slope of the transition. With a larger q, agents have a less binary homophily.

The social distance d_{ij} between agent i and agent j is defined as a function of the dissimilarity D in terms of their attributes $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$. And we give each attribute a a weight w_a to quantify the relative importance of the attribute to the social distance. As such, d_{ij} is defined as the weighted sum of $D(\boldsymbol{X}_i, \boldsymbol{X}_j)$ across all attributes A:

$$d_{ij} = \sum_{a=1}^{A} w_a D(\boldsymbol{X}_i^a, \boldsymbol{X}_j^a) \tag{3.3}$$

where

$$\sum_{a=1}^{A} w_a = 1 \tag{3.4}$$

For categorical attributes such as race, it is difficult to measure the dissimilarity between different categorical groups. Therefore, we assume different categorical

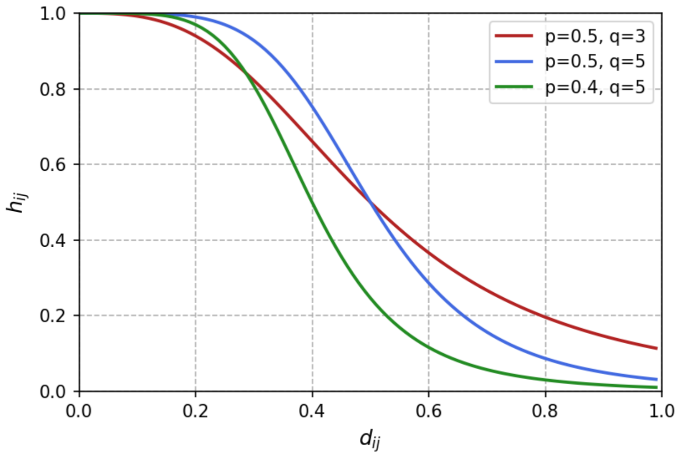


FIGURE 3.2: Homophily h_{ij} as a function of social distance d_{ij}.

groups have equally dissimilarity for simplicity. A binary function is applied to quantify the dissimilarity between agents' attributes. The dissimilarity between agent i and j with respect to attribute a is 1 if they are from different categorical groups. Otherwise, it is 0:

$$D(X_i^a, X_j^a) = \begin{cases} 0 & \text{if } X_i^a = X_j^a \\ 1 & \text{if } X_i^a \neq X_j^a \end{cases} \tag{3.5}$$

From above equations, we can see that when an attribute a has a larger weight w_a, the dissimilarity with respect to attribute a contributes more to the social distance. This leads to agents have less homophily for other agents with a different value of attribute a. Therefore, the weights of attributes are parameters affecting agents' homophily with respect to each attribute. The parameters used in the model are listed in Table 3.2.

TABLE 3.2: Model parameters and description

Parameters	Description
α	Weight for utility from direct links
β	Weight for utility from mutual links
γ	Weight for utility from indirect links
c	Cost of maintaining a link
p	The shift parameter of the homophily function
q	The slope parameter of the homophily function
w_a	Weight for attribute a
σ	The scale parameter of the distribution of noise

3.1.2 Process overview

In the process of network formation, agents randomly meet other agents. They are assumed to have complete information about the current network configuration and individual attributes of all agents. Agents make decisions based on the utility maximization principles. However, their preferences are subject to noise. The model still meets the conditions for convergence proposed by Angelo Mele so that generated networks can converge to a stationary equilibrium. The pseudo code overview of the model is shown in Algorithm 1.

Meeting process

At each time step a pair of agents $\{i, j\}$ are randomly selected from the population. Agent i can revise its connection to agent j in order to improve the utility it receives from the network. We assume that any pair of agents can meet, and the probability of any meeting is unbiased. Let $m^t = \{i, j\}$ represent the meeting m at time step t between agents i and j. Therefore the probability of $m^t = \{i, j\}$ is:

$$Pr(m^t = \{i, j\}) = \frac{1}{n(n-1)} \tag{3.6}$$

The assumption of unbiased meeting is not entirely realistic because similar people usually meet more often. However, unbiased meetings allows us to isolate the effects of students' homophily on their friendship choices. In future work, we plan to explore biased meetings where similar people have a higher chance to meet each other.

Utility maximization

In a meeting $m^t = \{i, j\}$, agent i revises the link g_{ij} to maximize its utility, while taking into account the current network configuration. Although agents make every decision based on the utility maximization principle, their preferences are subject to noise. The noise term is used to model unobservable events in real life that can influence agents' preferences, such as agents' moods [11]. Therefore, agents' friendship choices are not deterministic under noise.

Agents receive idiosyncratic noise denoted by ϵ. We assume that ϵ follows a *Type I extreme value* distribution and it is independent and identically distributed among links and across time [51]. We assume that the location parameter of the *Type I extreme value* distribution is 0, and the scale parameter of the distribution is a variable denoted by σ:

$$\epsilon \overset{i.i.d}{\sim} F(0, \sigma) \tag{3.7}$$

Agent i adds the link to j if and only if:

$$U_i(g^t_{ij} = 1, g^{t-1}_{-ij}, \boldsymbol{X}) + \epsilon^t_1 > U_i(g^t_{ij} = 0, g^{t-1}_{-ij}, \boldsymbol{X}) + \epsilon^t_0 \tag{3.8}$$

and i deletes the link to j if and only if:

$$U_i(g^t_{ij} = 1, g^{t-1}_{-ij}, \boldsymbol{X}) + \epsilon^t_1 < U_i(g^t_{ij} = 0, g^{t-1}_{-ij}, \boldsymbol{X}) + \epsilon^t_0 \tag{3.9}$$

When the equality holds, g^t_{ij} stays unchanged.

Convergence

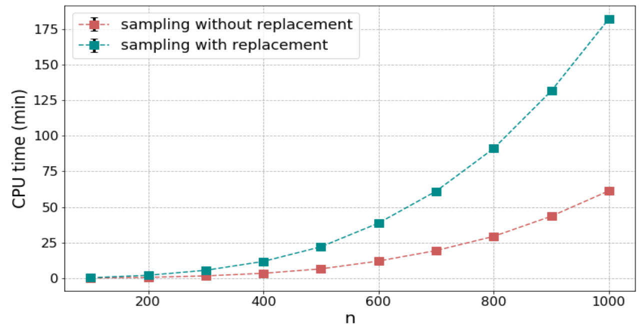
As mentioned before, Angelo Mele has proven that the model will converge to a unique stationary equilibrium under two conditions. The first condition is met by assuming that the probability of any meeting is nonzero. The second condition is satisfied by assuming that the idiosyncratic noise follows a *Type I extreme value* distribution, and it is independent and identically distributed among links and across time. Therefore, according to Mele's theory, the model will converge to a unique stationary equilibrium in the long run. In simulations, we do observe the convergence behavior of the model. And we also find that the scale parameter of noise has an influence on the stationary equilibrium of the model (see more details in Appendix A).

We set two criteria for stopping the model: a maximum time step t_{max} and a minimum fraction of unchanged links f_{min}. The model will stop either when the time step t reaches t_{max} or when the fraction of unchanged links is greater than f_{min}. However, it is possible that the fraction of unchanged links in the network reaches f_{min} before the model converges to its equilibrium. Therefore, we also set a variable t_{min} to prevent the model from stopping before the time step reaches t_{min}.

3.1.3 Computational cost

One challenge in the implementation of the model is the computational cost for large networks. However, using an efficient sampling method can greatly reduce computational cost, especially for large networks. For the meeting process, we compared two methods to sample meeting pairs: the first method (sampling with replacement) is to randomly draw two agents from the population; the second method (sampling without replacement) is to cycle through each possible meeting in a random order. Sampling without replacement is expected to be more efficient in our model as each meeting is guaranteed to be equally represented.

From Figure 3.3a, we see that the average CPU time for one simulation increases vastly as network size grows. And the CPU time by using sampling without replacement is shorter than using sampling with replacement: the speedup by sampling without replacement is more than three. Comparing the networks generated by the two sampling methods in Figure 3.3b, we find that the two sampling methods generate networks with the same total utility at convergence. This means that the two sampling methods lead to the same equilibrium. Therefore, sampling without replacement is effective and more efficient compared to sampling with replacement. It is applied to the model throughout the thesis.



(a)

(b)

FIGURE 3.3: Comparison between two sampling methods. (a) shows the CPU time required to reach convergence with respect to network size. The CPU used is an Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz, and the CPU time is obtained by using 20 hyper-threaded cores. (b) shows the total utility of networks generated by the two sampling methods at convergence. This figure is plotted using 100 agents and 100 repetitions.

Algorithm 1: Process overview of the model

```
Result: A friendship network
g^0 = a sparse initial random network
m = all n(n − 1) possible meeting pairs
for t = 1 to t_max do
    t_cycle = t mod |m|
    if t_cycle = 0 then
        if counter/|m| ≥ f_min and t ≥ t_min then
            terminate
        else
            shuffle(m)
            counter = 0
        end
    end
    {i, j} = m^{t_cycle}
    g_ij^initial = g_ij^t
    if U_i(g_ij^t = 1, g_{-ij}^{t-1}, X) + ε_1^t > U_i(g_ij^t = 0, g_{-ij}^{t-1}, X) + ε_0^t then
        g_ij^t = 1
    end
    if U_i(g_ij^t = 1, g_{-ij}^{t-1}, X) + ε_1^t < U_i(g_ij^t = 0, g_{-ij}^{t-1}, X) + ε_0^t then
        g_ij^t = 0
    end
    if g_ij^initial = g_ij^t then
        counter += 1
    end
end
```

Chapter 4

Sensitivity Analysis

Sensitivity analysis (SA) is a commonly used tool to quantify the sensitivity of the model output to the input parameters. It can also help in exploring the complex behavior of agent-based models. In this section, both global and local sensitivity analyses are performed. We use two methods (Shapley effects and Sobol indices) for global sensitivity analysis, and apply the one-factor-at-a-time (OFAT) method for local sensitivity analysis. The Shapley effects and Sobol indices of the model parameters are calculated first to give us an insight into the relative importance of each parameter to the model output. And then the OFAT method is applied to further explore the effects of each influential parameter on the model behavior.

To perform sensitivity analysis, a hypothetical population of agents is generated. Each agent is identified by three attributes, denoted by a_1, a_2, and a_3. These attributes can be any ethnic or socioeconomic characteristics such as race, gender, and the parental education. We assume that each attribute has two different categorical groups (e.g., male/female, high/low educated). The model developed in Chapter 3 is implemented to generate friendship networks. And the degree of segregation with respect to each attribute in the network is measured by the assortativity coefficient described in Chapter 2.

4.1 Global sensitivity analysis

For the global sensitivity analysis, the inputs of the model are the parameters listed in Table 4.1, and the outputs of the model are the assortativity by each attribute. Having no prior knowledge of the distributions of the model parameters, we assume that they are drawn from uniform distributions. The range of c used in simulations is from 0.1 to 0.9 because outside of these bounds, it would be too easy or too hard for agents to make friends. We expect that more interesting and complex behavior occurs within this range. Likewise, parameter p and q are also drawn from a range where the model shows complex behavior. Also, the upper bound of σ can not be too large, otherwise, agents' preferences will be obfuscated by noise.

For weight parameters (α, β, γ, w_1, w_2, and w_3), the range of the distribution is from 0 to 1. However, these parameters are not independent, because we assume that

$$\alpha + \beta + \gamma = 1 \quad \text{and} \quad w_1 + w_2 + w_3 = 1$$

Because the Sobol method [44] requires all the model inputs to be independent, we can not apply it to the dependent parameters in the model. Therefore, we use i.i.d parameters α', β' and γ' as model inputs instead of α, β and γ. In the model, α, β

and γ are replaced by the following formulas:

$$\alpha = \frac{\alpha'}{\alpha' + \beta' + \gamma'} \qquad \beta = \frac{\beta'}{\alpha' + \beta' + \gamma'} \qquad \gamma = \frac{\gamma'}{\alpha' + \beta' + \gamma'}$$

In this way, the sum of the three weights is still 1, while parameters α', β', and γ' are independent. The same replacement is also applied to w_1, w_2 and w_3.

On the contrary, the Shapley effects allow dependent model inputs. Therefore, the dependent parameters (α, β, γ, w_1, w_2, and w_3) are used for calculating the Shapley effects. And these parameters are drawn from Dirichlet distributions. Table 4.1 lists the model inputs and the their distributions used for the Sobol indices and the Shapley effects respectively.

TABLE 4.1: Model inputs and their distributions for global SA

Parameters	Description	Distribution for Shapley effects	Distribution for Sobol indices
α β γ	Weights for utility from direct, mutual and indirect links, $\alpha + \beta + \gamma = 1$	Dir($\boldsymbol{\theta}$), $\boldsymbol{\theta}$ = (1, 1, 1)	
w_1 w_2 w_3	Weights for attribute a_1, a_2 and a_3, $w_1 + w_2 + w_3 = 1$	Dir($\boldsymbol{\theta}$), $\boldsymbol{\theta}$ = (1, 1, 1)	
α'	Weight for utility from direct links		U(0, 1)
β'	Weight for utility from mutual links		U(0, 1)
γ'	Weight for utility from indirect links		U(0, 1)
$w_1{}'$	Weights for attribute a_1		U(0, 1)
$w_2{}'$	Weights for attribute a_2		U(0, 1)
$w_3{}'$	Weights for attribute a_3		U(0, 1)
c	Cost of maintaining a link	U(0.1, 0.9)	U(0.1, 0.9)
p	The shift parameter of the homophily function	U(0.1, 0.9)	U(0.1, 0.9)
q	The slope parameter of the homophily function	U(1, 10)	U(1, 10)
σ	The scale parameter of the distribution of noise	U(0, 0.5)	U(0, 0.5)

4.1.1 Sobol indices

The Sobol method is a variance-based global sensitivity analysis. It decomposes the variance of the model output into fractions attributable to each model input. Let Y denote the model output, and K be the set of model inputs. The first-order index of input K_i is defined as:

$$S_i = \frac{Var(\mathbb{E}[Y|K_i])}{Var(Y)} \tag{4.1}$$

The first-order index measures the expected reduction in the output variance if K_i is fixed, and it tells us the direct effects of K_i on the variance of the output Y [44]. With a large value of S_i, the variation of the expected model output strongly relies on K_i, indicating that Y is sensitive to K_i.

Let $K_{\sim i}$ denote all inputs but K_i. Then the total-order index of input K_i is defined as:

$$T_i = 1 - \frac{Var(\mathbb{E}[Y|K_{\sim i}])}{Var(Y)} \tag{4.2}$$

The value of T_i is equal to the expected variance left when all parameters but K_i are fixed. It accounts for the total effects that K_i has on the variance of the model output, including both direct effects and all interaction effects [21]. Therefore, the total-order index of a model input is equal to its first-order index when the input has no interactions with other inputs.

The first-order and total-order Sobol indices of the model parameters are shown in Figure 4.1. For the assortativity by attribute a_1, the scale parameter of noise and the weight of a_1 show for both the first-order and total-order the highest Sobol indices. Next comes the shift parameter of the homophily function p and the cost of maintaining a friendship c. The first-order indices of w_2' and w_3' are almost zero, meaning that the weight of attributes a_2 and a_3 have almost no direct effects on the assortativity by a_1. However, they have a higher value of total indices, suggesting that the weight of attribute a_2 and a_3 can indirectly influence the assortativity by a_1. Besides, it is shown that α, β, γ, and q are not influential to the assortativity with respect to a_1, either directly or indirectly.

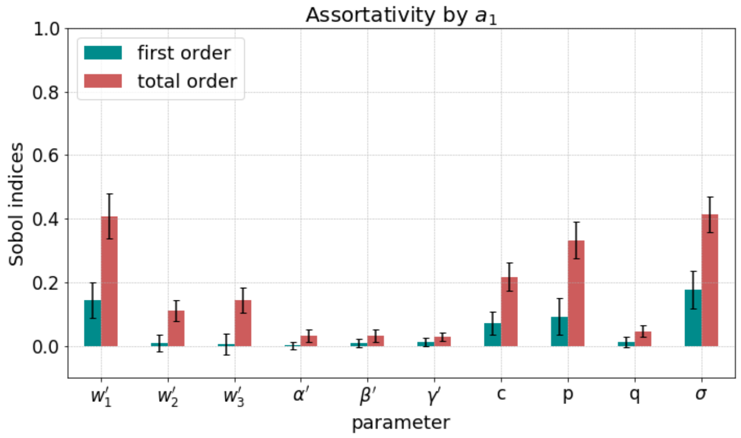


FIGURE 4.1: The first-order and total-order Sobol indices for each model parameter. The model output is the assortativity with respect to attribute a_1. The Sobol indices are calculated by 12,000 samples generated by Saltelli's sampling scheme [43]. The error bar indicates the 95% confidence interval for each parameter.

4.1.2 Shapley effects

Although replacing the dependent inputs allows us to use the Sobol method, we are still interested in the dependent parameters because they are the ones directly used in the model. Therefore, the Shapley effect is applied to attribute the variance of the model output to the model inputs including both dependent and independent parameters.

The Shapley effect [50] for sensitivity analysis is based on a solution concept used in game theory, called the Shapley value[47]. This concept defines a way to fairly distribute the total gains of a cooperative game to the contribution of each player. In its application to the sensitivity analysis, model inputs are regarded as players, and the total variance of the model output is regarded as the total gains of the game. The total variance of the model output is attributed to each model input by the Shapley effect. Specifically, the Shapley effect is the average expected marginal contribution of a model input to the variance of the model across all possible combinations. Let $K = \{1, 2, ..., k\}$ be the set of the model inputs, and S denote a subset of K. $v(S)$ measures the variance of the model output attributed by the uncertainty of the inputs in S. The Shapley effect of the i-th model input is calculated by:

$$Sh_i = \sum_{S \subseteq K \setminus i} \frac{S!(k - |S| - 1)!}{k!} \big(v(S \cup i) - v(S)\big) \tag{4.3}$$

For computing the Shapley effects of the model parameters, we implement the Monte Carlo algorithm proposed by Song et al [50], which greatly reduces the computational cost. The model inputs are drawn from their distributions listed in Table 4.1. Besides, a dummy parameter is introduced into the model. Since it has no actual functionality in the model, it can be used as a reference for the accuracy of the results. The dummy parameter is drawn from a uniform distribution ranging from 0 to 1.

The Shapley effects of the model parameters are shown in Figure 4.2. Similar to the results from the Sobol indices, the weight of attribute a_1 and the scale parameter of noise are the most influential factors to the assortativity by a_1. However, while the Shapley effects show that the weight of a_1 is more influential to the assortativity by a_1 than the scale of noise, the Sobol indices of them are almost the same. The shift parameter of the homophily function p is still the third most important contributor to the variance of the assortativity, followed by the cost c. Besides, the parameters α, β, γ and q show similar or less Shapley effects as the dummy parameter, therefore, their effects are not significant. In conclusion, the ranking of the parameters by their contribution to the variance of the assortativity, is for the most part the same between the Sobol indices and the Shapley effects.

The total-order Sobol indices of the independent inputs w'_1, w'_2 and w'_3 are similar to the Shapley effects of the dependent inputs w_1, w_2 and w_3. This indicates that even though w'_1, w'_2 and w'_3 are independently sampled, in the model they are still mapped to w_1, w_2 and w_3. Therefore, they operate in the same way on the model output.

4.2 Local sensitivity analysis

The global sensitivity analysis quantifies the contribution of each model parameter to the variance of the model output. However, it does not reveal the specific relationships between them. Therefore, the OFAT method is applied. The OFAT method varies one parameter at a time while keeping the others fixed, and it reveals the relationship between the varied parameter and the output, e.g. linear or nonlinear relationship. It can also show tipping points where the output changes

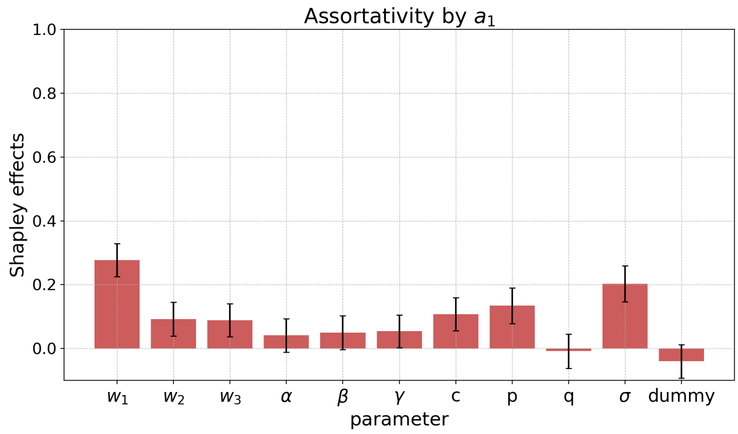


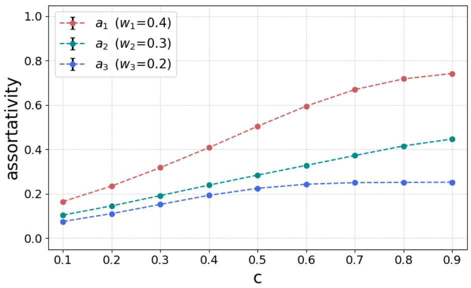
FIGURE 4.2: The Shapley effects of the model parameters. The model output is the assortativity with respect to attribute a_1. The variance of the assortativity is estimated using 20,000 samples. The error bar indicates the 95% confidence interval for each parameter.

drastically with respect to a small change in the varied parameter [10]. The parameters α, β, γ and q are not discussed in the OFAT method, becasue they show almost no effects on the assortativity in the global sensitivity analysis.

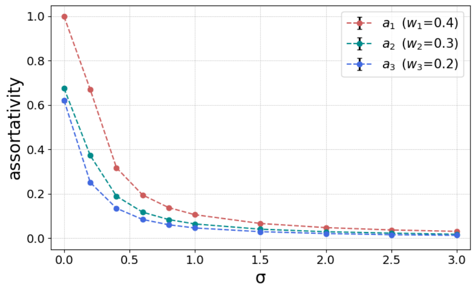
From Figure 4.3, we can see the effects of each parameter on the assortativity. Figure 4.3a shows that as the cost of maintaining a friendship c increases, the assortativity with respect to the attributes a_1, a_2 and a_3 also increases. In addition, the assortativity by an attribute is higher when the weight of the attribute is larger. This is more obvious when the cost of making friends is expensive. This suggests that when maintaining a friendship is expensive, agents are less likely to make friends that are not similar, especially with respect to the attributes that are important to social distance. In contrast to the cost, the scale parameter of noise σ is negatively correlated to the assortativity. This is shown in Figure 4.3b, where the assortativity decreases as the scale of noise grows. And the assortativity converges to zero when the noise scale becomes so large that it obfuscates the preferences of agents.

In Figure 4.3c, we see that the assortativity by an attribute becomes higher as the weight of the attribute increases. And when the weight of an attribute is zero, the assortativity by this attribute is also zero. As discussed in Chapter 3, the weights of attributes can represent agents' homophily with respect to each attribute. When the weight of an attribute is high, agents have a high homophily for the attribute, and they are less likely to make friends of the different categorical groups.

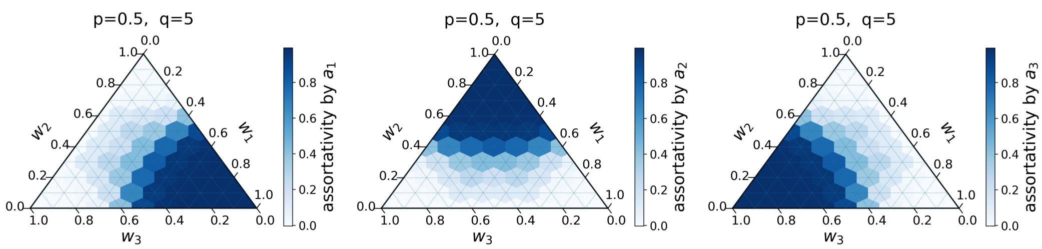
Finally, Figure 4.3d shows the relationship between the assortativity and the shift parameter of the homophily function: p. By increasing p, we see a general decrease in the assortativity. This is because p is a parameter that controls the tolerance of agents to social distance. The larger the p is, the more tolerant agents are to dissimilar friends, therefore the lower the assortativity is in the network.



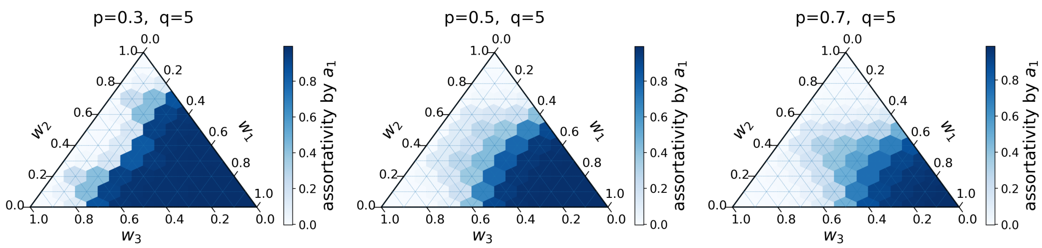
(a)



(b)



(c)



(d)

FIGURE 4.3: OFAT results for the influential model parameters. (a) shows the relationship between c and the assortativity with respect to three attributes with different weights; (b) shows the effects of σ on the assortativity; (c) shows how the weights of attributes impact the assortativity by each attribute; (d) shows the effects of the shift parameter of the homophily function p on the assortativity.

Chapter 5

Model Calibration and Evaluation

This section aims to evaluate the performance of the model on segregation estimation. First we calibrate the model so that the output of the model can fit the observed data in real life. After that, we evaluate the performance of the calibrated model by fitting it to a test set and measuring the goodness of fit. Due to the limitation of obtaining friendship network data from Dutch schools, we use the US school data instead. The validity of using US school data to the Dutch context is discussed in Chapter 7.

5.1 Data description

The data used for model evaluation is from Wave I of Add Health[19], a study of adolescents in junior high or high schools in the United States during the 1994-95 school year. $90,118$ students representing 145 schools in 84 communities responded to an in-school questionnaire, in which each student was asked to nominate up to five male and five female friends. The Add Health data also collected the individual characteristics of each student, including sex, race and grade.

We used 45 schools whose size is between 100 and 1000, which contain $25,758$ students in total. Table 5.1 shows the attributes of students and their categories collected in the data. Since there are relatively little unreported values for each attribute, we replace them with the median value of the attribute of the school for simplicity.

TABLE 5.1: Student attributes and their categories.

Attribute	Categories	Missing values (%)
Sex	1=male, 2=female	0.722
Race	1=white, 2=black, 3=Hispanic, 4=Asian, 5=mixed/other	1.168
Grade	7, 8, 9, 10, 11, 12	0.772

In Figure 5.1, we show the distribution of each attribute among those schools. We can see that in most schools, the gender ratio is around 1, and white students form a majority. However, there are some schools where black or Hispanic students form a majority instead. Some schools have no representation for certain grades.

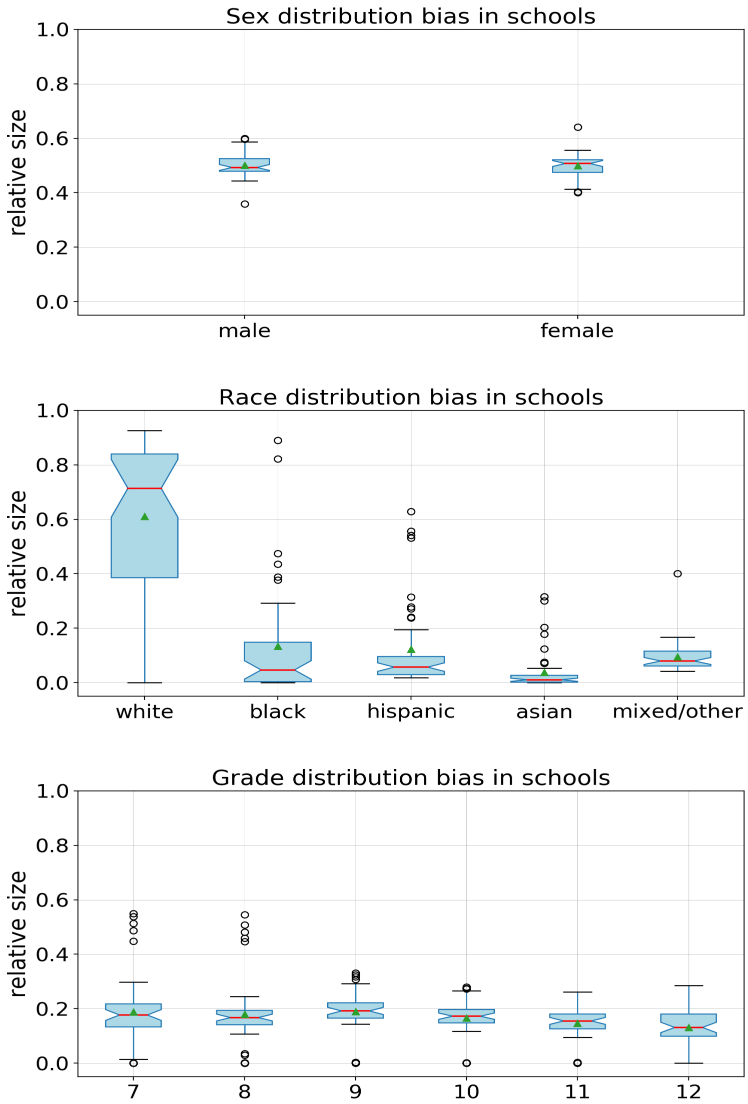


FIGURE 5.1: The distribution of each attribute in the selected 45 schools. The red line represents the median value and the green dot represents the mean value of the distribution.

Looking into the assortativity distribution across schools for the three attributes (Figure 5.2), we find that grade has the highest assortativity coefficient, and race has the highest variance in assortativity. For sex, because each student was asked to nominate up to 5 male and 5 female friends, the assortativity is expected to be lower than the real value. Which means, there might be a bias in sex assortativity.

5.2 Model calibration

Model calibration is carried out by tuning model parameters to obtain a best fit between simulated data and observed data. Searching for a parameter setting that gives the best fit requires an optimization algorithm. For which we choose Evolutionary Algorithms (EAs) because unlike traditional optimization approaches, EAs can solve problems with no explicit formulation between inputs and outputs. And they only require a fitness function to evaluate solutions. Furthermore, as EAs search

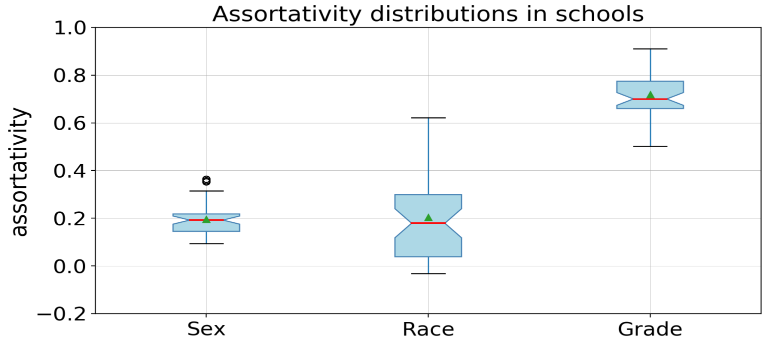


FIGURE 5.2: The distribution of assortativity by each attribute. The red line represents the median value and the green dot represents the mean value of the distribution.

the solution space in parallel, they are able to avoid being trapped by local optimal solutions [14]. Although EAs do not guarantee to find the globally optimal solution, a good solution is sufficient for our task. Apart from an optimization algorithm, a training set is needed for model calibration, and a test set is needed for assessing how well the calibrated model performs.

5.2.1 Differential evolution with global and local search

An efficient evolutionary framework called GL-DE [63] has been used for model calibration. It includes two modules, an exploration module that focuses on global search and an exploitation module that focuses on local search around the best-so-far solution. Both modules makes use of differential evolution (DE), a variant of EAs.

Similar to other EAs, the DE algorithm first generates an initial population with Z random candidate solutions $\boldsymbol{\theta}_i$ ($i = 1, 2, ..., Z$), and then evaluates each candidate with regard to a fitness measure. After that, it iteratively creates offspring by applying mutation and crossover. Parents are replaced by their offspring in the new generation if the fitness of the offspring is better. The uniqueness of DE is to use differential mutation. It creates new candidate solutions by combining existing solutions according to the following formula [14]:

$$\boldsymbol{\theta}' = \boldsymbol{\theta}_{r1} + F \cdot (\boldsymbol{\theta}_{r2} - \boldsymbol{\theta}_{r3}) \tag{5.1}$$

where the new mutant $\boldsymbol{\theta}'$ is produced by combining three randomly selected candidates: $\boldsymbol{\theta}_{r1}$, $\boldsymbol{\theta}_{r2}$ and $\boldsymbol{\theta}_{r3}$. F is a scalar controlling the rate of mutation.

In the GL-DE framework, the entire population is split into two sub-populations, a fraction f is for the exploration module and a fraction $1 - f$ is for the exploitation module. The exploration module uses a DE operator that creates new mutants according to the above formula. This DE operator is good at maintaining diversity within the population so that it can focus on searching for solutions from the global solution space. The DE operator used in the exploitation module creates new mutants based on the best-so-far solution $\boldsymbol{\theta}_{best}$:

$$\boldsymbol{\theta}' = \boldsymbol{\theta}_{best} + F \cdot (\boldsymbol{\theta}_{r1} - \boldsymbol{\theta}_{r2}) \tag{5.2}$$

Therefore it is good at local search and fine-tuning $\boldsymbol{\theta}_{best}$. The flowchart of the GL-DE framework is shown in Figure 5.3.

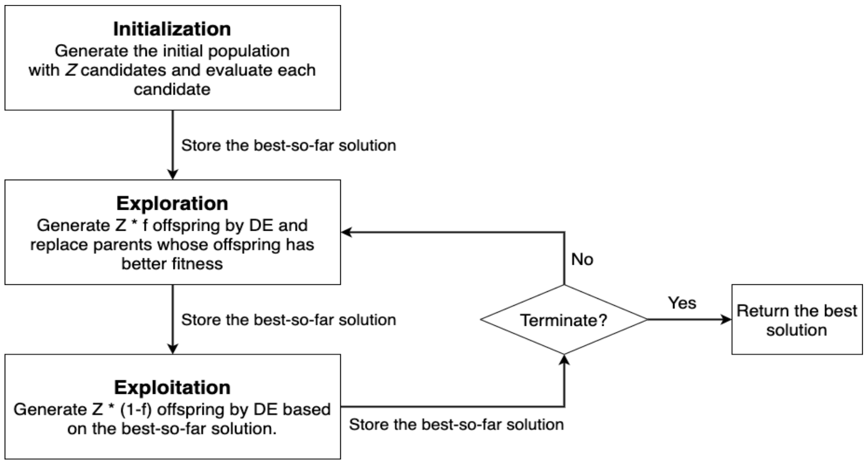


FIGURE 5.3: The flowchart of the GL-DE framework.

5.2.2 Implementation

To implement the framework, we need to define our problem, solution candidates and choose a fitness measure. Our problem is to find the best model parameter settings leading to the least difference between simulated assortativity and observed assortativity. Therefore, solution candidates in this problem are sets of model parameters in the parameter space. The fitness function should measure the overall error between simulated assortativity and observed assortativity across all schools in the training set. Having limited influence on the assortativity along attributes, the value of parameter α, β and γ can be fixed in order to reduce the search space.

The mean absolute error (MAE) can act as the fitness function. It measures the average of the absolute difference between simulated assortativity $\boldsymbol{M}_s$ and observed assortativity $\boldsymbol{O}_s$ across all schools s in the training set S.

$$fitness(\boldsymbol{\theta}) = \frac{\sum_{s \in S} ||\boldsymbol{M}_s(\boldsymbol{\theta}) - \boldsymbol{O}_s||}{|S|} \tag{5.3}$$

Since the assortativity by sex is probably biased through the data collection method, it might not reflect the real level of sex segregation in schools, therefore fitting this metric might introduce bias in the fit. So we only take race and grade into account, which means $\boldsymbol{M}_s$ and $\boldsymbol{O}_s$ are vectors with two entries: the assortativity by race and by grade. The difference between $\boldsymbol{M}_s$ and $\boldsymbol{O}_s$ is measured by Euclidean distance for simplicity.

Considering the high computational cost of model calibration, the training set consists of 20 schools, and the test set includes the other 25 schools. Schools are randomly allocated to the two sets.

5.3 Model performance

The performance of the calibrated model is assessed by the test set. Given the individual attributes of students in a school, we use the calibrated model to generate a simulated network. And then we compare the simulated network to the real network with regard to the assortativity by race and grade. The Wilcoxon signed-rank test is performed to determine whether the simulated data and the observed data are significantly different.

The null hypothesis of the Wilcoxon signed-rank test is that the distribution of the difference between observed data and simulated data is symmetric around zero. According to the results from the Wilcoxon signed rank test shown in Figure 5.4, the p-values for both attributes are larger than 0.05, which means we have no strong evidence against the null hypothesis. Which implies we can not reject the null hypothesis.

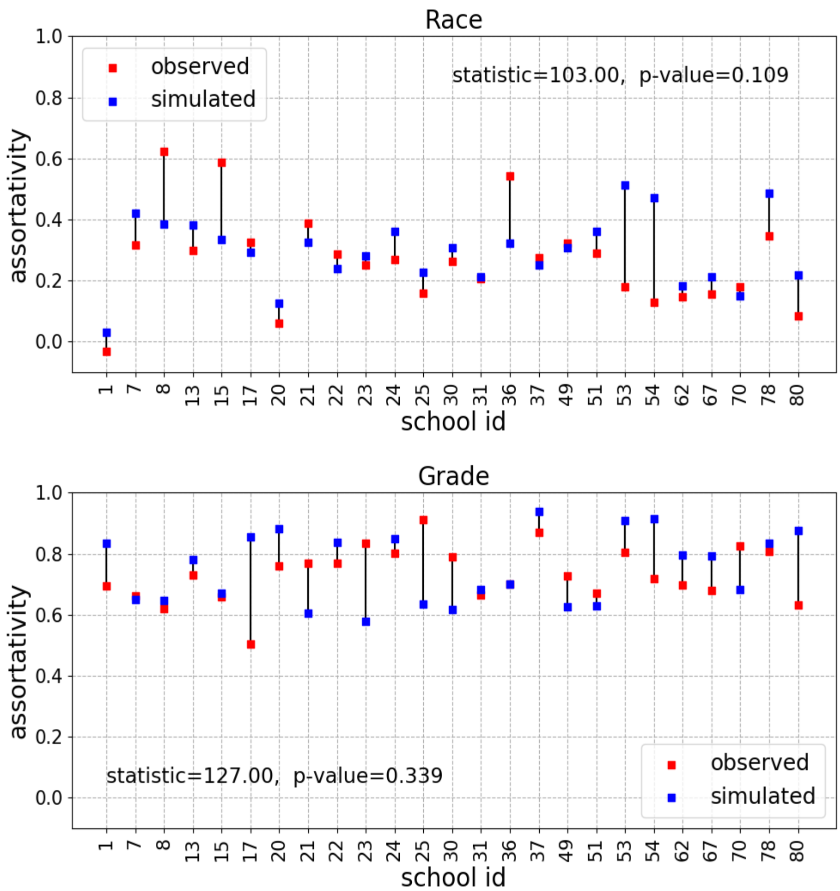


FIGURE 5.4: The performance of the calibrated model on the test set. The goodness of fit is measured by the Wilcoxon signed-rank test.

The mean absolute error (MAE) and the root mean square error (RMSE) shown in Figure 5.5 demonstrate that the simulated error in the assortativity by grade is slightly higher than the error in the assortativity by race. The MAE of the assortativity by race is 0.10, while the MAE of the assortativity by grade is 0.11. By comparing the MAE to the RMSE of the two attributes, we can see that there are not many large errors in the simulated assortativity.

In addition to the accuracy of the model, we are also interested in when the model over- or underestimates assortativity in schools. In Figure 5.4, we observe some schools with a high level of assortativity by race: these schools are school 8, 15 and 36. From the Add Health data, we find that all of these schools contain two

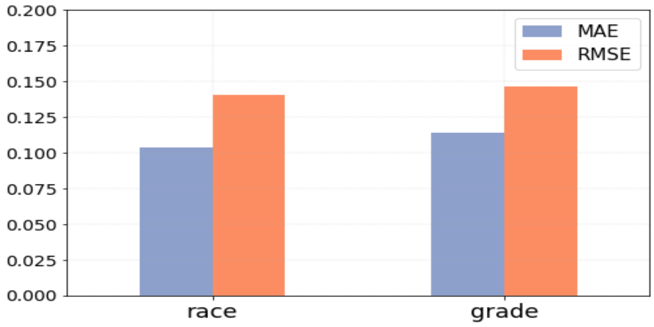


FIGURE 5.5: The mean absolute error (MAE) and the root mean square error (RMSE) of the simulated assortativity by race and grade.

dominant racial groups of similar size: white and black students. The model underestimates the assortativity by race in these schools. On the contrary, in schools with a low level of assortativity by race (school 1, 20 and 80), we find only one dominating racial group: white students. The assortativity regarding race in these schools is overestimated by the model. It seems that the model underestimates the racial segregation in some schools with two competitive racial groups. This would suggest that students' racial homophily becomes stronger when there is another racial group with a competitive size and the model underestimates racial segregation when there is only one large racial group. This might force minorities in these schools to make friends with the majority in order to avoid being isolated.

Another limitation of the model is found by comparing school 7 with 8. These two schools are similar in size, and they both have two dominant racial groups. However, the assortative mixing by race in the two schools is very different. We can explain the difference by looking at the racial compositions of these schools: in school 7, Hispanic students and mixed students are the two largest racial groups, while in school 8, white students and black students are the two largest groups. Since the model only considers whether two students come from the same racial group when it calculates social distance (see Equation 3.5), it ignores the distinction between different pairs of racial groups. In real life, the dissimilarity between Hispanic students and mixed students might be lower than that between white and black students. Likewise, in school 53 and 54 where the two biggest racial groups are white students and Hispanic students, the assortativity by race is lower than that in school 8. Above observations implies that students' racial homophily is probably related to the racial composition of schools. This relationship is not found in the assortativity regarding grade.

Chapter 6

Experiments and Results

In this section the model is applied to conduct experiments and answer the research questions. We set up a hypothetical school with 300 students and vary the racial composition of the school. We take race as an example, however, the model can be applied to other characteristics such as socioeconomic status as well. Our experiments start with two racial groups, and then with three groups. From Chapter 5, we know that schools with one dominating racial group usually have low assortativity with respect to race. However in these schools students also do not have many opportunities to contact others from different racial groups and to make intergroup friendships. Therefore instead of only looking into the assortativity with respect to race, we also take into account the fraction of interracial friendships out of all friendships. In order to show whether the experiment results replicate real-world patterns, we will compare them with the Add Health data.

6.1 The effects of school composition

6.1.1 School compositions with two racial groups

In the first experiment, we mix students of two racial groups: rg_1 and rg_2 in the hypothetical school. We want to explore how the racial composition of the school influences the assortativity by race as well as interracial friendships. On top of that, we also want to study the racial competition theory [57, 49], which says there is a correlation between students' racial homophily and the racial composition of schools. As the size of the racial minorities grows, racial majorities might feel threatened by minorities, leading to students having a higher racial homophily. However, students' racial homophily can be reduced if schools provide contacts under the right conditions [27, 52]. Therefore, both positive and negative correlations between students' racial homophily and school composition are possible. To study both cases, we correlate the weight of race w_r with the size of rg_2.

In Figure 6.1, we experiment with different correlations between w_r and the size of rg_2. When the weight of race is 0 (see Figure 6.1a), there is no assortativity with respect to race in the network because students have no preference for same-race friends. Also as the size of rg_2 increases, it provides more opportunities for students of rg_1 to make interracial friendships, therefore the fraction of interracial friendships in rg_1 increases. Meanwhile, students of rg_2 have more opportunities to make same-race friends, therefore the fraction of interracial friendships in rg_2 decreases. Because students have no racial homophily, when the sizes of two races are equal, the fraction of interracial friendships in both racial groups converges to 0.5.

By setting the weight of race to 0.5 in Figure 6.1b, students have preference for same-race friends. Therefore we see a significant increase in the assortativity by race, and relatively fewer interracial friendships compared to Figure 6.1a. In this case, increasing the size of rg_2 significantly intensifies the assortativity with respect race, and it only slightly promotes interracial friendships for rg_1.

Now a correlation between w_r and the size of rg_2 is introduced. When the correlation is positive, increasing the size of rg_2 will enhance students' racial homophily. This leads to a drastic increase in the assortativity with respect to race, as shown in Figure 6.1c. And when the size of rg_2 is about 0.25 the assortativity is the most sensitive to the change in the size of rg_2. Also the fraction of interracial friendships slightly increases because of more contact opportunities for rg_1, but then it drops back to zero because racial homophily starts to dominate students' friendship choice. This experiment demonstrates that if the racial competition theory holds in schools, increasing the size of racial minorities will aggravate racial assortativity in students' friendship networks. On the contrary, if the correlation between w_r and the size of rg_2 is negative (see Figure 6.1d), mixing students with different backgrounds will both reduce the assortativity and promote interracial friendships for both rg_1 and rg_2. This shows that it could be important to study the correlation between students' homophily and school composition before mixing students in schools.

6.1.2 School compositions with three racial groups

Now we mix three racial groups: rg_1, rg_2 and rg_3 in the hypothetical school. In this experiment, we assume that the weight of race has no correlation with the racial composition of the school. And the value of w_r is set to 0.5.

The borders in Figure 6.2 show compositions with two racial groups. It is clear that when there are two racial groups and their group sizes are equal, the assortativity by race reaches a maximum. The assortativity with respect to race is low when there is one racial group dominating the other groups, as shown in the corners of the plot. The fraction of interracial friendships is also low in this situation. With three racial groups that have similar sizes, the fraction of interracial friendships is the highest, however the assortativity with respect to race is relatively high as well. In this experiment, there is no such a composition leading to both a low assortativity with respect to race and a high fraction of interracial friendships. Therefore, having three racial groups similar in size seems to be the best compromise between a relatively low level of assortativity and a relatively high fraction of interracial friendships.

6.1.3 School compositions with a "bridge" group

In the previous experiment, the dissimilarity regarding race between any two different races is 1 (see Equation 3.5). In real life however, the dissimilarity between races that share similar languages and cultures is usually smaller than races that have very different languages and cultures. This leads to the next experiment, in which we introduce a racial group rg_3 that shares similarities with two other groups: rg_1 and rg_2. Specifically, the dissimilarity with respect to race between rg_3 and the other two races is 0.5, while the dissimilarity between rg_1 and rg_2 is still 1. The purpose of the experiment is to see whether rg_3 can act as a bridge, facilitating friendships across races.

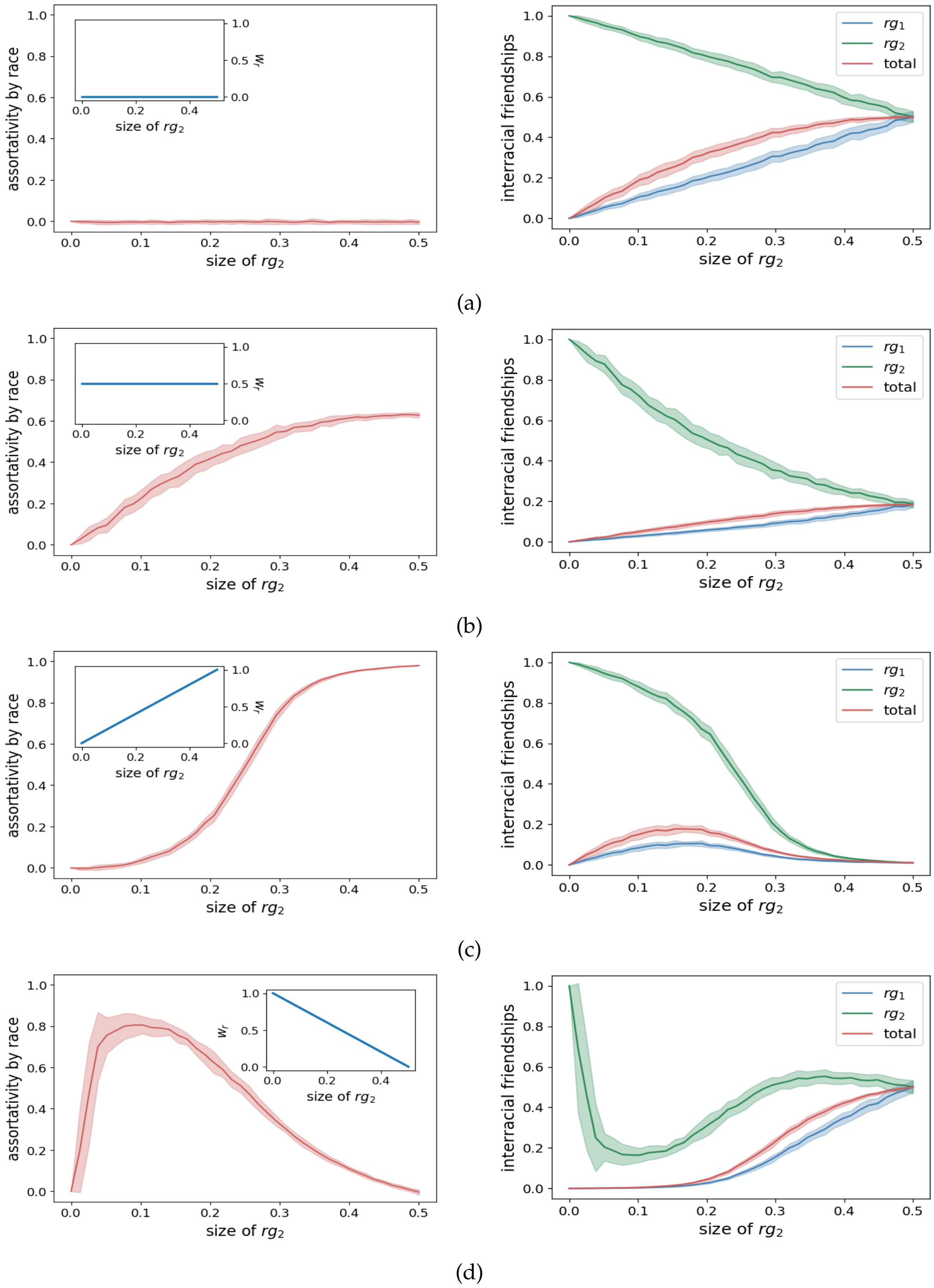


FIGURE 6.1: The assortativity by race and the fraction of interracial friendships with respect to the size of rg_2. The relationship between w_r and the size of rg_2 has no correlation, a positive correlation, and a negative correlation respectively.

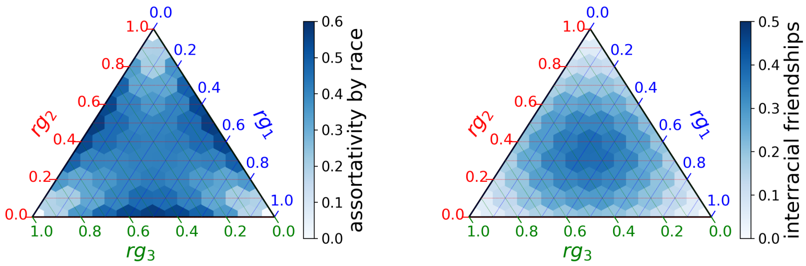


FIGURE 6.2: The assortativity by race and the fraction of interracial friendships with respect to the racial composition of the school. The three axes represent the size of rg_1, rg_2 and rg_3 respectively.

Comparing Figure 6.3 to Figure 6.2, we find that when the size of rg_3 is not zero, the assortativity with respect to race is generally lower, while the fraction of interracial friendships is generally higher. When the size of rg_3 is 0.4 and the other two groups have sizes of 0.3, the fraction of interracial friendships peaks, and the assortativity by race is still relatively low. This shows that introducing a bridge group into mixed schools can reduce the assortativity and promote interracial friendships.

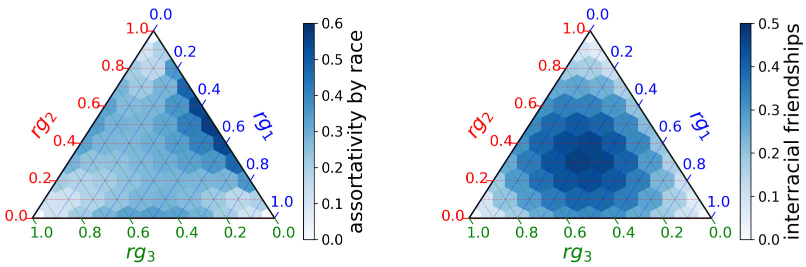
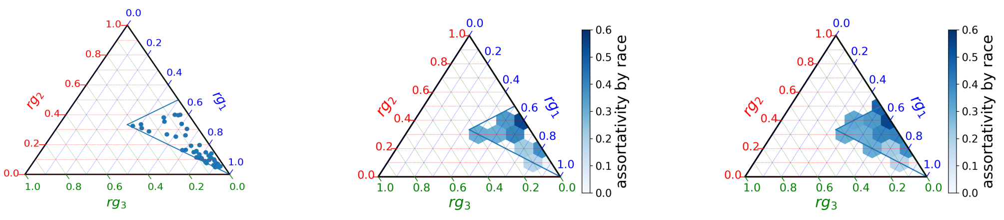


FIGURE 6.3: The assortativity by race and the fraction of interracial friendships with respect to the racial composition of the school. The three axes represent the size of rg_1, rg_2 and rg_3 respectively. The dissimilarity between rg_3 and the other two groups is 0.5, but it is 1 between rg_1 and rg_2.

6.2 Simulation vs. Add Health data

The above simulations reveal how school racial compositions influence the assortativity with respect to race and the fraction of interracial friendships in students' social networks. However, we are also interested in whether the simulations can replicate real-world patterns. The Add Health data introduced in Chapter 5 allows us to make a comparison between the simulated patterns and the real patterns observed in sampled schools. Among the 45 schools used in Chapter 5, all schools have three or more racial groups, which allows us to replicate Figure 6.2. For this reason the Add Health data has been processed to get the assortativity and the fraction of interracial friendships for each school. For schools containing more than three racial groups, we keep the three major racial groups and rescale their sizes so that we obtain a normalized composition of the three major racial groups of each school.

Because the three racial groups in each school are not always the same, it is not possible to define rg_1, rg_2 and rg_3 as three specific races while keeping enough



(a) (b) (c)

FIGURE 6.4: Data processing for the Add Health data. The three axes represent the size of each racial group. The blue lines indicate compositions where rg_1 is the largest group, rg_2 is the second largest, and rg_3 is the smallest group. (a) shows the compositions of the 45 schools in the data, with dots representing each school. (b) shows the assortativity by race for each composition. (c) shows the assortativity by race for the missing values obtained by interpolation as well.

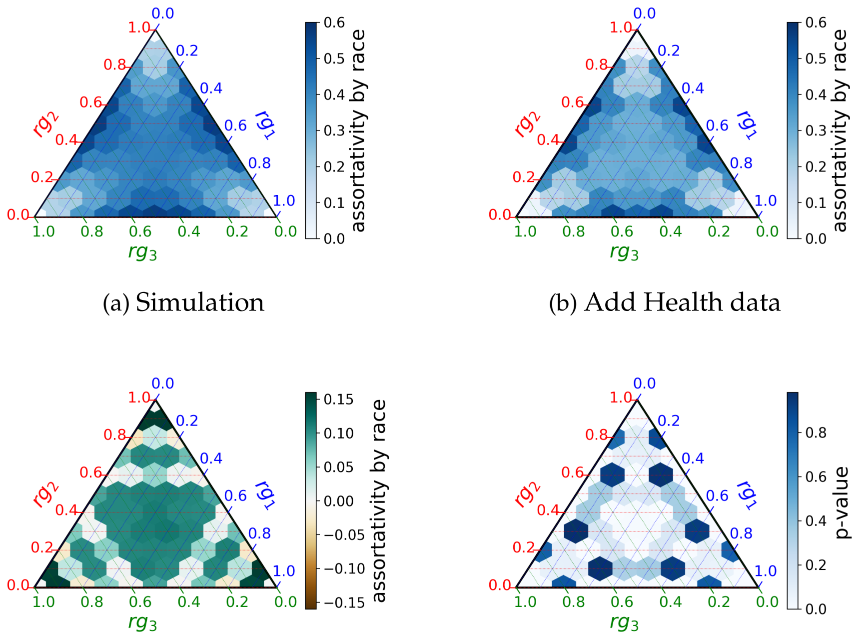
data points. Therefore we only keep the information of the racial composition of each school, where the largest racial group is rg_1, the second largest is rg_2 and the smallest of the three is rg_3. In Figure 6.4a, we show the racial compositions of the 45 schools in the Add Health data. Then we round each data point in Figure 6.4a to fit them to the nearest grid cell in the ternary plot. When multiple data points get rounded to the same grid cell, we take their average value of the assortativity, as shown in Figure 6.4b. However some compositions are still missing from the Add Health data. For those compositions we interpolate their values by taking the average value of their neighbors (see Figure 6.4c).

Now we have obtained the values of the assortativity for all possible compositions where rg_1 is the largest group, rg_2 is the second largest, and rg_3 is the smallest group. To create a complete picture such as Figure 6.2, we can copy the value of the assortativity of each composition to the compositions with the same combination of relative racial group sizes. At the end, we end up at the Figure 6.5, which is symmetric like the simulation. For the fraction of interracial friendships, we do the same processing (see Figure 6.6).

Comparing the simulation to the Add Health data in Figure 6.5, we find that it has a similar pattern in terms of the assortativity by race. In the simulation, the assortativity is high when schools have two racial groups of similar size, and it is low when one race dominates. This is also observed in the Add Health data. However, for the case where only one race dominates, the simulated assortativity is higher than the Add Health data, and the difference between them is around 0.15 (see Figure 6.5c). When schools have three comparable races, the simulation gives a higher assortativity than the Add Health data. For some racial compositions, the difference between the simulation and the Add Health data is approximately 0.

Figure 6.5d shows the p-value of the t-test with the null hypothesis that the means of the simulation and the Add Health data are equal. For compositions with a low p-value, the difference between the simulation and the Add Health data is more significant. However, because the sample size of the Add Health data is not large, the p-value might not be very reliable.

For the fraction of interracial friendships, the pattern of the simulation is less similar to that of the Add Health data (see Figure 6.6). However, they both reach their maximum when the three racial groups are similar in sizes, and reach their



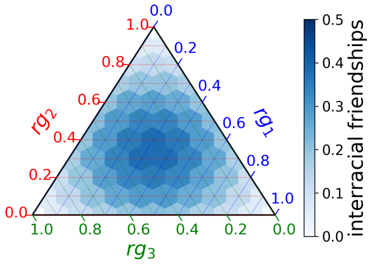
(a) Simulation

(b) Add Health data

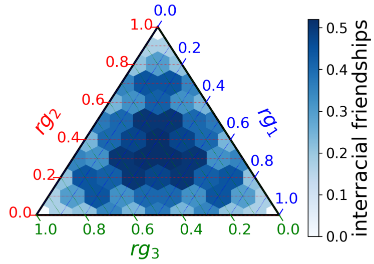
(c) Difference

(d) p-value

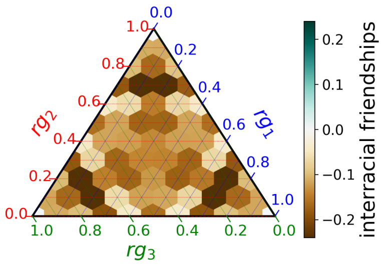
FIGURE 6.5: Comparison between the simulation and the Add Health data with respect to the assortativity by race. (a) shows the simulated assortativity generated by the model for each racial composition. (b) shows the average assortativity of schools with the same composition in the Add Health data after interpolation. (c) shows the absolute difference between (a) and (b). (d) shows the p-value of t-test for each composition, the p-value is 0 for interpolated values or when there is only 1 sampled school.



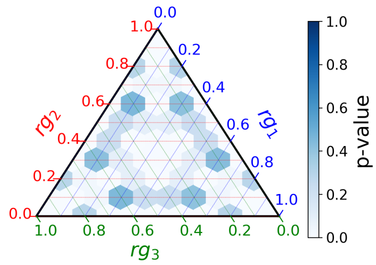
(a) Simulation



(b) Add Health data



(c) Difference



(d) p-value

FIGURE 6.6: Comparison between the simulation and the Add Health data with respect to the fraction of interracial friendships. (a) shows the simulated values generated by the model for each racial composition. (b) shows the average fraction of interracial friendships of schools with the same composition in the Add Health data after interpolation. (c) shows the absolute difference between (a) and (b). (d) shows the p-value of t-test for each composition, the p-value is 0 for interpolated values or when there is only 1 sampled school.

minimum when only one race dominates. From Figure 6.6c, we can see that the differences between the simulated data and the Add Health data are mostly negative, which means the simulation underestimates the fraction of interracial friendships. And compared to the assortativity, the fraction of interracial friendships has lower p-values, which implies that the difference between the simulation and the Add Health data for the fraction of interracial friendships is more significant.

Chapter 7

Conclusions

7.1 Main findings

This thesis focuses on understanding friendship segregation within schools. Specifically, it aims to examine whether there is a relationship between school ethnic composition and friendship segregation, and what is the optimal composition for mixed schools. To answer the questions, we generate a hypothetical school and mix students from two or three arbitrary racial groups into the school. It is notable that the model can be applied to any attributes like socioeconomic status. We focus on race because we have data to validate the model. The main findings of the thesis are as follows.

First of all, from the results of the experiments, we find that friendship segregation is correlated to school composition. In mixed schools with two dominant racial groups, the degree of assortativity by race reaches a maximum, and it decreases when there are three racial groups. This is in line with Moody's finding that friendship segregation peaks in moderately heterogeneous schools but declines at the highest heterogeneity levels [38].

Through a series of experiments, we show that homophily, contact theory and competition theory all have effects on friendship segregation. The assortativity increases when students' homophily remains the same while the group size of minority students grows. It indicates that when students have preferences for same-race friends, increasing the group size of minority students allows them to make their desired number of friends within their own group, and thereby increasing the assortativity within the school. In addition, we also show that the assortativity increases as the size of minority students becomes large, under the condition that students' homophily is positively correlated to the minority group size. This means that if the competition theory holds in schools, the assortativity is expected to become higher within mixed schools with comparable sizes of majority and minority students. On the contrary, if students' homophily decreases as the size of minority students grows, the assortativity is expected to be reduced by mixing more minority students within schools. These results suggest that simply exposing students to diverse ethnic peers does not necessarily eliminate segregation and promote interracial relations. Friendship segregation is aggravated if students have higher homophily in mixed schools. Therefore, efforts should be taken to prevent students from forming competitive relationships or restricting their friendships to their own group.

When mixing three racial groups in schools, the assortativity becomes less salient, especially when one of them can act as a "bridge". This might be explained by two aspects. First of all, mixing more racial groups into a school reduces the opportunities for students to make same-race friends. Therefore students need to make more interracial friendships to get their desired number of friends. Second of all, when schools have two racial groups that are comparable in size, students' racial identity and racial homophily are more likely to be salient. It is mitigated as they see more racial groups. According to this theory, an effective way of mixing students with different backgrounds into schools is to increase ethnic diversity while avoiding to aggravate the racial identity or homophily of students.

Second, in the results of our experiments, there is no optimal composition which leads to both the lowest level of assortativity and the highest fraction of interracial friendships. When there is only one dominant group, the assortativity is the lowest, however, the fraction of interracial friendships is also low. This finding suggests that it is critical to determine which measure of friendship segregation is appropriate in what context.

Finally, from the results of sensitivity analysis, we find that the cost of maintaining a friendship is an influential parameter to the assortativity. In the model, we assume that cost is constant, which means that theoretically one can make infinite friends as long as the extra utility is larger than the added cost. However, this is not in line with Dunbar's number, which is a suggested cognitive limit to the number of stable social relationships one can maintain [13]. Dunbar finds that the number of relationships that one can monitor at the same time is limited by their information processing capacity, and that is restricted by the brain size of primates. According to this theory, the cost of maintaining a friendship should increase as the number of friendships increases. This can be a new hypothesis in future work. Another interesting hypothesis is that the cost of maintaining a friendship for minority students is more expensive than for majority students in mixed schools, because minority students usually pay an additional cost to integrate into the native environment. If this is true, then minority students are expected to have fewer friendships than majority students.

7.2 Validity of using US school data to the Dutch context

Because we use friendship network data from US schools instead of Dutch schools, we need to compare the differences between the two contexts, which can help us understand potential differences in segregation within Dutch schools from the US case. In our research, friendship segregation is assumed to be driven mainly by three mechanisms: homophily, contact theory and competition theory. Therefore, we can reason about the difference in the segregation between the US data and the Dutch context through the three mechanisms.

The contact structure is determined by the ethnic composition of a school. Comparing the US school data to the the ethnic groups in Dutch primary schools, we find that the ethnic groups are not the same, and the relative sizes of the ethnic groups are also not similar. However this difference in composition has no effect on our results, since in our experiments we consider only two and three arbitrary ethnic groups, and analyze the segregation across all possible compositions. Likewise, the difference in the competition theory between the US and Dutch contexts has no influence on our results.

In both contexts, the dissimilarity varies for different combinations of ethnic groups since there might be some overlap between ethnic groups. However, we assume that the dissimilarity between any two ethnic groups is the same, which makes the results applicable to both contexts. If the dissimilarity between two ethnic groups is changed to be continuous, we might see differences between the US and the Dutch contexts.

However, the ethnic homophily of students in Dutch schools is expected to be less compared to US schools, according to Baerveldt et al. [3]. The argument is that immigrants in the Netherlands came voluntarily, therefore they are less opposed to the majority, compared to the US where native Americans and Afro-Americans show more oppositional relationships. If this argument is true, then the ethnic homophily of students in Dutch schools is lower than that in US schools. We would expect the level of friendship segregation within Dutch schools to be lower than our results.

7.3 Limitations and further work

Besides the possible hypotheses mentioned in the main findings, several improvements can be made in future work. In chapter 5, we find that the model underestimates assortativity when two racial groups are comparable in size, and it overestimates assortativity when one racial group dominates. This is possibly because the model ignores the relationship between students' homophily and school ethnic composition. In the future, this relationship could be taken into consideration.

Moreover, the model assumes that the dissimilarity between two ethnic groups is the same, which is not consistent with the observation from the real world. In reality, ethnic groups are not really discrete, and they usually intersect with other characteristics, creating a mosaic of groups and leading to more complexity.

In addition, for studies of school segregation, context is really important. Therefore, to study friendship segregation in Dutch schools, the collection of friendship network data from Dutch schools might shine more light on friendship dynamics in this context.

Appendix A

Supplementary Results

A.1 Convergence

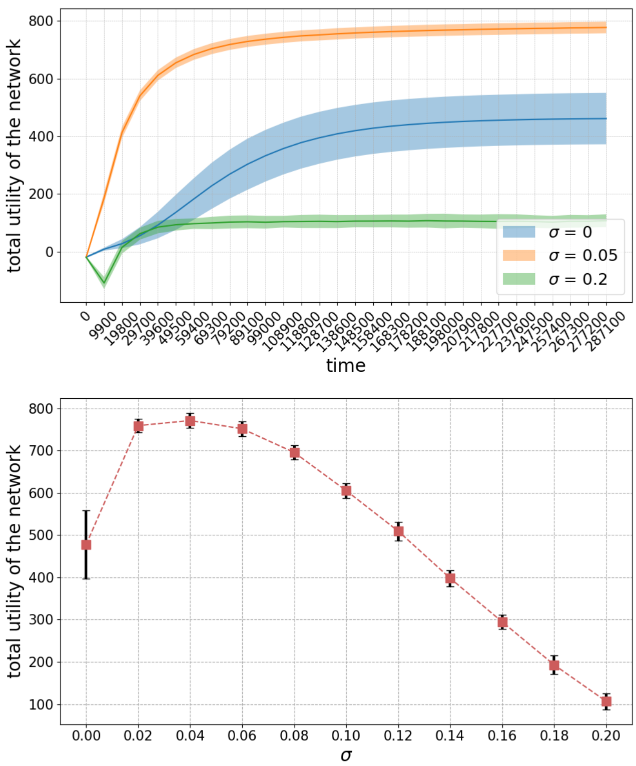


FIGURE A.1: The impacts of the scale parameter of the noise distribution σ on the equilibrium of the network. The top graph shows that the network converges over time, and the equilibrium is influenced by the scale parameter of noise. The bottom graph indicates that there is a switch point where the total utility of the network reaches an optimum. For these figures we use 100 agents and 100 repetitions.

A.2 Sensitivity analysis

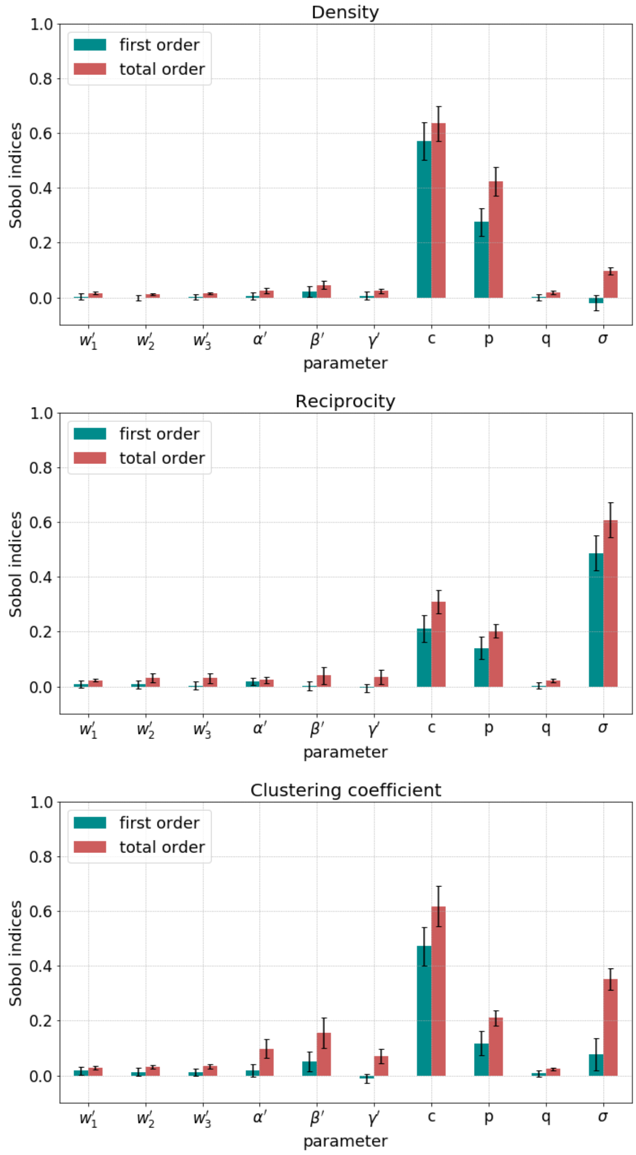


FIGURE A.2: The first-order and total-order Sobol indices for each model parameter. The model outputs are density, reciprocity and clustering coefficient respectively. The Sobol indices are calculated by 12,000 samples generated by Saltelli's sampling scheme [43]. The error bar indicates the 95% confidence interval for each parameter.

Bibliography

[1] Réka Albert and Albert-László Barabási. "Statistical mechanics of complex networks". In: *Rev. Mod. Phys.* 74 (1 2002), pp. 47–97. DOI: 10.1103/RevModPhys.74.47. URL: https://link.aps.org/doi/10.1103/RevModPhys.74.47.

[2] Tobias Stark Andreas Flache. "Preference or opportunity? Why do we find more friendship segregation in more heterogeneous schools?" In: (2009).

[3] Chris Baerveldt et al. "Ethnic boundaries and personal choice. Assessing the influence of individual inclinations to choose intra-ethnic relationships on pupils' networks". In: *Social Networks* 26.1 (2004), pp. 55–74.

[4] Albert-László Barabási and Márton Pósfai. *Network science*. Cambridge: Cambridge University Press, 2016. ISBN: 9781107076266 1107076269. URL: http://barabasi.com/networksciencebook/.

[5] Hubert M. Blalock. *Toward a Theory of Minority-Group Relations*. 1967.

[6] Michał Bojanowski and Rense Corten. "Measuring segregation in social networks". In: *Social Networks* (2014). ISSN: 03788733. DOI: 10.1016/j.socnet.2014.04.001.

[7] Shyam Boriah, Varun Chandola, and Vipin Kumar. "Similarity measures for categorical data: A comparative evaluation". In: (2008). DOI: 10.1137/1.9781611972788.22.

[8] Willem R Boterman. "School segregation in the free school choice context of Dutch cities". In: *Understanding School Segregation: Patterns, Causes and Consequences of Spatial Inequalities in Education. London: Bloomsbury* (2018), pp. 155–178.

[9] Willem R Boterman. "The role of geography in school segregation in the free parental choice context of Dutch cities". In: *Urban Studies* 56.15 (2019), pp. 3074–3094.

[10] Guus ten Broeke, George van Voorn, and Arend Ligtenberg. "Which sensitivity analysis method should i use for my agent-based model?" In: *JASSS* (2016). ISSN: 14607425. DOI: 10.18564/jasss.2857.

[11] Ennio Cascetta. "Random Utility Theory". In: *Transportation Systems Analysis: Models and Applications*. Boston, MA: Springer US, 2009, pp. 89–167. ISBN: 978-0-387-75857-2. DOI: 10.1007/978-0-387-75857-2_3. URL: https://doi.org/10.1007/978-0-387-75857-2_3.

[12] W A V Clark, F M Dieleman, and L de Klerk. "School Segregation: Managed Integration or Free Choice?" In: *Environment and Planning C: Government and Policy* 10.1 (1992), pp. 91–103. DOI: 10.1068/c100091. eprint: https://doi.org/10.1068/c100091. URL: https://doi.org/10.1068/c100091.

[13] Robin IM Dunbar. "Neocortex size as a constraint on group size in primates". In: *Journal of human evolution* 22.6 (1992), pp. 469–493.

[14] A. E. Eiben and J. E. Smith. *Natural Computing Series Introduction to Evolutionary Computing*. 2015. ISBN: 978-3-662-44873-1.

[15] Stephen Gorard. "The complex determinants of school intake characteristics and segregation, England 1989 to 2014". In: *Cambridge Journal of Education* 46.1

(2016), pp. 131–146. DOI: 10.1080/0305764X.2015.1045446. eprint: https://doi.org/10.1080/0305764X.2015.1045446. URL: https://doi.org/10.1080/0305764X.2015.1045446.

[16] Thomas Pettigrew Gordon Willard Allport Kenneth Clark. "The nature of prejudice". In: *Addison-wesley* 25 (1954).

[17] Peter Gramberg. "School Segregation: The Case of Amsterdam". In: *Urban Studies* 35.3 (1998), pp. 547–564. DOI: 10.1080/0042098984907. eprint: https://doi.org/10.1080/0042098984907. URL: https://doi.org/10.1080/0042098984907.

[18] Maureen T. Hallinan and Richard A. Williams. "The Stability of Students Interracial Friendships". In: *American Sociological Review* 52.5 (1987), p. 653. DOI: 10.2307/2095601.

[19] Kathleen Mullan Harris and Richard J. Udry. *National Longitudinal Study of Adolescent to Adult Health (Add Health) Wave I, 1994-1995*. Version V3. 2015. DOI: 10.15139/S3/11900. URL: https://doi.org/10.15139/S3/11900.

[20] M. Hewstone and R. Brown. "Contact and Conflict in Intergroup Encounters". In: 1986.

[21] Toshimitsu Homma and Andrea Saltelli. "Importance measures in global sensitivity analysis of nonlinear models". In: *Reliability Engineering and System Safety* (1996). ISSN: 09518320. DOI: 10.1016/0951-8320(96)00002-6.

[22] Mieke van Houtte and Peter A.J. Stevens. "School ethnic composition and students' integration outside and inside schools in Belgium". In: *Sociology of Education* (2009). ISSN: 00380407. DOI: 10.1177/003804070908200302.

[23] Kara Joyner and Grace Kao. "School Racial Composition and Adolescent Racial Homophily". In: *Social Science Quarterly* 81.3 (2000), pp. 810–825. ISSN: 00384941, 15406237. URL: http://www.jstor.org/stable/42864005.

[24] Agnieszka Kanas, Frank van Tubergen, and Tanja Van der Lippe. "The role of social contacts in the employment status of immigrants: A panel study of immigrants in Germany". In: *International Sociology* 26.1 (2011), pp. 95–122. DOI: 10.1177/0268580910380977. eprint: https://doi.org/10.1177/0268580910380977. URL: https://doi.org/10.1177/0268580910380977.

[25] Fariba Karimi et al. "Homophily influences ranking of minorities in social networks". In: *Scientific Reports* 8.1 (2018), p. 11077. DOI: 10.1038/s41598-018-29405-7. URL: https://doi.org/10.1038/s41598-018-29405-7.

[26] Sjoerd Karsten et al. "School choice and ethnic segregation". In: *Educational policy* 17.4 (2003), pp. 452–477.

[27] Vladimir Khmelkov and Maureen Hallinan. "Organizational Effects on Race Relations in Schools". In: *Journal of Social Issues* 55 (Dec. 1999), pp. 627 –645. DOI: 10.1111/0022-4537.00139.

[28] David Krackhardt and Robert N. Stern. "Informal Networks and Organizational Crises: An Experimental Simulation". In: *Social Psychology Quarterly* 51.2 (1988), pp. 123–140. ISSN: 01902725. URL: http://www.jstor.org/stable/2786835.

[29] Helen Ladd, Edward Fiske, and Nienke Ruijs. "Parental choice in The Netherlands: Growing concerns about segregation". In: (Jan. 2009).

[30] Helen F Ladd, Edward B Fiske, and Nienke Ruijs. "Parental choice in the Netherlands: Growing concerns about segregation". In: *National Conference on School Choice, Vanderbilt University*. 2009.

[31] Bram Lancee. "The economic returns of bonding and bridging social capital for immigrant men in Germany". In: *Ethnic and Racial Studies* 35.4 (2012), pp. 664–683. DOI: 10.1080/01419870.2011.591405. eprint: https://doi.org/

10.1080/01419870.2011.591405. URL: https://doi.org/10.1080/01419870.2011.591405.

[32] Lars Leszczensky and Sebastian Pink. "Ethnic segregation of friendship networks in school: Testing a rational-choice argument of differences in ethnic homophily between classroom- and grade-level networks". In: *Social Networks* 42 (2015), pp. 18–26. ISSN: 0378-8733. DOI: https://doi.org/10.1016/j.socnet.2015.02.002. URL: https://www.sciencedirect.com/science/article/pii/S0378873315000076.

[33] Bart H. H. Golsteyn Lex Borghans and Ulf Zölitz. "Parental Preferences for Primary School Characteristics". In: *The B.E. Journal of Economic Analysis & Policy* 15.1 (2015), pp. 85–117. DOI: doi:10.1515/bejeap-2014-0032. URL: https://doi.org/10.1515/bejeap-2014-0032.

[34] Eleanor E. Maccoby and Carol Nagy Jacklin. "Gender Segregation in Childhood". In: ed. by Hayne W. Reese. Vol. 20. Advances in Child Development and Behavior. JAI, 1987, pp. 239–287. DOI: https://doi.org/10.1016/S0065-2407(08)60404-8. URL: https://www.sciencedirect.com/science/article/pii/S0065240708604048.

[35] J. Miller Mcpherson and Lynn Smith-Lovin. "Homophily in Voluntary Organizations: Status Distance and the Composition of Face-to-Face Groups". In: *American Sociological Review* 52.3 (1987), p. 370. DOI: 10.2307/2095356.

[36] Miller Mcpherson, Lynn Smith-Lovin, and James M Cook. *BIRDS OF A FEATHER: Homophily in Social Networks*. Tech. rep. 2001. URL: www.annualreviews.org.

[37] Angelo Mele. "A Structural Model of Segregation in Social Networks". In: *SSRN Electronic Journal* (2013). DOI: 10.2139/ssrn.2294957.

[38] James Moody. "Race, school integration, and friendship segregation in America". In: *American Journal of Sociology* (2001). ISSN: 00029602. DOI: 10.1086/338954.

[39] Ted Mouw and Barbara Entwisle. "Residential Segregation and Interracial Friendship in Schools."" In: *American Journal of Sociology - AMER J SOCIOL* 112 (Sept. 2006). DOI: 10.1086/506415.

[40] Thomas F Pettigrew and Linda R Tropp. "A meta-analytic test of intergroup contact theory". In: *Journal of personality and social psychology* 90.5 (2006), pp. 751–783. ISSN: 0022-3514. DOI: 10.1037/0022-3514.90.5.751. URL: https://doi.org/10.1037/0022-3514.90.5.751.

[41] "Reducing prejudice via direct and extended cross-group friendship". In: *European Review of Social Psychology* 18.1 (2007), pp. 212–255. URL: https://doi.org/10.1080/10463280701680297.

[42] Erdos RENYI. "On random graph". In: *Publicationes Mathematicate* 6 (1959), pp. 290–297.

[43] Andrea Saltelli. "Making best use of model evaluations to compute sensitivity indices". In: *Computer Physics Communications* (2002). ISSN: 00104655. DOI: 10.1016/S0010-4655(02)00280-1.

[44] Andrea Saltelli et al. *Global Sensitivity Analysis. The Primer*. 2008. ISBN: 9780470059975. DOI: 10.1002/9780470725184.

[45] S. Schalk-Soekar, F.J.R. van de Vijver, and M. Hoogsteder. "Attitudes toward multiculturalism of immigrants and majority members in the Netherlands". English. In: *International Journal of Intercultural Relations* 28.6 (2004), pp. 533–550. ISSN: 0147-1767.

[46] Peer Scheepers, Mérove Gijsberts, and Marcel Coenders. "Ethnic Exclusionism in European Countries. Public Opposition to Civil Rights for Legal Migrants

as a Response to Perceived Ethnic Threat". In: *European Sociological Review* 18 (Mar. 2002). DOI: 10.1093/esr/18.1.17.

[47] L. S. Shapley. "17. A Value for n-Person Games". In: *Contributions to the Theory of Games (AM-28), Volume II*. Ed. by Harold William Kuhn and Albert William Tucker. Princeton University Press, 2016, pp. 307–318. DOI: doi:10.1515/9781400881970-018. URL: https://doi.org/10.1515/9781400881970-018.

[48] Wesley Shrum, Neil H. Cheek, and Saundra MacD. Hunter. "Friendship in School: Gender and Racial Homophily". In: *Sociology of Education* 61.4 (1988), pp. 227–239. ISSN: 00380407, 19398573. URL: http://www.jstor.org/stable/2112441.

[49] Sanne Smith et al. "Ethnic Composition and Friendship Segregation: Differential Effects for Adolescent Natives and Immigrants". In: *American Journal of Sociology* 121.4 (2016), pp. 1223–1272. DOI: 10.1086/684032. eprint: https://doi.org/10.1086/684032. URL: https://doi.org/10.1086/684032.

[50] Eunhye Song, Barry L. Nelson, and Jeremy Staum. "Shapley effects for global sensitivity analysis: Theory and computation". English (US). In: *SIAM-ASA Journal on Uncertainty Quantification* 4.1 (2016), pp. 1060–1083. ISSN: 2166-2525. DOI: 10.1137/15M1048070.

[51] Trueck Stefan and Rachev Svetlozar T. "Chapter 2 - Rating and Scoring Techniques". In: *Rating Based Modeling of Credit Risk*. Ed. by Trueck Stefan and Rachev Svetlozar T. Academic Press Advanced Finance. Boston: Academic Press, 2009, pp. 11–30. DOI: https://doi.org/10.1016/B978-0-12-373683-3.00003-8. URL: https://www.sciencedirect.com/science/article/pii/B9780123736833000038.

[52] Rhiannon N. Turner et al. "Reducing prejudice via direct and extended cross-group friendship". In: *European Review of Social Psychology* 18.1 (2007), pp. 212–255. DOI: 10.1080/10463280701680297. eprint: https://doi.org/10.1080/10463280701680297. URL: https://doi.org/10.1080/10463280701680297.

[53] Eva van der Heijden and H.A.G. de Valk. "Schooluitval onder tweede generatie jongeren". English. In: *Jaarrapport Integratie 2018*. Centraal Bureau voor de Statistiek (CBS), Nov. 2018, pp. 191–219. ISBN: 978-90-357-2148-7.

[54] Paul Vedder. "Black and white schools in the Netherlands". In: *European Education* 38.2 (2006), pp. 36–49.

[55] Thirumalaisamy P Velavan and Christian G Meyer. "The COVID-19 epidemic". In: *Tropical medicine & international health* 25.3 (2020), p. 278.

[56] Lotte Vermeij, Marijtje AJ Van Duijn, and Chris Baerveldt. "Ethnic segregation in context: Social discrimination among native Dutch pupils and their ethnic minority classmates". In: *Social networks* 31.4 (2009), pp. 230–239.

[57] Miranda H.M. Vervoort, Ron H.J. Scholte, and Peer L.H. Scheepers. "Ethnic composition of school classes, majority-minority friendships, and adolescents' intergroup attitudes in the Netherlands". In: *Journal of Adolescence* (2011). ISSN: 01401971. DOI: 10.1016/j.adolescence.2010.05.005.

[58] *Voorrangsregels*. URL: https://www.schoolwijzernijmegen.nl/uitleg/voorrangsregels.

[59] G. Walraven. "The Netherlands: interventions to counteract school segregation and facilitate integration in education". In: (2013).

[60] Ke Wang, Amy Rathburn, and Lauren Musu. "School Choice in the United States: 2019. NCES 2019-106." In: *National Center for Education Statistics* (2019).

[61] Peter Wolfgram. "Leerlingen, basisscholen en hun buurt". In: (2009). URL: http://docplayer.nl/7717803-Leerlingen-basisscholen-en-hun-buurt-een-onderzoek-naar-de-samenstelling-van-schoolpopulaties-en-buurtpopulaties-peter-wolfgram.html.

[62] Stephen Worchel and William G. Austin. *The social psychology of intergroup relations*. Brooks/Cole Publishing Company, a division of Wadsworth, Inc., 1979, 33–47.

[63] Jinghui Zhong et al. "Density-based evolutionary framework for crowd model calibration". In: *Journal of Computational Science* (2015). ISSN: 18777503. DOI: 10.1016/j.jocs.2014.09.002.