UNIVERSITY OF KENT

DOCTORAL THESIS

Complex Contagion of Desirable Behavior in Adolescent Social Networks – a Simulation Model

Author: Tobias JORDAN
Words: 45.173
Academic school: School of Computing

Year of Submission: 2022
Pages: 163

Abstract This thesis proposes a network spreading model for the simulation of complex contagion processes in social networks. Current models for political decision support often fail in reproducing macro phenomena that emerge from micro behavior. The approach aims at overcoming those shortcomings related to restrictions of current Dynamic Stochastic General Equilibrium (DSGE)-models to rational homogeneous individuals on the ground of the connection of Network Science and Agent-based Simulation. Hereby special attention is drawn to applications in the field of Conditional Cash Transfer Programs (CCT). Using a case study that concerns the educational commitment of adolescents in northeastern Brazil, a step by step description of model setup is given. The contribution to the current state of the research is hereby fourfold. A novel approach to model the diffusion of educational commitment among adolescents (the effort they put into learning) as a Coordination-Game is proposed and it is demonstrated that it adequately represents reality. Moreover, the problem of missing data is addressed in this thesis from the perspective of a modeler that aims at creating meaningful large-scale network simulations. Adaptions of existing link-prediction and network generation approaches as well as a combination of both are proposed as a new, well performing method to impute missing links in social networks, stemming from surveys and online sources. It is shown that both, the "Boundary Specification Problem" and the "Fixed Choice Effect" can be tackled successfully with this techniques. Moreover, the thesis proposes an implementation of a Learning Classifier System (LCS)-based decision module for the agents within the simulation model. This novel adoption of the well known approach provides the agents with bounded rationality and hence enables more realistic simulations. For the first time, it is demonstrated that this decision module mimics human reasoning about educational commitment well. Eventually, an adaption of the standard Genetic Algorithm is proposed and developed for the task of parameter estimation and fitting the simulation model to real data. It is demonstrated that the Genetic Algorithm is well suited for this task.

Contribution to Science We have developed a novel approach to model the diffusion of educational commitment among adolescents as a Coordination-Game. Hereby the stage is set for the incorporation of Coordination-Game like imitation processes to a more sophisticated model of the spread of educational commitment.

We have provided solutions for the problem of missing data between isolated components from social network surveys, stemming from "Fixed-Choice-Effect" and "Boundary Specification Problem", enabling simulations on a global network model. The presented research contributes to the literature with the proposal of three approaches that are capable of generating interconnected networks with different features.

We have proposed a novel Learning Classifier implementation that provides the agents in an Agent-based Model (ABM) with bounded rationality and enables simulations of the diffusion of educational attitudes via social networks.

We have provided a Genetic Algorithm that successfully estimates the input parameters for the simulation model so that the model closely reproduces real data. Hereby, attention is drawn to the possibility of decentral estimation of ABM and it is demonstrated that decentral estimation approaches yield much better results than attempts to global parameter estimation.

Impact Statement The simulation model that is proposed here helps to better understand complex spreading processes in social networks and respective multiplier- and spillover effects of public policies. This includes the analysis, evaluation and potential alteration of Conditional Cash Transfer Programs (CCT) and other policies concerning education and social equality. By providing insights in the process of diffusion and enabling the realistic simulation of what-if scenarios, this research has the potential to contribute to the construction of human capital and the eradication of poverty and other United Nations Sustainable Development Goals such as the good quality of education and reduced inequality.

UNIVERSITY OF KENT

DOCTORAL THESIS

Complex Contagion of Desirable Behavior in Adolescent Social Networks – a Simulation Model

Author:
Tobias JORDAN

Supervisor:
Prof. Philippe DE WILDE
Co-Supervisor:
Prof. Fernando BUARQUE DE LIMA NETO

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the
School of Computing

January 2022

Declaration of Authorship

I, Tobias JORDAN, declare that this thesis titled, "Complex Contagion of Desirable Behavior in Adolescent Social Networks – a Simulation 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: 21.08.2022

"Ninguém educa ninguém, ninguém se educa a si mesmo, os homens se educam entre si, mediatizados pelo mundo "

Paulo Freire

UNIVERSITY OF KENT

Abstract

Division of Computing, Engineering and Mathematical Sciences
School of Computing

Doctor of Philosophy

Complex Contagion of Desirable Behavior in Adolescent Social Networks – a Simulation Model

by Tobias JORDAN

This thesis proposes a network spreading model for the simulation of complex contagion processes in social networks. Current models for political decision support often fail in reproducing macro phenomena that emerge from micro behavior. The approach aims at overcoming those shortcomings related to restrictions of current Dynamic Stochastic General Equilibrium (DSGE)-models to rational homogeneous individuals on the ground of the connection of Network Science and Agent-based Simulation. Hereby special attention is drawn to applications in the field of Conditional Cash Transfer Programs (CCT). Using a case study that concerns the educational commitment of adolescents in northeastern Brazil, a step by step description of model setup is given. The contribution to the current state of the research is hereby fourfold. A novel approach to model the diffusion of educational commitment among adolescents (the effort they put into learning) as a Coordination-Game is proposed and it is demonstrated that it adequately represents reality. Moreover, the problem of missing data is addressed in this thesis from the perspective of a modeler that aims at creating meaningful large-scale network simulations. Adaptions of existing link-prediction and network generation approaches as well as a combination of both are proposed as a new, well performing method to impute missing links in social networks, stemming from surveys and online sources. It is shown that both, the "Boundary Specification Problem" and the "Fixed Choice Effect" can be tackled successfully with this techniques. Moreover, the thesis proposes an implementation of a Learning Classifier System (LCS)-based decision module for the agents within the simulation model. This novel adoption of the well known approach provides the agents with bounded rationality and hence enables more realistic simulations. For the first time, it is demonstrated that this decision module mimics human reasoning about educational commitment well. Eventually, an adaption of the standard Genetic Algorithm is proposed and developed for the task of parameter estimation and fitting the simulation model to real data. It is demonstrated that the Genetic Algorithm is well suited for this task.

Contribution to Science

- We have developed a novel approach to model the diffusion of educational commitment among adolescents as a Coordination-Game. Hereby the stage is set for the incorporation of Coordination-Game like imitation processes to a more sophisticated model of the spread of educational commitment.
- We have provided solutions for the problem of missing data between isolated components from social network surveys stemming from "Fixed-Choice-Effect" and "Boundary Specification Problem", enabling simulations on a global network model. The presented research contributes to the literature with the proposal of three approaches that are capable of generating interconnected networks with different features.
- We have proposed a novel Learning Classifier implementation that provides the agents in an Agent-based Model (ABM) with bounded rationality and enables simulations of the diffusion of educational attitudes via social networks.
- We have provided a Genetic Algorithm that successfully estimates the input parameters for the simulation model so that the model closely reproduces real data. Hereby, attention is drawn to the possibility of decentral estimation of ABM and it is demonstrated that decentral estimation approaches yield much better results than attempts to global parameter estimation.

Impact Statement The simulation model that is proposed here helps to better understand complex spreading processes in social networks and respective multiplier- and spillover effects of public policies. This includes the analysis, evaluation and potential alteration of Conditional Cash Transfer Programs (CCT) and other policies concerning education and social equality. By providing insights in the process of diffusion and enabling the realistic simulation of what-if scenarios, this research has the potential to contribute to the construction of human capital and the eradication of poverty and other United Nations Sustainable Development Goals such as the good quality of education and reduced inequality.

Acknowledgements

Acknowledgments

Throughout the writing of this dissertation I have received a great deal of support and assistance.

First of all, I would like to thank my supervisor, Prof. Philippe De Wilde. His support and advice were essential in formulating the research questions and his feedback on each step brought my work to a higher level. Prof. De Wilde guided me through the research process and let me benefit from his vast experience. I could not think of a better supervisor.

Moreover, I own particular thanks to Polytecnic School of Pernambuco at Universidade de Pernambuco (UPE), in person to Prof. Fernando Buarque de Lima Neto for the outstanding support from the beginning until the end of this research project. He not only aroused my interest for the subject of Agent-based Social Simulation but also introduced me to the persons and institutions that were key for the presented research. Prof. Buarque accompanies my personal and professional development with words and deeds since more than ten years and significantly shaped the path I took.

This path would have been definitely different without the support from Dr. Michael Craanen. Everything started with our journey to Recife many years ago.

Also, I very much appreciated the guidance received from the members of my supervisory team at University of Kent, Prof. Howard Bowman and Prof. Peter Rodgers.

Special thanks are due to Fundação Joaquim Nabuco, in particular to Dr. Isabel Raposo and Dr. Michela B. Camboim Gonçalves Feitosa for providing data and domain expertise.

In addition, this research was partially funded by Andrea von Braun Stiftung. Herewith I would like to express my sincere thanks for this support.

Contents

List of Figures

List of Tables

List of Abbreviations

ABC	**A**gent **B**ased **C**omputational Economics
ABM	**A**gent **B**ased **M**odel
ABSS	**A**gent **B**ased **S**ocial **S**imulation
CCT	**C**onditional **C**ash **T**ransfer
DSGE	**D**ynamic **S**tochastic **G**eneral **E**quilibrium
FUNDAJ	**FUNDA**ção **J**oaquim Nabuco
GMM	**G**aussian **M**ixed **M**odel
LCS	**L**earning **C**lasssifier **S**ystem
MAR	**M**issing **A**t **R**andom
MCAR	**M**issing **C**ompletely **A**t **R**andom
MNAR	**M**issing **N**ot **A**t **R**andom
PSO	**P**article **S**warm **O**ptimization
SA	**S**imulated **A**nnealing
SMD	**S**imulated **M**inimum **D**istance

For Amelie and Carol

Chapter 1

Introduction

"Ninguém educa ninguém, ninguém se educa a si mesmo, os homens se educam entre si, mediatizados pelo mundo." [1]

Paulo Freire wrote those words in his famous "Pedagogy of the Oppressed", expressing that education must not regard the student as an empty vessel that needs to be filled with wisdom. Instead, according to Freire, education occurs in a dialogue between student and teacher and particularly in the dialogue among students.

This dialogue-centered approach points out how important interaction with peers is for educational success. Education therewith does not only depend on the content that is being taught but on how peers discuss this content among each other and especially if and how the group develops an attitude toward education. However, to benefit from the described peer effect and the benefits of education, pupils must be able and willing to attend school.

1.1 Motivation for the Research - Policies for Poverty Eradication in Brazil

The necessity of school attendance is one of the core ideas of Conditional Cash Transfer (CCT) programs that have been emerging all over the world within the last two decades. The CCT-Programs usually combine a cash transfer to poor families, commonly obtaining incomes nearby the poverty line or below, under the condition that differing human capital building services are accessed. Such services may comprise health or education. The objective is unanimously twofold: First, the eradication of extreme poverty by direct cash transfer and second, avoiding the inter-generational transmission of poverty fostering the education and health of the following generations [2].

The Brazil Government-funded "Bolsa Familia" Program is worldwide one of the largest Conditional Cash Transfer Programs, reaching more than 46 Million people. The Brazilian National Congress approved in 2021 a successor program for "Bolsa Familia" named "Auxilio Brazil". After intense debates in both chambers (Federal Senate and Chamber of Deputies), the program can be interpreted as an extension of the existing program. Participating families receive a monthly benefit according

to family size and monthly houshold income. Extremely poor families are hereby eligible for a fixed benefit without conditions, while moderately poor families are obliged to care for the school attendance, and vaccinations of their children, as well as for the pre- and postnatal preventive medical checkups of women. The aim of the policy is to achieve two goals with one program: fight hunger and poverty and build human capital, educating the youngest to break the vicious circle of poverty [2].

The CCT programs and particularly "Bolsa Familia" have been studied extensively and strong evidence on the positive effect of those programs on school enrollment exists. Moreover, it has been shown that CCTs have a negative effect on school dropout rates and likewise, positive effects on the average length of schooling of participants [3], [4].

To a lesser extent, yet not less convincing, hints have been found that those positive effects of the CCTs are not limited to the eligible families and individuals. School enrollment of significantly affected neighborhoods also rises due to peer effects and positive educational outcomes of participants as well a penalties applied to noncompliance with the conditions [5], [6]. Literature also refers to those peer effects with the term “spillover”. This indicates that beyond the direct effects of schooling and learning, the role of the school as a social space where connections and friendships are formed and opinions and attitudes are built, cannot be underestimated and the effects of intra- and inter school networks should be taken into consideration when designing and evaluating public measures such as the Conditional Cash Transfer Programs.

The above gave rise to the idea for the research in this thesis. In networks of individuals, network effects can be identified statistically and it has been proven that they heavily depend on network properties and the nature of whatever is spreading through the network. Departing from the idea of designing and evaluating CCTs, the public administrator may also want to examine whether those network effects are present and important for the policy in question and if so, how they could be used to generate positive outcomes. For related questions that emerge during the process of policy design, the state of the art General Equilibrium Models (DSGE-Models) [7] may be too static and may not account for the strong effect that local clusters and the heterogeneity of the society have on the overall outcomes of public policies. Hence, a modeling approach better suited to simulate the complexity of the studied societies and more flexible for the policy-maker, should be developed.

1.2 Theoretical Grounding and Approach

General Equilibrium Models [7] represent the most popular paradigm for macroeconomic simulation and thereby the most popular instrument for political decision support. However, those models are based on strong neo-classical assumptions like rational decision making, perfect market behavior, and perfect information for all actors. These assumptions do not hold in the real world and lead to a stereotype average consumer, that is the rational individual or Homo Oeconomicus. Criticism of Homo Oeconomicus became louder during the last decade due to the unrealistic assumptions of the underlying model and the failure of rational individual-based models in predicting problems such as the big economic crisis at the beginning of the 21st century [8]. These assumptions also suppose that our highly heterogeneous societies can be understood by investigating the behavior of rational average individuals and their communication and group behavior. It is claimed that irrationality does not exist, or at least not affect the crowd's behavior [9].

To better understand and predict human behavior, the concept of Agent-based modeling came up as an alternative for economists. Agent-based models use autonomous acting, communicating computer programs, the so-called agents, that can decide in a bounded rational way [10]. Agents within these models may resemble individuals, consumers, or juridical persons like companies. The element of bounded rationality is here represented in the form of a decision mechanism that enables the agents to strive for optimal decisions but that also reflects the irrational bias that real decision-makers usually face. This bias can be introduced by distinct means such as, for instance, providing limited processing capacities, rules of thumb and other heuristic decision making approaches, or providing the agents only with a share of the information related to the given decision. In combination with the representation of heterogeneous individual characteristics and inter-temporal effects, the element of bounded rationality triggers the emergence of macro phenomena from micro behavior that may not be expected from simulations based upon the paradigm of the rational individual. Agent-based models thereby are enabled, to better model human heterogeneity and thus create a more sophisticated image of reality.

In parallel, and accompanying the findings on spillovers of CCTs (see Section 1.1), the study of social networks revealed that our peers and even peers of peers influence our behavior [11]. It is understood that besides behavior, other diffusion processes occur on social networks, such as the spreading of contagion and information. However, it has been shown that the diffusion pattern depends on the nature of whatever is diffusing throughout the network.

Literature distinguishes broadly between simple- and complex contagion [12]. Simple contagion describes processes where a single contact between an infected and a susceptible individual triggers contagion. This spreading pattern can be observed for the contagion of diseases and rumors. On the other hand, complex contagion encompasses situations where a single contact is not sufficient to trigger a diffusion

process. Rather multiple exposures are needed before contagion occurs. Such complex contagion can, among others, certainly be assumed for the decision to attend school classes or to dedicate at school.

Several studies with complex contagion processes also revealed that network structure affects complex contagion differently [13]. While weak ties or single connections between otherwise unconnected components play a major role in the spreading of simple contagion, they appear to impede the diffusion of complex contagion. Based upon those findings it has been investigated how contagious processes may be stopped by the removal of links and influential nodes from the network (e.g. spreading of riots and diseases). However, in certain contexts, the opposite strategy may be of interest, particularly if the contagious behavior is desirable. A policymaker, who is responsible for a CCT as "Bolsa Familia", for example, might want to investigate if the positive spillover effects can be enforced by rearranging social networks of children. For instance, such rearrangement could be achieved, bringing children from different social contexts together through exchange programs or social activities.

Hence, models that are capable of simulating such spreading processes coherently and realistically may be of great help for decision-makers when considering how to impede undesirable behavior or reinforce desirable behavior spreading.

Therefore, this thesis addresses the process of drafting and implementing a simulation model for complex spreading on social networks and provides solutions for pending problems inherent to this process. A model on the ground of the connection of Network Science and Agent-based Simulation is proposed as a solution. In this thesis our contribution to the current state of the research is hereby fourfold:

The first challenge when developing such a simulation model is the choice of the underlying spreading pattern. A problem-specific approach must be found that represents well the mechanism that triggers the diffusion of the behavior in question. Therefore, this thesis gives guidance on the sensitive choice of a fitting model environment in Chapter 2[1]. Theoretical background on the representation of spreading processes is given and the suitability of a Coordination-Game mechanism [15] is studied in depth.

Another persisting problem for modelers is the availability of adequate data and - more importantly- the quality and completeness of those data. Particularly when it comes to social network data, the available data is generally restricted to a certain environment. For example, network data is usually solely available for all members of the entity where the network study was conducted. Hereby, entities may be thought of as companies, associations or schools for example. Even when relying on data from online network sources such as social online networks, purchasing networks, etc., the available data is restricted to the users of such services. For coherent large-scale simulation models, however, interconnected networks including comprehensive information about the incorporated individuals are needed in a

[1] The contents of this Chapter have been previously published in the proceedings of a peer reviewed conference [14]

larger scale. Hence, the problem of missing data is addressed in this thesis from the perspective of a modeler that aims at creating meaningful large-scale network simulations. Hereby the possibilities of both areas of research, network sciences, and Agent-based modeling are evaluated and put together. Chapter 3[2] introduces the research question and the current state of the research in greater detail.

Moreover, as stated above, this approach stems from Agent-based modeling and network science and overcomes the restriction of current DSGE-Models to the rational individual. To this purpose, the thesis proposes an implementation of a Learning Classifier System (LCS)-based decision module for the agents within the simulation model. This approach provides the agents with bounded rationality [10] and hence enables more realistic simulations. Chapter 4[3] explains the approach and evaluates the suitability for the given scenario. Moreover, a literature review on Agent-based models and their applications is given.

Finally, the challenge of finding the right parameters for such large-scale complex network simulation is addressed. Difficulties in calibration and parameter estimation are pointed out. It is argued that heuristic estimation approaches are better suited for such tasks than indirect inference approaches. The heuristic parameter estimation is presented at the basis of a Genetic Algorithm and particular obstacles with this approach as well as possible solutions are pointed out in Chapter 5.

1.3 Practical Example

The theoretical contributions are presented alongside a practical example that concerns the educational commitment of adolescents in Brazil. The payoff of educational efforts generally materializes after a considerable time gap and may not be recognized clearly by adolescents. This is why the educational commitment of adolescents depends heavily on peer behavior. Nevertheless, high educational commitment is clearly a desirable behavior and hence it may be of public interest to understand how diffusion of that behavior can be promoted. This holds especially in the context of the design and evaluation of CCTs.

It is also known that educational success and commitment are distributed heterogeneously among the society, where socioeconomic status correlates with the attitudes towards education.

Bringing together groups of adolescents with different socioeconomic backgrounds, for example from private and public schools may be a method to re-shape social networks. Desirable attitudes towards education may then spread to regions within the social network they would not have reached otherwise. To that purpose, it may

[2] The contents of this Chapter have been previously published as an article in a peer reviewed journal [16]

[3] The contents of this Chapter have been previously published in the proceedings of a peer reviewed conference [17]

be helpful for public administration to know, how the re-connection of networks could affect the spreading of desirable behavior throughout social systems.

To facilitate the conduction of such what-if simulations we propose a simulation-based model of behavior spreading among individual agents that reconnect on the ground of homophily[4] and network structure when brought together.

1.4 Contribution to Science

As stated above, four major contributions to the current state of research are made. A novel approach to model the diffusion of the behavior in question is adequately represented as a Coordination-Game. Moreover, the problem of missing data is addressed in this thesis from the perspective of a modeler that aims at creating meaningful large-scale network simulations. Adaptions of existing link-prediction and network generation approaches as well as a combination of both are proposed as a new, well performing method to impute missing links in social networks that stem from surveys and online sources. It is shown that both, the "Boundary Specification Problem" and the "Fixed Choice Effect" can be tackled successfully with this techniques. Moreover, the thesis proposes an implementation of a Learning Classifier System (LCS)-based decision module for the agents within the simulation model. This different adoption of the well known approach provides the agents within the simulation model with bounded rationality and hence enables more realistic simulations. For the first time it is demonstrated that this decision module mimics human reasoning about educational commitment well. Eventually, an adaption of the standard Genetic Algorithm is proposed and developed for the task of parameter estimation and fitting of the simulation model to real data. Further, it is demonstrated that the Genetic Algorithm is well suited for this task. Summarizing, the following contributions are made:

- A novel approach to model the diffusion of educational commitment among adolescents as a Coordination-Game. Hereby the stage is set for the incorporation of Coordination-Game like imitation processes to a more sophisticated model of the spread of educational commitment.
- Solutions for the problem of missing data between isolated components from social network surveys stemming from "Fixed-Choice-Effect" and "Boundary Specification Problem" are provided, enabling simulations on a global network model. The presented research contributes to the literature with the proposal of three approaches that are capable of generating interconnected networks with different features.

[4] Homophily is the principle that a contact between similar people occurs at a higher rate than among dissimilar people [18]

- A novel Learning Classifier implementation is presented that provides the agents in an Agent-based Model with bounded rationality and enables simulations of the diffusion of educational attitudes via social networks.
- A Genetic Algorithm is proposed that successfully estimates the input parameters for the simulation model so that the model closely reproduces real data. Hereby, attention is drawn to the possibility of decentral estimation of ABM and it is demonstrated that decentral estimation approaches yield much better results than attempts to global parameter estimation.

Impact Statement The simulation model that is proposed here helps to better understand complex spreading processes in social networks and respective multiplier- and spillover effects of public policies. This includes the analysis, evaluation and potential alteration of Conditional Cash Transfer Programs (CCT) and other policies, concerning education and social equality. By providing insights in the process of diffusion and enabling the realistic simulation of what-if scenarios, this research has the potential to contribute to the construction of human capital and the eradication of poverty and other United Nations Sustainable Development Goals such as the good quality of education and reduced inequality.

1.5 Data

The practical example presented above is based upon an extensive data-set stemming from the study *"Determinantes do desempenho escolar na rede de ensino fundamental do Recife"* [19]. The original study had the objective to estimate a linear hierarchic model in order to quantify the effect that schooling infrastructure projects have on the performance of school children. For this purpose, a survey was conducted by Fundação Joaquim Nabuco (FUNDAJ) in 2013, gathering data from more than 4000 pupils in public schools in the north-eastern Brazilian city Recife. The survey collected via questionnaire-based interviews information about socioeconomic status and family habits that reflect the importance of education and how the attitude towards education is transmitted from parents to children. Moreover, information about the neighborhood of the children and their integration into their neighborhood have been collected. The study applied questionnaires from three perspectives: pupil, parents/ legal guardians and teacher. Additionally, the status of health, self-esteem, performance at school and social relations of the children where assessed. Those data contain among others the social network of the pupils and their performance in the subject maths at the beginning and at the end of a school year. Children were asked to nominate their five best friends. If the friends went to the same class, a questionnaire was sent to those students as well. In this way, a network containing 4191 students was generated. However, 573 students that did not nominate any friend within their class were removed from the data-set, leading to a total number

of 3618 vertices. Since students could not nominate friends outside their school, the network is subdivided into 219 clusters from 122 schools.

FUNDAJ provides a full description of the applied questionnaires and the collected items as well as the full data-set in an online repository [19]. The variables that have been applied for the research in this thesis are listed and described in appendices A and B.

Chapter 2

Finding a Spreading Model: Modeling Complex Contagion of Behavior and Spreading Processes in Social Networks

2.1 Why and How to Model Complex Contagion Processes

A first step towards creating a simulation model for the spread of behavior throughout social networks is to develop a schematic understanding of the spreading process in question and consequently an adequate computational implementation of that process. Current research proves that peer influence exists and also points out situations where the effects are stronger or weaker. However, to adequately model the effects of contagion and information spreading for simulation and prediction models, a coherent representation is required. This is the motivation for the following Chapter. A simple, but intuitive way of thinking of behavior spreading is to assume that individuals within a network tend to behave just like the majority of peers. Respectively, there is a threshold for the number of peers of an individual that need to present certain behavior for the individual in question also to adopt that behavior. Easley and Kleinberg have developed such a simple model under the term "Coordination-Game" [15], [20]. Pragmatically, the contagion of behavior within friendship networks may be implemented as a Coordination-Game. Assuming that individuals benefit from compliant behavior while opposing behavior generates zero pay-offs.

In order to evaluate the performance of the implementation of the Coordination-Game mechanism, it is subsequently implemented based on two different data-sets, each containing information about friendship ties, as well as time-dependent information about specific behaviors or behavioral outcomes. As an assessment of the suitability of the approaches, it is checked to what extent the models are capable of reproducing real data.

Synopsis. A review of the state of relevant research regarding the diffusion of behavior and information in social networks and a description of the data-sets used within this work can be found in Section 2.2. Section 2.3 presents the implementation of the Coordination-Game and its adaption to the different data-sets while Section 2.4 contains experimental setup and results. The experimental results are discussed in Section 2.5. The conclusion of this Chapter follows in Section 2.6.

2.2 Background and Data - Complex Contagion

This Section presents theoretical background from the literature about the genesis and the current state of research in the field of Social Network Sciences and particularly the findings concerning spreading and contagion processes. In detail, the Coordination-Game mechanism is presented [15]. In a second subsection, the data-sets that build the ground for the presented experiments are described.

2.2.1 Background on Spreading and Contagion Processes

Existing research indicates that human decisions, opinions, norms and behavior are influenced by the social environment [21]. Also, human social networks haven been shown to be an effective search tool [22]. Social influence and contagion as well as the spread of behavior and information through social networks have been documented in a wide range of cases [11]. For instance, diffusion of voting behavior [11] and obesity [23] has been proven statistically. Moreover, cooperative behavior has been shown to be contagious, though depending on tie structure and dynamics [24] and recent studies revealed contagiousness of emotions [25]. Other behaviors do not spread like sexual orientation [26]. The literature distinguishes broadly between simple- and complex contagion [12]. Simple contagion describes processes where a single contact between an infected and a susceptible individual triggers contagion. This spreading pattern can be found for the contagion of diseases and rumours. On the other hand, complex contagion encompasses situations where a single contact is not sufficient to trigger a diffusion process. Rather multiple exposures are needed before contagion occurs. Such complex contagion can certainly be assumed for the decision to attend school classes or to dedicate at school.

Several studies with complex contagion processes also revealed that network structure affects complex contagion differently [13]. While weak ties or single connections between otherwise unconnected components play a major role in the spreading of simple contagion, they appear to impede the diffusion of complex contagion

This indicates the existence of those effects on other individual behaviors of children and adolescents such as “commitment to school education", substance use, or sport. Marques [27] reveals the vast differences between the social networks of the poor and those of more wealthy people. Considering the above, this further encourages

the modeling and simulation of social network effects in order to understand social phenomena and to guide political decision-making. Related research has also been conducted in children and adolescent networks. For instance, roles of nodes within a network of school children have already been identified [28] and diffusion of social norms and harassment behavior in adolescent school networks have been empirically studied and evidenced [29].

The Coordination-Game Mechanism. It then has been reasonably shown that behavior, norms, information and opinions flow within social networks of adults and children. Approaches to model this diffusion come for example from the field of Social Psychology, like the concept of Social Influence Network Theory [30] from Friedkin. Another approach is the modeling as a Coordination-Game [15], [20]. Those models may be considered as advanced threshold models [31]–[33] that incorporate social network structure instead of simple crowd behavior. The Coordination-Game as implemented in [20] is characterized by the assumption that individuals benefit when their behavior matches the behavior of their neighbors in the network. Hereby a node within a network can adopt one of two behaviors A or B. The node receives pay-off a when equaling her behavior with a neighbor that adopts $behavior A$. b respectively denotes the pay-off a node receives when both, her and her neighbor adopt $behavior B$. When choosing different behaviors, nodes receive a pay-off of 0 (other implementations may introduce negative pay-offs for non-compliance). The total pay-off for each node can accordingly be calculated as presented in 2.1 and 2.2. Here P_i^a denotes the total pay-off for node i from choosing behavior A (respectively behavior B for P_i^b), d_i denotes the degree of node i and n_i^a (same for n_i^b) denotes the number of neighbors of node i adopting behavior A (respectively behavior B).

$$P_i^a = an_i^a \tag{2.1}$$

$$P_i^b = b(d_i - n_i^b) \tag{2.2}$$

This determines that the best strategy for node i is to choose behavior A if $an_i^a \geq bn_i^b$ and behavior B otherwise. Rearranging the inequality in 2.3, we get:

$$r \geq T \text{ with } T = \frac{b}{a+b} \text{ and } r = \frac{n_i^a}{d_i} \tag{2.3}$$

In the absence of knowledge of the individual pay-offs a and b, a global threshold T may be found experimentally, as shown in the remainder of this Chapter.

Coordination-Game mechanisms have been applied to a wide range of network Models. The principles of those mechanisms are simple, easy to implement and therefore particularly intuitive. This makes them quite interesting for modelers with the intention to create verifiable models that are accepted also in other fields. In order to illustrate how to verify if a Coordination-Game mechanism may apply to the

spreading process in question, the procedure is subsequently described using two exemplary data-sets.

2.2.2 Data-sets - Complex Contagion

The experiments are performed on two data-sets, both contain information about adolescent friendship ties, as well as about different types of behavior.

(*i*) The first data-set stems from the study *"Determinantes do desempenho escolar na rede de ensino fundamental do Recife"* [19]. The survey was conducted by Fundação Joaquim Nabuco (FUNDAJ) in 2013, gathering data from more than 4000 pupils in public schools in the north-eastern Brazilian city Recife. Those data contain among others the social network of the pupils and their performance in the subject maths at the beginning and at the end of a school year. Children were asked to nominate their 5 best friends. In this way, a network containing 4191 students was generated. However, 573 students that did not nominate any friend within their class were removed from the data-set, leading to a total number of 3618 vertices. Since students could not nominate friends outside their school, the network is subdivided into 219 clusters from 122 schools.

(*ii*) The second data-set is a selection of 50 girls from the social network data collected in the *Teenage Friends and Lifestyle Study* [34]. Here the friendship network, as well as behavior in sports and substance use of students from a school in Scotland were surveyed. The survey started in 1995 and continued for three years until 1997. Students were 13 years old when the study started. The study counted 160 participants of whom 129 participated during the whole study. The friendship networks were surveyed asking the pupils to name up to twelve friends. Pupils were also asked to report their behavior related to sports and smoking as well as alcohol and cannabis consumption. The question about sporting activity assessed if the pupil regularly took part in any sport or went to training for sport out of school (e.g. football, gymnastics, skating, mountain biking). The school was representative of others in the region in terms of social class composition. Though there are alternative data sources such as the study *Network and actor attributes in early adolescence* [35], the described excerpt of the *Teenage Friends and Lifestyle Study* provides sufficient network complexity for the presented analysis.

2.3 Complex Contagion of of Behavior Modeled as a Coordination-Game

The imitation of behavior of neighbors within the friendship network is modeled according to the Coordination-Game as presented in Section 2.2. As indicated in Section 2.2, no information is available about possible pay-offs a and b or eventual costs of transition. Hence the threshold T shall be found experimentally. This means that a vertex within the network changes her state over time depending on the state

of her neighbors. For simplicity, the vertices may adopt one of two different states according to the investigated behavior. Hereby one state indicates that the vertex adopted behavior A, the other possible state indicates the adoption of behavior B. For each iteration, the current ratio r_i is being calculated. Here a_i denotes the number of neighbors of node i that adopt behavior A and n_i denotes the total number of neighbors of node i.

$$r_i = \frac{a_i}{n_i} \tag{2.4}$$

If the perceived ratio r_i is higher than the global threshold T and the state of node i is B , the node changes her behavior towards behavior A. Conversely, if r_i is below T and node $i's$ behavior is A, she changes her behavior towards B.

One challenge, a modeler may encounter when implementing experimental set-ups for Coordination-Game mechanisms, is naturally the variety of data-sets. As an example, it is hereinafter demonstrated how a simple behavior spreading mechanism can be adapted to different data representations and scopes. As explained above, two quite distinct data-sets are subsequently reviewed as an instance. In accordance with the focus of this thesis, the subsequently applied data-sets are analyzed with respect to questions concerning the spreading of behavior in networks. For the FUNDAJ data-set, it is drawn on the behavior "commitment at school" with the aim to create a good representation of the processes that trigger pupils to dedicate themselves at school. The second data-set contains information about physical activity as well as substance use, which is why the focus of the implementation lies here on those behaviors.

2.3.1 Implementation of the Coordination-Game - FUNDAJ Data-Set

The only information available for more than one moment in time of the FUNDAJ survey is the mark of the pupils in the subject maths for the beginning and the end of the year. Although marks are not behavior in themselves, they stem among others from individual behavior such as doing homework, paying attention, studying frequently, etc.. Marks are therefore considered a good indicator for the behavior *commitment at school*. They are represented as numeric values between 0 and 100. In order to differentiate between two behaviors, students are classified as *good students* or *bad students* according to their mark. Students whose mark lies below the threshold *tm* are thereby classified as bad students and vice-versa. The setting of *tm* defines hereby the number of *good students* (positives) and *bad students* (negatives) and hence affects heavily if nodes are predominantly connected to positives or negatives. High values for *tm* generate large numbers of *bad students* and smaller numbers of *good students* and vice versa. The ratio r_i from Equation 2.4 is being calculated for each student at each iteration of the simulation. If required, the mark for the next time step m_{i+1} is being multiplied by the factor $1 + f$ in order to alternate the state of the node:

$$m_{i+1} = m_i(1 + f) \tag{2.5}$$

Parameter T sets the affinity of the nodes to change behavior. Thus, depending on the proportions of positives and negatives, it either yields a volatile or a stable system. Adaption parameter f also influences the stability of the system, where volatility increases with increasing values of f.

2.3.2 Implementation of the Coordination-Game - Scottish Teenage Friends and Lifestyle Study

The second data-set contains information about four different behaviors, which are practicing sports, drug (cannabis) use, alcohol use, and smoking behavior. Characteristic values differ slightly for the distinct behaviors, as there are for example two increments representing the intensity of sports but four increments for drug use intensity. For the implementation of an easily accessible simulation model, it may be helpful to simplify as much as possible. Thus, the characteristic values have been classified in order to obtain a simplified two-status situation. Table 2.1 presents the characteristic values and their classification as *behavior A*, all other values are accordingly classified as *behavior B*.

TABLE 2.1: **Classification of Characteristic Values for Behavior**

behavior	Characteristic values	Class. as behavior A if:
Sports	1 (non regular); 2 (regular)	≥ 2
Drugs	1 (non), 2 (tried once), 3 (occasional) and 4 (regular)	≥ 2
Alcohol	1 (non), 2 (once or twice a year), 3 (once a month), 4 (once a week) and 5 (more than once a week)	≥ 2
Smoke	1 (non), 2 (occasional) and 3 (more than once a week)	≥ 2

In contrast to the FUNDAJ-data, the representation of behavior by discrete values requires a slightly different imitation process. Hence, for the experiments with the Scottish school network data-set, the state transition of vertexes is discrete. This means that if a vertex changes state, it respectively raises the behavior value by 1 when aiming at adopting behavior A or, decreases the behavior value by 1 if it aims to adopt behavior B.

Information is available for three consecutive years. Hence, the starting value for each vertex in the Coordination-Game is its behavior in year one. The quality of the simulation is measured by comparing the state of the simulation after a certain number of iterations with the state of the real system after two years, here referred to as *benchmark t+1* or after three years, denominated as *benchmark t+2*.

Moreover, this data-set imposes another challenge to the modeler, since the friendship network of the girls in the study has been surveyed for each of the three years, the study lasted. This yields a somehow dynamic network with the three slightly different networks g_1 from the first survey, g_2 after one year, and g_3 after two years.

This implicated for the simulation that the neighbors that a vertex considers for the calculation of her state vary for different years. To achieve an accurate and comprehensive simulation, the network dynamics need to be incorporated into the simulation.

For the presented case, the network used to define the adjacent vertices of a node has been changed after completing 50% of iterations. In this case, experiments indicated that network combination of g_1 as representation for the friendship network in the period between year 1 and year 2, and g_2 representing the friendship network in the period from year 2 to year 3, outperformed the results for network combination (g_2, g_3). Therefore it is assumed that the more appropriate network combination is the former. Thus, experiments and results presented in the remainder of this paper refer to network combination (g_1, g_2).

2.4 Experiments with the Coordination-Game Mechanism

Creating simulation models based on data from existing surveys is a frequent task for those whom aim at creating simulation models for behavior spreading. Data is easy to acquire and does not require long preceding surveys or other preliminary research. However, it comes with the cost, that survey design does usually not match exactly with the aim of the simulation model.

In absence of fitting accompanying information, aligning simulation data with given data is often the only way to calibrate model parameters and assess the quality of the created simulation model. Also in the presented cases, this approach has been applied.

2.4.1 Experimental Setup

Experiments were run for the two Coordination-Game settings with varying parameters in order to find a parameter setting that leads to plausible results. As for the simulation with FUNDAJ-data, the simulation was conducted with all combinations of the parameters T(global threshold) and f(adaption parameter) for $T, f \in [0, 0.2, 0.4, 0.6, 0.8, 1]$ and tm (classification of marks) with $tm \in [20, 40, 60, 80]$.

For simulations with the Scottish data-set the parameter T was set to values $T \in [0.0, 0.05, 0.1, \cdots, 1.0]$.

2.4.2 Quality Measurement

In order to assess the quality of the respective simulation, four distinct quality measures were applied: (*i*) match quality, (*ii*) ROC-curves (*iii*) graph-based quality measures, and (*iv*) average estimation error.

(*i*) The most intuitive measure for the simulation quality is to compare the state of each vertex v_s after a certain number of simulation iterations with her state in reality

v_r in *benchmark t+1* or *benchmark t+2*. Hereby, *match* denotes the case when $v_s = v_r$ and accordingly the case $v_s \neq v_r$ is denoted as $miss - match$. This quality measure is named $match - quality$ and denoted as q for the rest of this Chapter. The match quality q of the simulation can then be assessed as in 2.6, where n denotes the total number of vertices:

$$q = \frac{\sum_{i=1}^{n} match_i}{n} \tag{2.6}$$

However, for skewed attribute distributions, this measure favors estimates with high numbers of positive or respectively negative estimates and hence fails to mirror the quality of the simulation when the distribution of attributes is skewed.

(*ii*) The ROC-metric [36] sets the number of true positives (*Recall*) in relation to the number of false positives (*Fallout*). *Recall* is the ratio of correctly estimated positives values, the $true - positives$ and the total number of positive values n^p. *Fallout* denotes the ratio between wrongly estimated positive values $false - positives$ and the total number of negative values n^n.

$$Recall = \frac{true - positives}{n_p} \tag{2.7}$$

$$Fallout = \frac{false - positives}{n_n} \tag{2.8}$$

The ROC-curve displays respectively *Recall* values for each simulation on the ordinate and *Fallout* values on the abscissa. Values above the diagonal of the graph indicate the existence of a signal and values below the diagonal may be interpreted as noise. Thus, this metric provides a clearer picture of simulation quality. Best estimates can be found mathematically maximizing the $Youden - Index$ [37] y as presented in 2.9.

$$y = Recall - Fallout \tag{2.9}$$

(*iii*) For global analysis, it might not be necessary to simulate the state of each vertex correctly, as long as the system state can be predicted adequately. Thus, as a third quality measure, *behavior distribution in friendship-patterns* was implemented. Hereby friendship patterns in the network are defined using a modified version of NEGOPY [38]. According to NEGOPY, vertex types are defined as isolate, dyad, liaison, and group member. As this example deals with undirected networks, no tree-nodes are classified. Thus, the nodes are divided into four subgroups ($j = 4$). According to Richards [38], an isolate is an individual with maximum one friend. Two persons connected only to each other are denoted as dyad. Liaisons are individuals with more than 50% connections to members of different groups. Liaisons can also be nodes that are mostly connected to other liaisons and with less than 50% links to group members. A composition of minimum three individuals is referred to as group if the individuals share more than 50% of their linkage, build a connected component and stay connected if up to 10% of the group members are removed.

For measuring the quality, the number of positive vertices n_k^p in each friendship-pattern class k is calculated after each iteration of the simulation. Subsequently, the error e_k is calculated as the difference between n_k^p of simulated and real values. The $average-error$ e denotes the weighted average error of the simulation and n_k the number of vertices in friendship-pattern k :

$$e = \frac{\sum_{k=1}^{j} e_k n_k}{\sum_{k=1}^{n} n_k} \tag{2.10}$$

(*iv*) The average estimation error ϵ assesses the average difference between simulated values for behavior and real behavioral outcomes. Here n denotes the total number of nodes in the simulation, while the difference between simulation and reality for node i is represented by ϵ_i.

$$\epsilon = \frac{\sum_{i=1}^{n} \epsilon_i}{n} \tag{2.11}$$

2.4.3 Results

This subsection contains the results from the experiments presented earlier. First results for the experiments with FUNDAJ data and subsequently results for experiments with the Scottish data-set are presented.

Results - FUNDAJ Data-Set

Figures 2.1, 2.2 and 2.3 illustrate the results for simulations with FUNDAJ data for 15 iterations. Figure 2.1 contains ROC-curves for the experimental results with varying settings of tm, T, and f. For each investigated value of mark threshold tm, the Figure illustrates an individual ROC-curve. The dashed lines indicate the ROC-level of the respective setting for tm before starting the simulation. Thus only parameter settings leading to ROC-values situated above the respective dashed line can be considered as settings that improve the quality of the simulation. The colored lines in Figure 2.2 represent the development of quality indicators q and e for distinct parameter settings and also indicate the average estimation error ϵ during the runtime of the simulation.

The results with the highest $Youden-Index$ in simulations with FUNDAJ data-set are indicated by arrows pointing from the respective parameter settings for mark-threshold tm, global threshold T and adaption parameter f in parentheses as (tm,T,f) in Figure 2.1. The results for q, e and ϵ of those most promising parameter settings are presented in Figures 2.2 and 2.3.

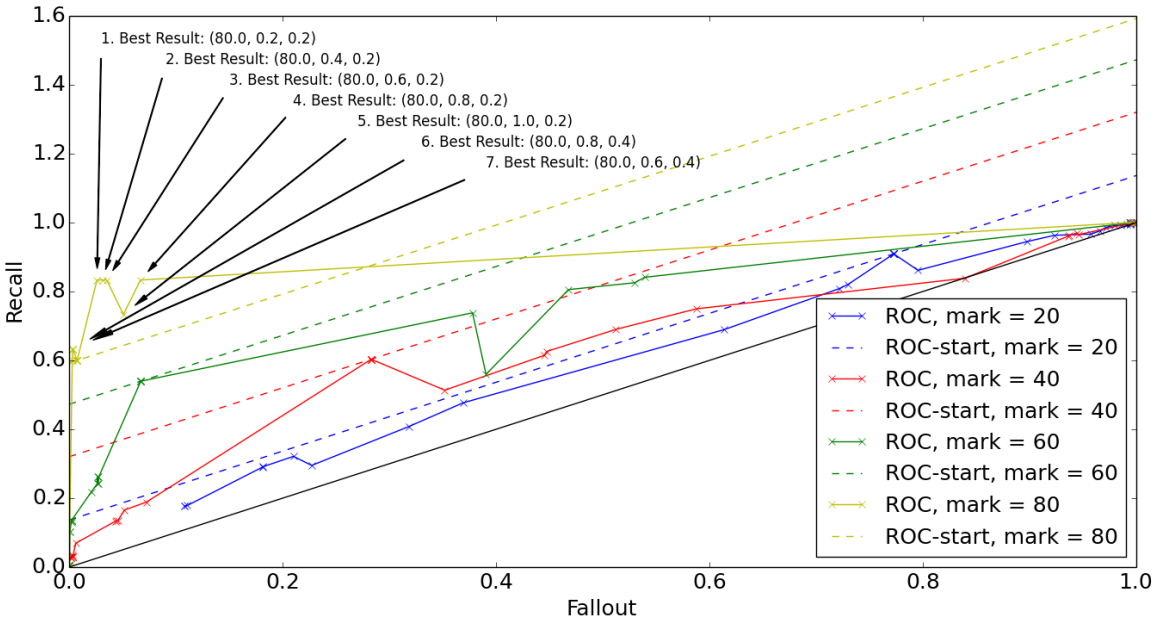


FIGURE 2.1: **ROC-Curve for Coordination-Game Simulations with FUNDAJ Data - 15 iterations.** The Figure presents the ROC-Curves for experiments with varying threshold of marks tm, classifying the pupils as *good students*, if their mark is greater than tm or *bad students* if their performance is below tm and for varying settings of T and f, as pointed out in parentheses (tm, T, f). $Recall = \frac{true-positives}{n^p}$; $Fallout = \frac{false-positives}{n^n}$. The results with the highest $Youden - Index$ are indicated by arrows pointing from the respective parameter setting. Simulations with those settings provide ROC-levels above the ROC-levels of the respective setting for tm before starting the simulation. Hence they indicate the existence of a signal, rather than a random process.

The more detailed analysis of the five parameter settings that were performing best in ROC-curve analysis in Figure 2.2 yields increasing e and increasing estimation error ϵ while q continuously decreases.

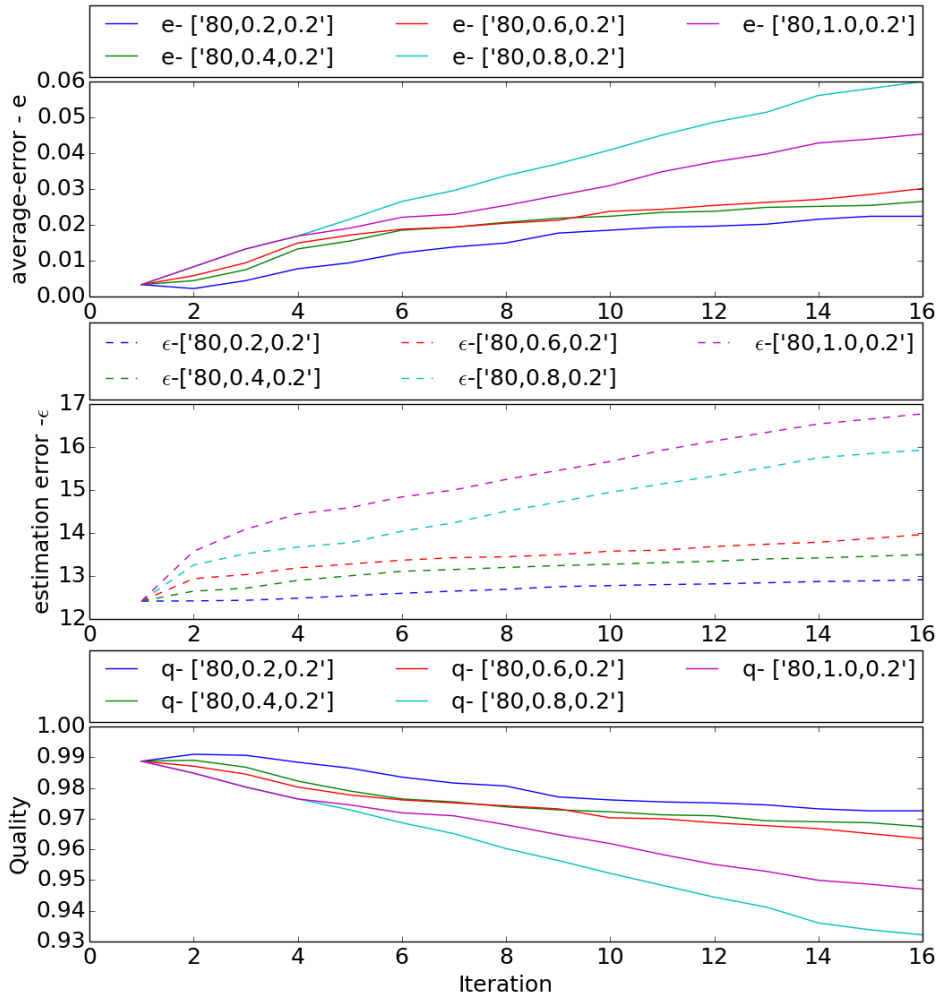


FIGURE 2.2: **Analysis for Coordination-Game Simulations with FUNDAJ Data.** The Figure presents results for 15 iterations of simulations with the best performing parameter settings from Figure 2.2 with the contagion model-f = 0.2. Indicators for the quality of the simulations evolve negatively throughout the run-time

However, as presented in Figure 2.3 the second-best performing parameter settings from ROC-curve analysis lead in general to decay of e and significant growth of q whereas at least one setting (80,0.2,0.4) also decreases estimation error ϵ slightly.

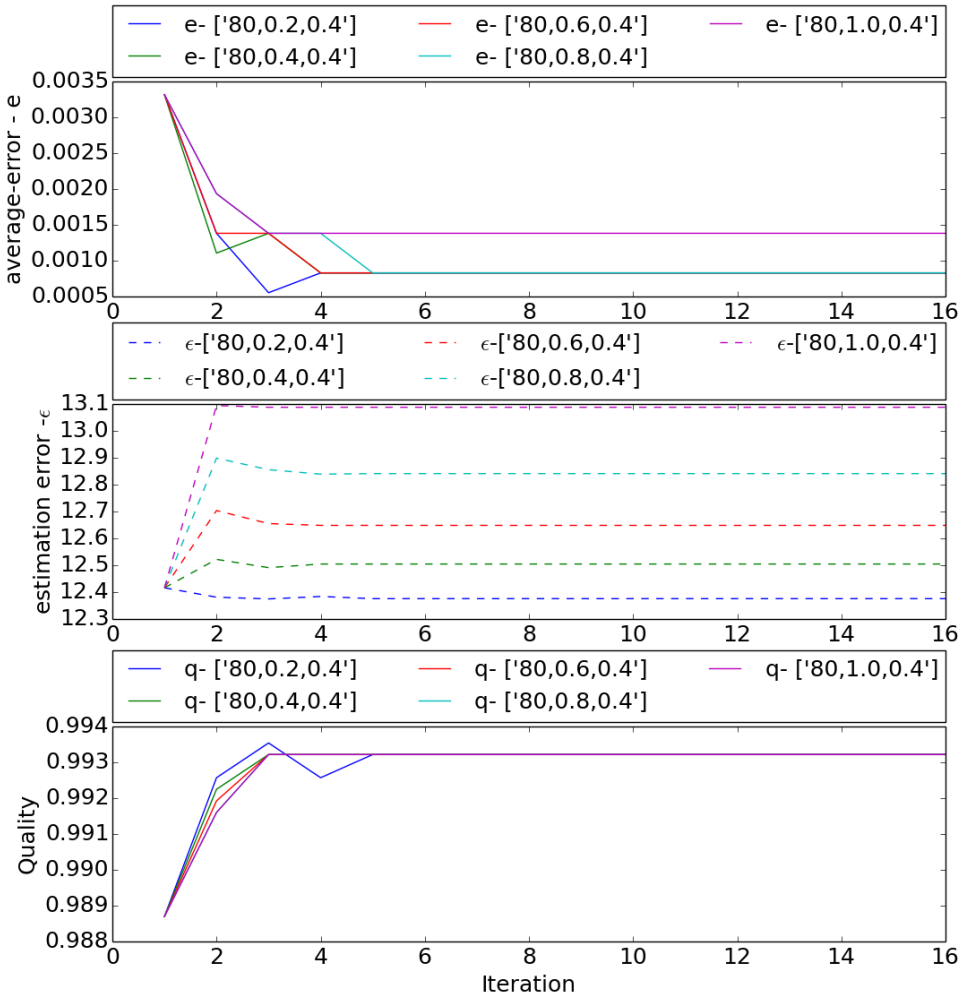


FIGURE 2.3: **Analysis for Coordination-Game simulations with FUNDAJ data.** The Figure presents results for 15 iterations of simulations with the second-best performing parameter settings from Figure 2.2 with the contagion model-f = 0.4. Indicators for the quality of the simulations evolve positively throughout the run-time

Results - Scottish Teenage Friends and Lifestyle Study

Figures 2.4, 2.5, 2.6, 2.7 and 2.8 illustrate the results for experiments with the Scottish data-set for 50 iterations for each of the investigated behaviors. The solid lines in Figure 2.4 illustrate the *Recall-Fallout* relation for varying parameter settings and for different behaviors. The black diagonal line in this graph indicates *Recall-Fallout* ratios that represent random processes, while the dashed lines indicate the ROC-level of the start situation. Recall that only parameter settings leading to ROC-values situated above the respective dashed line can be considered as settings that improve the quality of the simulation.

Since experiments with Scottish data were run with two different networks as explained in Section 2.3, analysis of q, e and ϵ in Figure 2.5, 2.6, 2.7 and 2.8 contain

blue lines, indicating the values calculated in relation to *benchmark t+1* and red lines, representing the results calculated in relation to *benchmark t+2*. ROC-curve for the simulation of diffusion of the behavior *sport* in Figure 2.4 is very close to the diagonal of the graph, indicating that the simulation is rather a random process. Furthermore, ROC-values cannot reach the ROC-level of the baseline indicated by the dashed line. However, there are two values for t that yield ROC-values above the diagonal of which $t = 0.55$ generates the most promising results. Hence, q, e and ϵ development are analyzed over the whole run-time in Figure 2.5. It is observable, that e in $t + 1$ indicated by the blue line decreases significantly until the 25th iteration, which is when the network g_t is replaced by network g_{t+1}. After the 25th iteration, e in $t + 2$ decreases heavily. ϵ decreases slightly for benchmark $t + 2$ but increases if compared to benchmark $t + 1$. Although decreasing for the first five iterations, q remains stable during the following 20 iterations and slightly improves after 25 iterations.

ROC-curve for smoking behavior in Figure 2.4 yields positive results for t 0.35, 0.4, and 0.45, significantly outperforming the initial ROC-value indicated by the dashed line. A deeper examination of q, e and ϵ development during run-time in Figure 2.6 shows that as compared with benchmark $t + 1$ neither q, nor e or ϵ develop positively. Though, compared with benchmark $t + 2$ a strong improvement of q, as well as a significant decrease of e and a slight decrease of ϵ is observable.

The t values indicated by the ROC-curve for Alcohol-use in Figure 2.4 do not reach the initial ROC-level and yield decreasing q and increasing e until the underlying network is changed after 25 iterations, initiating a slight improvement of those values for both benchmark values as presented in Figure 2.7. Nevertheless, q never reaches a value higher than the start value, also e does not drop under its start value and ϵ remains on an equal level. ROC-curve for Drug-use in Figure 2.4 yields positive results for t 0.35, 0.4 and 0.45, slightly exceeding the initial ROC-value. Figure 2.8 presents decreasing e and ϵ, as well as increasing q over the run-time for benchmark value $t + 2$, while all quality measures develop negatively for benchmark $t + 1$.

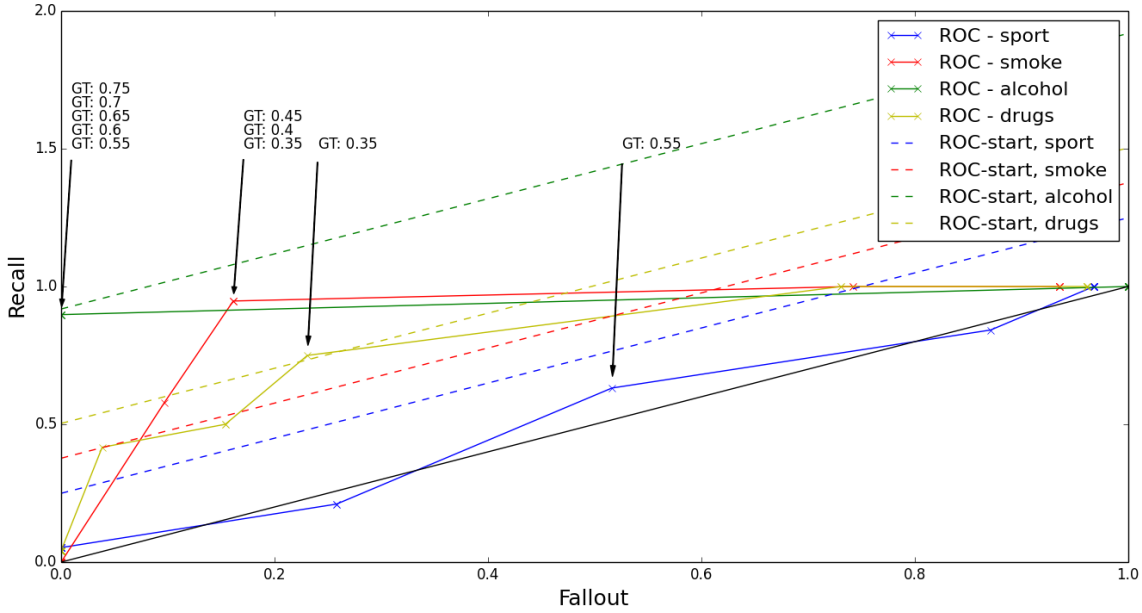


FIGURE 2.4: **ROC-curves for Coordination-Game Simulations With Scottish Data-Set.** The Figure presents results for 50 Iterations of the contagion model for for varying behaviors and for varying settings of T. $Recall = \frac{true-positives}{n^p}$; $Fallout = \frac{false-positives}{n^n}$. Simulations with the behaviors $sport$, $smoking$ and $drug - use$ yield ROC-Levels that outperform the initial ROC-value indicated by the dashed line and hence indicate that the simulations possess predictive power.

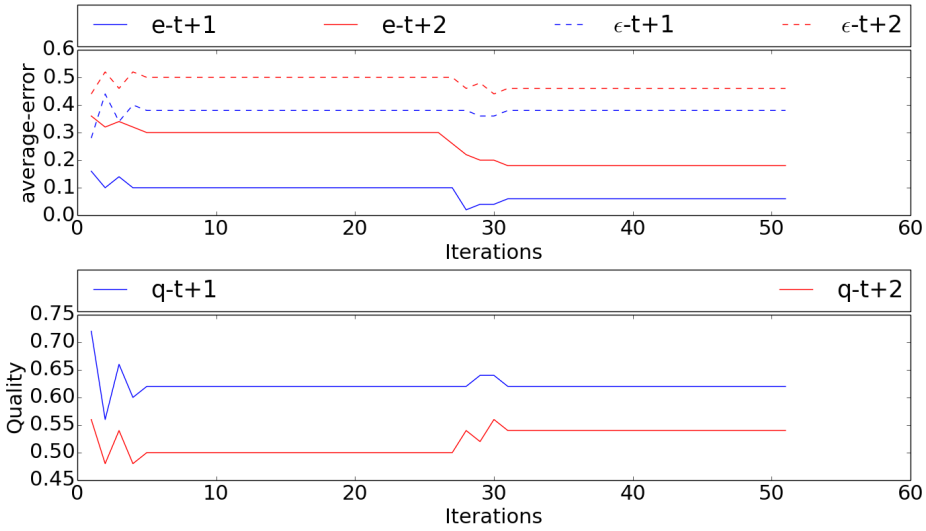


FIGURE 2.5: **Analysis for Coordination-Game Simulations With Scottish Data-Set.** The Figure presents results for 50 Iterations of the contagion model for the behavior Sport - $T = 0.55$. It is observable that performance indicators evolve positively particularly for simulations with the network g_{t+1}.

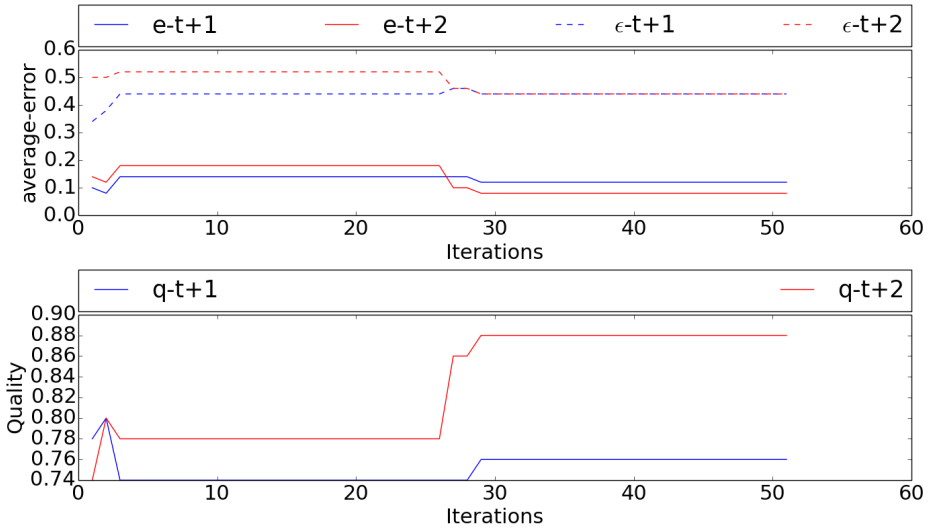


FIGURE 2.6: **Analysis for Coordination-Game Simulations With Scottish Data-Set.** The Figure presents results for 50 Iterations of the contagion model for the behavior Smoking - $T = 0.45$. Compared with benchmark $t + 2$ a strong improvement of q, as well as a significant decrease of e and a slight decrease of ϵ is observable.

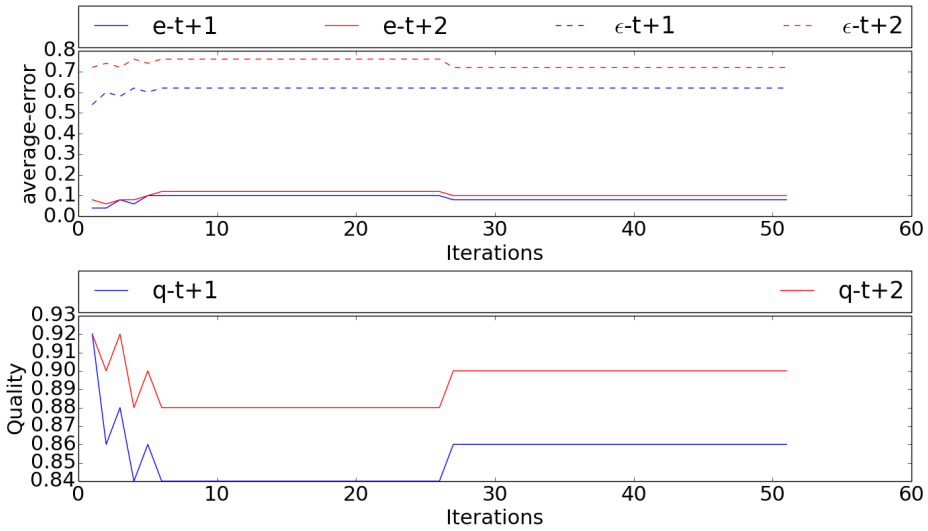


FIGURE 2.7: **Analysis for Coordination-Game Simulations With Scottish Data-Set.** The Figure presents results for 50 Iterations of the contagion model for the behavior Alcohol use - $T = 0.65$. Note that q never reaches a value higher than the start value, also e does not drop under its start value and ϵ remains on an equal level.

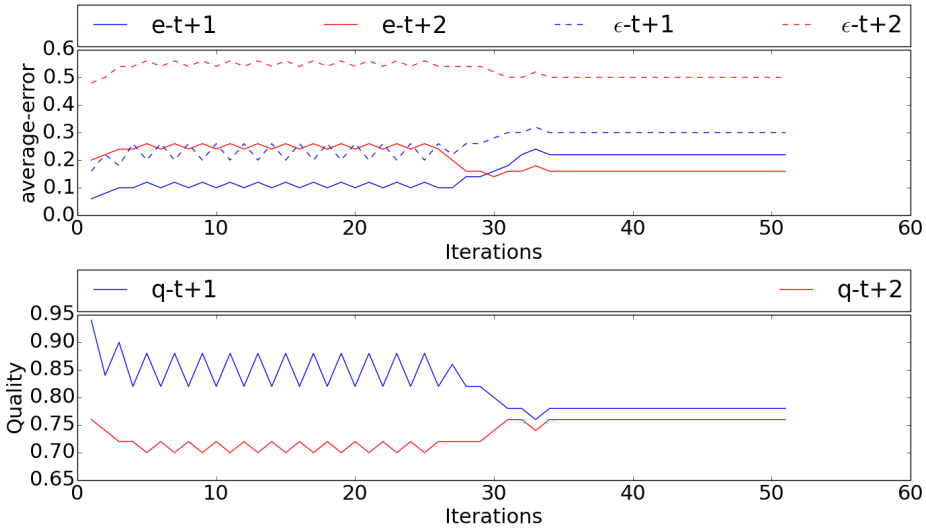


FIGURE 2.8: **Analysis for Coordination-Game Simulations With Scottish Data-Set.** The Figure presents results for 50 Iterations of the contagion model for the behavior Drug-use - $T = 0.35$. The Figure yields decreasing e and ϵ, as well as increasing q over the run-time for benchmark value $t + 2$ and hence confirms the promising performance of the model for this behavior.

2.5 Discussion - How Did the Model Perform?

This Section describes and discusses the outcomes of the experiments conducted with the FUNDAJ data-set and the data-set from Scottish Teenage Friends and Lifestyle Study.

2.5.1 FUNDAJ Data-Set

As pointed out in Section 2.4, the parameter setting (80,0.2,0.4) performs best as under this setting average-error e is being more than halved (approximately 75%). For this setting also match quality q increases slightly, ϵ shows a small decay, and $Youden - Index$ improves. This indicates that the setting reasonably approximates the real system state. However, simulation is not very adequate in estimating individual behavior. Thus it might be argued that diffusion of marks can be reasonably modeled as a Coordination-Game if the researcher is willing to disregard individual states and is interested in the global state of the network instead. Results further indicate that 15 iterations under the given parameter setting are well suited to approximate one school year.

2.5.2 Scottish Teenage Friends and Lifestyle Study Data-Set

Simulating the Coordination-Game spread for behavior *sport* with $T = 0.55$ yields a relatively small $Youden - Index$ and cannot improve the ROC-level of the initial situation. However, the development of $average - error$ for benchmark t+1 and t+2

yields improvement of the overall state of the network, while decreasing $match - quality\ q$ and increasing ϵ for both benchmarks. Although improving the estimation of the general network state, setting $T = 0.55$ cannot improve the estimation quality and can therefore not be considered a good setting for T.

As for simulating the spread of behavior *smoke* throughout the given network, strong evidence for the suitability of parameter T = 0.35, 0.4, and 0.45 has been found in the ROC-curve. Run-time analysis of e, ϵ, and q indicate that the parameter setting when run on network g_1 cannot reproduce the contagion process during the first year, since the former grows while the latter declines for the first 25 iterations. However considering benchmarks $t + 2$ and network g_2, all three quality indicators support the hypothesis that spreading occurs as a Coordination-Game with $t = 0.35, 0.4$ or 0.45. Recall that children were around age 13 when the study started, this discrepancy may be explained by the nature of the behavior smoking, which probably has a higher attraction to children aged 14 to 15 than to children aged 12 to 13.

Similar but not as striking evidence can be found when examining behavior $drug - use$. As $drug - use$ has been explicitly surveyed as the use of cannabis, this seems coherent, since tobacco use does commonly precede cannabis use. Conversely, for the behavior $alcohol - use$, results are not clear. ROC-curves indicate that parameter settings yielding reasonable estimates of the real situation exist. Yet, run-time analysis of those cases show that those promising parameter settings do not lead to an improvement of the estimation. Hence, it is argued that for alcohol-use there is no evidence that contagion of behavior can be modeled as a Coordination-Game within the given data-set. This might also be related to the age of the students, since parents' influence might be stronger during this period. Additionally due to the restriction of available data to female students the lack of spreading could be gender related.

2.6 Conclusion - Complex Contagion of Behavior

This Chapter presents the implementation of a Coordination-Game mechanism for simulating the spreading process of behavior throughout social networks. Simulation has been run on two different data-sets, the FUNDAJ study with school children from the metropolitan area of Recife and the Scottish friends and lifestyle study. The spread of behavior "commitment to school education" represented by the marks of the pupils in the FUNDAJ study, as well as the behaviors "Substance use" for tobacco, drugs and alcohol and the behavior "practicing sports" as surveyed in the Scottish data-set have been investigated.

Here, good indications have been found that a Coordination-Game mechanism underlies the spread of behavior "commitment to school education" as well as "smoking" and "drug-use". Comparable evidence for the behavior "alcohol-use"could not be found. Results for behavior "practicing sports" were not clear. Similar studies

found that the Coordination-Game mechanism also underlies the behavior "alcohol-use" [39], [40]. The missing evidence for this behavior in this work may stem from the nature of the data-set, since participating individuals were below 16 years of age until the end of the survey. Moreover only female pupils participated. Since male adolescents are more susceptible to early alcohol-use, this could be an explanation for the lack of evidence, for that particular aspect.

This Chapter serves as a first step in simulating the Complex Contagion of behavior throughout social networks, since it provides evidence that (1) there is an underlying game-environment for the agents within the social system and (2) that it can be modeled as a Coordination-Game. However, the players of this game, the bounded rational agents [41] might be equipped with decision finding mechanisms that better approximate human decision making. Though driving the social systems from a real start situation towards the state in reality after one or respectively two years, the investigated deterministic mechanism still leads to a considerable difference between the real and the simulated system. Hence, it seems that a deterministic mechanism is not fully capable of simulating human bounded rationality and the lack of information humans face within their decision process. Besides this, eventual noise within the data and external influences may not be represented by a deterministic mechanism. Further, the problem of missing data, particularly missing links has not been addressed yet. The following Chapters therefore deal with the missing data problem and more elaborated decision mechanisms for the individual agents. Thus, the following Chapters aim at a better representation of human decision making within a Coordination-Game setting. In addition, the following Chapters and Sections expand the binary behavioral variable which is an extreme simplification for the on continuous scales measured nuances of human behavior such as sports activities, drug- and alcohol consumption or school performance.

Chapter 3

Dealing with Missing Data: Solutions to the Boundary Specification Problem in Social Network Surveys

3.1 Introduction to the Missing Data Problem

So far, it has been examined how suiting spreading models for the process of complex contagion of behavior through social networks may be identified. To create a coherent simulation of such processes however, potentially missing information within the available data need do be considered. This Chapter addresses the challenge that network surveys of close contact networks such as close friendships and, to a lesser extent, also data gathered from online sources, generally suffer from missing data. Missing data may stem from a survey design, restricted to a certain type of participants, to certain relations between participants or from the focus on certain places such as schools, classes, companies or offices. This issue is also referred to as the "Boundary specification problem" [42]. Another persisting problem in social network surveys is the restriction of the number of contacts to choose. This "Fixed choice effect" [43] appears when survey participants are asked to nominate a certain number of contacts. Surveys including a large number of participants may not be capable of capturing relations that could exist between survey participants from distinct places or of different entities, such as pupils from different schools or employees from different companies. Hence, those large social network surveys often appear to consist of many disconnected components. The availability of a high volume of network data creates the possibility to investigate local network effects based on a considerable number of empirical data. However, in large societies some trends, behaviors or norms may emerge in one part of the society and then diffuse to other parts. There may be local circumstances in one component that prevent individuals within that component from adopting whatever is spreading throughout the network and hence impede it from becoming a global trend. Moreover, there is evidence that network structure and network heterogeneity heavily affect global

behavioral outcomes of network diffusion processes [44], [45]. Therefore, in order to understand and simulate global network effects on a population or society level, it is desirable to develop mechanisms that coherently estimate possible connections between those isolated components or between network components from distinct surveys. Hereby it may be possible to create a global network that features real world properties. The data that build the ground of the practical example accompanying the theoretical findings of this work stem from the study *"Determinantes do desempenho escolar na rede de ensino fundamental do Recife"* [19]. In this study both effects, "Boundary specification problem" and "Fixed choice effect" are present and need to be addressed in order to create a holistic simulation model.

This Chapter addresses the issues of missing data, investigating the performance of techniques from the fields of network generation [46] and link-prediction [47] as well as a combination of both, in filling the informational gap that frequently occurs between isolated components in social network surveys.

3.2 Background - Missing Data

The following literature review reveals how the problem of missing information in social network data has been tackled from different disciplines. Missing information may be classified as missing completely at random (MCAR) if the missing value does neither depend on other missing values, nor on observable values, missing at random (MAR) if the missing value does not depend on other missing values and missing not at random (MNAR) when the reason for the missing information can be found in the information itself [48]. It is found that there are several fairly well performing methods to deal with both, (i) the total absence of information about links between individuals in the network (MNAR), and (ii) the randomly missing information about links within a network (MAR) or (MCAR). Nevertheless, to the best of the authors knowledge it has not been studied yet how systematically missing data between isolated components from social network surveys (MNAR) may be inferred (or imputed) in order to enable simulations on a global network model. This type of missing data is denoted as missing not at random (MNAR), since the missing information about an existing link between two nodes may depend on the nominated friend node: if the friend is included in the boundary specifications, the link is being recorded, if not the information is missed out.

3.2.1 Missing Data in the Social Sciences

Traditional solutions to the missing data problems "Fixed Choice Effect" and "Boundary Specification Problem" are applied in survey planning, dealing with the careful definition of the survey group [42], [43], [49].

The “Fixed Choice Effect" seems to disturb assortativity measures and degree distributions which may explain the frequent deviation of those measures when comparing surveyed social networks to other known social networks [49]. The problem of missing data after completing data collection has been approached by social sciences mainly under the term imputation [50]. The set of applied mechanisms incorporates for example the estimation of missing reciprocal ties in directed networks (reconstruction) [51], the replacement of incomplete respondents by similar others (hot deck imputation) [50], or using the concept of preferential attachment (assortativity) [52].

3.2.2 Link-prediction in Complex Networks

The missing data problem for social networks is a recent issue for researchers dealing with large social networks from online sources. Here links may be omitted due to privacy restrictions, or missing because observed networks tend to be dynamic. The task to predict links to be established in the future or links that have been omitted due to other reasons is here called the “Link-prediction Problem" [47]. Unsupervised measures for link prediction build on the “similarity" of nodes in terms of network properties as for example the number of common neighbors or draw from common properties of social networks such as assortativity [47]. Other approaches are for example supervised random walks [53], methods based on community structure [54] or on mutual information [55]. Furthermore, the problem to predict links between individuals that are not part of the same data-set or platform has been successfully tackled using machine learning techniques such as classifier systems [56].

A special case of the “Link-prediction Problem" occurs when no previous data of the network is available, and the network structure is to be re-build based on other information about the nodes. This problem may arise in co-purchasing networks or recommendation networks where information about the nodes is available, but connections between them are omitted or simply not informed [57]. This special situation requires a different approach for link-prediction, since network based measures are not applicable due to the total absence of links. A well performing mechanism to tackle this problem is a two phase bootstrapping method [57]. Here a bootstrap probabilistic graph is being estimated from the node properties in a first step, assigning a probability for the existence of each possible link in the network. Subsequently network based measures are applied to the bootstrap probabilistic graph in order to reinforce the probability of links to exist. Finally, the researcher defines a probability threshold t so that all links with probability $p \geq t$ are estimated as existing links. For a review of link-prediction techniques see [58].

Similar to the aforementioned bootstrapping method, the benefits from joining information derived from the network structure and information stemming from the individual vertices have been recognized in recent work on link-prediction [59]. Here, the feature of dynamic networks is additionally addressed just as in related recent

publications dealing with temporal link-prediction of persistent relationships as opposed to link-prediction in networks with discrete events [60].

3.2.3 Network Generation

Scientists in the fields of Complex Network Sciences and Social Simulation have extensively studied models to create networks featuring characteristics that can be found in real complex networks. Especially in social networks there is consensus that models that generate social network representations should assure that generated networks feature limited size and heterogeneous right-skewed distribution of degree allowing for a "cut-off' for higher degrees in case of close relationships [61]. The networks should furthermore incorporate high clustering, low density, positive assortativity by degree, and short path lengths [62]. The literature about social networks of students suggests that pupils tend to have lots of contacts in their own classroom and fewer within other classrooms. Furthermore, this pattern may be found at larger scales i.e, many contacts within school, fewer contacts between schools [63].

Moreover, recent studies with location-based social networks suggest that on a global scale, distance matters for the likelihood of the existence of links between any two nodes [64]. Those studies further indicate that the relation between link-probability $P(l)$ and distance of any two nodes follows approximately a law $P(l) \sim d^{-\alpha}$, where the exponent α lies between 0.5 and 2 for different networks [65], [66].

An approach to generate complex networks has been the use of random graph theory [67]. However, those graphs failed to exhibit the scale-free property or to have right-skewed degree distributions [68]. Models to generate networks with more plausible features from scratch have been developed on the ground of preferential attachment, where the probability of a new vertex being attached to an existing vertex depends on the degree of the existing vertex [68]–[70]. Other approaches focus on local interaction of nodes and equip the models with mechanisms that represent human behavior such as inviting and visiting each other [71] or incorporate the idea of social proximity and agents moving and meeting within a "social space" [72].

The idea of social proximity is also an essential part of the social circle model for generating large artificial social networks for social simulations [62]. Here a "Social Reach" is defined, allowing the agents to connect to other agents within their "Social Reach", when the relation reciprocates. The "Social Reach" may be interpreted as a social distance, set for example by the number of common friends or the similarity of interests between two agents, but can also be interpreted as a physical distance like for example the distance between the domiciles of two individuals. It is established that this approach enables the researcher to create networks from scratch that exhibit characteristics that match the aforementioned network properties. Similar to this is the "Waxman model" [73] which does not define a fixed social reach, but employs an exponential decay model to create links depending on the local proximity of nodes.

3.3 Analysis: Problem Description and Link Prediction Approaches

3.3.1 Practical Example: Missing-Data in the FUNDAJ Data-Set

The problem context shall be illustrated using the example of the aforementioned large social network stemming from a survey among basic school pupils in the Brazilian city of Recife. The problem context is described in *"Determinantes do desempenho escolar na rede de ensino fundamental do Recife"* [19]. The study has the objective to estimate a linear hierarchic model in order to quantify the effect that schooling infrastructure projects have on the performance of school children. The survey was conducted by Fundação Joaquim Nabuco (FUNDAJ) in 2013, gathering data from more than 4000 pupils in public schools in the northeastern Brazilian city Recife. The data contain among others the social network of the pupils and their performance in the subject of maths at the beginning and at the end of the school year. Children were asked to nominate their five best friends. If the friends went to the same class, a questionnaire was sent to these students as well. In this way, a network containing 4191 students was generated. However, since the friendship nominations were only traced if the students were within the same class, the network is subdivided in 219 disconnected components from 122 schools. The schools are distributed over the districts of Recife according to population size of the respective district. Hence, very large districts are represented by more schools than smaller districts. Figure 3.1 illustrates the distribution of the isolated components throughout the city of Recife. The pupils of each school are represented by points , where pupils from the same school are assigned equal color. The location of the pupils is defined by a Fruchtermann-Reingold algorithm [74], centered around the location of their school on the map. The area of the graph is hereby given as n^2, where n denotes the number of pupils of the respective school. To provide better insight to the micro structures of this network, Figure 3.2 presents a close-up view of an extract of the network.

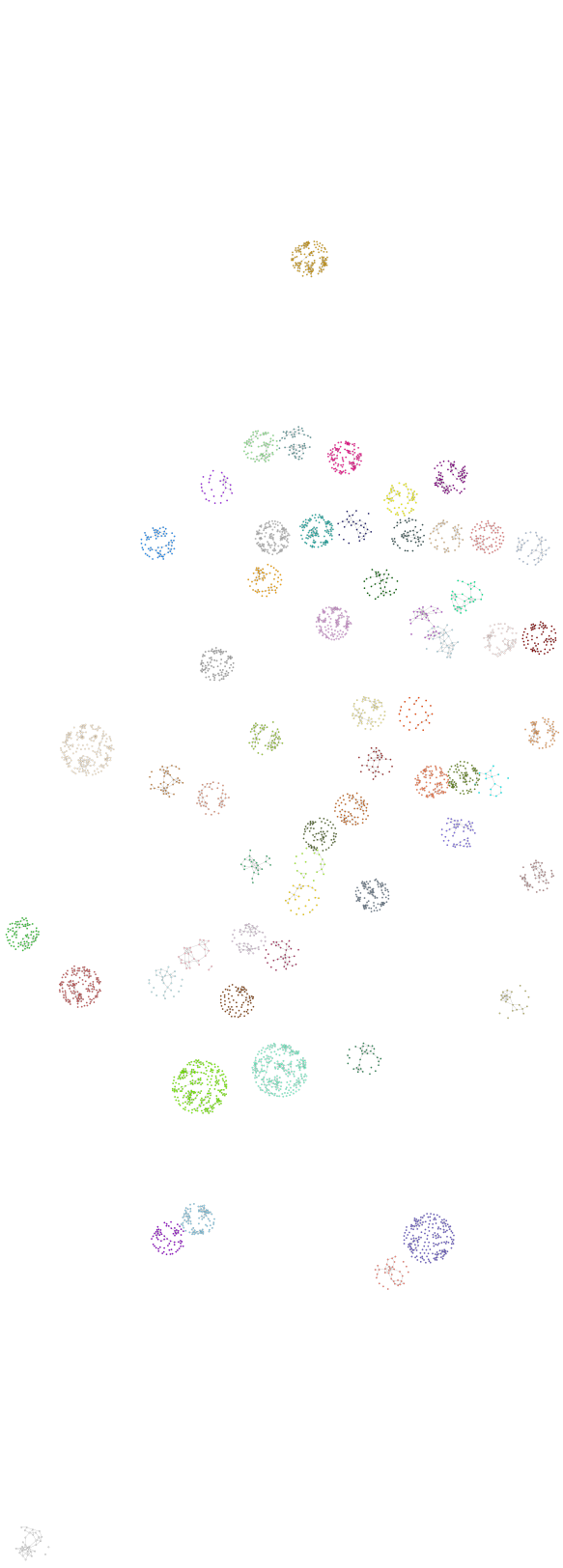


FIGURE 3.1: **School-clusters from Recife.** The Figure presents the individual pupils that participated in the FUNDAJ-Survey. Pupils are colored according to their school. Location of pupils is assigned by a Fruchtermann-Reingold algorithm [74] centered around the location of their school within the city of Recife. Grey lines indicate friendships between pupils as registered by the survey. As social networks where solely surveyed within schools, isolated components appear for each school.

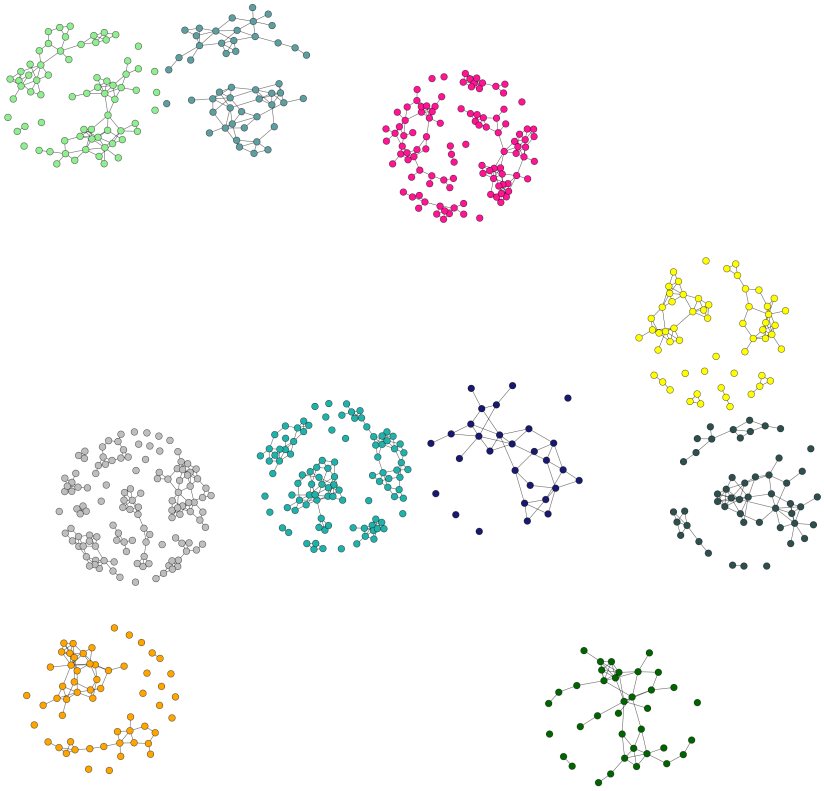


FIGURE 3.2: **Extract of School-clusters from Recife.** The Figure presents an extract of the individual pupils that participated in the FUNDAJ-Survey. Pupils are colored according to their school. Location of pupils is assigned by a Fruchtermann-Reingold algorithm [74] centered around the location of their school within the city of Recife. Grey lines indicate friendships between pupils as registered by the survey. As social networks where solely surveyed within schools, isolated components appear for each school.

Recife, as most large Brazilian cities, is characterized by a marked social divide between districts [75]. Hence districts and the people living in them are differently affected by governmental social welfare programs as for example the “Bolsa Familia" Program [2]. In order to simulate indirect effects that stem from those social welfare programs, for example the spreading of behavior through friendship or kinship networks, a model of the global structure of the social networks of the population is required. The following Sections present and compare different approaches to deal with this missing data problem that appears frequently due to boundary specification in social network surveys.

3.3.2 Approaches for Network Extrapolation

Three different approaches to impute friendship connections between the isolated components of the original semi-connected network are employed. Hereby it is important to note that approaches aim at creating a good model of the global network rather than actually finding the individual connections. The first approach stems

from the “Social Circles" model proposed by Hamill and Gilbert [62], the second approach employs a link-prediction technique based on a two phase bootstrapping procedure. Eventually, elements from both techniques are joined in a third, combined approach.

Social Circles/Waxman Model Approach: The approach to the problem described in the previous paragraph, was inspired by Hamill and Gilbert's social circle model [62]. Hereby the social distance is defined as as the physical distance between any two pupils. However, unlike Hamill and Gilbert, our challenge is not to create a social network from scratch, but to insert links between the isolated components of an already existing social network data-set. Hence the social circles approach must be adapted to the underlying problem.

As the location of the individuals is available only on district level, a stand-alone social circles approach would lead to very densely connected components because all individuals that went to school in the same district were within one social circle and hence connected to each other. Thus, instead of purely adopting the social radius, an exponential decay model based on a probability function related to the social distance is used. This approximates the approach to the well known “Waxman Model" [73]. In contrast do Hamill and Gilbert's work, heterogeneity of the span of social networks is not introduced by drawing social reach from a probability distribution, but by assigning a probability to each possible connection. Hereby the probability decreases with increasing distance and increasing accumulated degree of the two nodes to be connected. Hence, the attachment of new edges between nodes that have been unconnected or sparsely connected before is favoured. This may seem to oppose the assortativity assumption. However, isolated and sparsely connected nodes within the data-set do not necessarily indicate that those individuals are disconnected in reality but rather that their friendship nominations have been outside their school class and are thus missing information. By controlling the probability of attaching a new link with the accumulated degree of two nodes, the attachment of new links to nodes that feature not at random missing information (MNAR, defined in previous Section) about friendship ties becomes more likely. This operator differs from the “Waxman Model", where the density of the links is controlled by a global parameter. Although this approach is very close to the “Waxman Model", it is addressed with the term “social circles approach" in the remainder of this Chapter, because the use of the distance between pupils has been inspired by Hamill and Gilbert's social circle model.

The approach that is described above, has been implemented as follows. For every possible connection that does not yet exist, the following procedures are executed:

A connection probability is computed according to Equation 3.1. Hereby P_{link} denotes the connection probability; d stands for the distance in meters between the

two nodes; k denotes the accumulated number of neighbors of both nodes and c is an adjustable parameter.

$$P_{link}(u,v) = e^{\frac{-d(u,v)k}{c}} \tag{3.1}$$

As P_{link} is being affected by the accumulated degree of the respective vertices, the order of link-estimation matters. In order to favor links between close vertices, $P_{link}(u,v)$ is being calculated for randomly picked u and a semi-random v. Semi-random in this case means that v is chosen randomly among the vertices of the closest school to u that has not been chosen yet. After computing P_{link}, a random threshold $r \in [0,1]$ is generated for each possible link. If P_{link} surpasses the threshold, the new connection is effectively created.

Cold Start Link-Prediction - Bootstrapping Approach: The described social circle approach makes use of the "social distance" between two individuals and, implementing the circle concept, implicitly generates expected network properties. However, it does not make use of information that is implicitly stored in the network structure. We for example know that transitivity or triadic closure is a common phenomenon in social networks. This means that the probability that an edge exists between two nodes increases with the number of common neighbors of those nodes. In order to use this implicitly available information for creating the missing links between the isolated components of the network data, we implement as a second approach a two phase bootstrapping algorithm as proposed in [57]. The two phase approach allows for estimating probabilities for all potentially existing and not yet nominated friendships in the network and subsequently applying graph based measures to reinforce the probabilities for the existence of links.

Phase I makes use of individually available information about the nodes, according to the work of Leroy et. al. [57]. Yet, Leroy et. al. deal with data from Flickr. Thus, they use the common membership in thematic groups of two individuals to estimate a probability for the existence of a link between them in the first phase.

In the underlying case, such obvious similarities are not available. However, information is available about the district of the domicile and of the school of the pupils. The data-set also contains information about leisure activities such as membership in sports associations or religious organizations of the children, about their integration into their neighborhood and how they arrive at school. It can be claimed that friendships between children that emerge outside the school environment are frequently established either in the neighborhood of the domicile, during leisure activities or on the way to school. Therefore, this information is used to estimate probabilities for the existence of links that have not been recorded within the network survey conducted at class level. Similar to Leroy et. al, groups are defined as co-occurrence of geographical and behavioral properties of the nodes, such as doing the same activity within the same district, or sharing a means of public transport on the way to school.

In order to assign probabilities $P_{link_1}(u, v|u, v \in \gamma)$ for the existence of a friendship between two individuals that share a group γ, we compute the share of links between the nodes that belong to the respective common group γ within the original data according to Equation 3.2. Hereby e_γ denotes the number of links originating from vertices within the group and l_γ denotes the number of links between vertices within the group. Please note that e_γ also contains links to vertices outside the group.

$$P_{link_{1_\gamma}}(u, v) = \begin{cases} \frac{l_\gamma}{e_\gamma}, & if u, v \in \gamma. \\ 0, & otherwise. \end{cases} \tag{3.2}$$

Considering that the membership in each common group may be the cause for the friendship between the two individuals in question, the outcome “friendship" occurs if a link is established within at least one of the common groups of the two individuals. Hence, to estimate the existence of a friendship, it is computed the probability that at least one outcome occurs within a set of outcomes of size $2^k - 1$, where k is the number of distinct groups. Each outcome has the from $(x_1, ...x_k)$, where $x_k \in [0, 1]$, here 1 indicates that a link was established within the respective group.

The total probability for the existence of a link between a pair of nodes (u, v) can then be computed as the sum of the probabilities provided by each outcome within the set of outcomes according to Equation 3.3. All groups used may be reviewed in Table 3.1. The respective probabilities $P_{link_1}(u, v|u, v \in \gamma)$ vary significantly with group size. Thus, $P_{link_1}(u, v|u, v \in \gamma)$ increases significantly with decreasing group size. A link between two individuals from a small district with few other pupils that share a common activity is hence more likely than a link between two individuals from a large district.

It needs to be recognized that drawing $P_{link_1}(u, v|u, v \in \gamma)$ from the observable data may introduce bias, as friendships can only be observed between pupils that visit the same class. It seems to be reasonable, however that common groups and common interests play a major role in the establishment of friendships also on class level and hence the available data may be considered a good indicator for the weight of common groups on link creation.

$$P_{link_1}(u, v) = \sum_{i=1}^{2^k-1} \prod_{q=1}^{k} x_{iq} \tag{3.3}$$

TABLE 3.1: **Definition of Groups**

Name	Condition
Sports	Pupils live in the same district and regularly practice sports
Church	Pupils live in the same district and frequent church or religious services
Transport	Pupils live in the same district, go to school in the same district and use the same public transportation on the way to school

The table presents combinations of activities and locations that define a group. Pupils that share a group are assigned a certain probability that a link exists between them.

Phase II foresees the application of graph-based measures. "*Common_Neighbors*" [47] measure has been shown to perform well as graph-based measure for reinforcing probabilities of links that will potentially exist in the second phase of the bootstrapping algorithm [57]. Therefore the implementation contains the adaption of "*Common_Neighbors*" for the cold-start link-prediction problem according to [57]. Consequently probability scores $score(u,v)$ are derived, adding the probability $P_{link_2}(u,v)$ derived from "*Common_Neighbors*" measure to the probability $P_{link_1}(u,v)$ calculated in **Phase I** of the bootstrapping method. $P_{link_2}(u,v)$ is hereby computed according to Equation 3.4 for the pair of nodes (u,v) as the sum of the probabilities of each node y within the graph U being linked to both, u and v.

$$P_{link_2}(u,v) = \sum_{y \in U} P_{link_1}(u,y) \times P_{link_1}(v,y) \tag{3.4}$$

Subsequently, the scores calculated in **Phase I** and **Phase II** are converted to probability values inline with the work of Leroy et.al. [57], using a simple logarithmic function as presented in Equation 3.6.

$$score(u,v) = P_{link_1}(u,v) + P_{link_2}(u,v) \tag{3.5}$$

$$P_{link}(u,v) = \frac{log(score(u,v)+1)}{log(max(score(u,v)|u,v \in U)+1)} \tag{3.6}$$

In order to define the links that finally exist, a threshold $r \in [0,1]$ is applied. Links that have been estimated to exist with a probability $P_{link}(u,v) \geq r$ are considered as existing links.

Combined approach: The bootstrapping approach employs more information about the nodes than the social circle approach does. Hence, we expect that individual link-prediction is more accurate with the bootstrapping approach. However, the effectiveness of this approach is restricted by availability and granularity of information. In the presented implementation, for example, groups are defined based on

the district of residence of a pupil and her personal activities. This is a very strong restriction as it implies that friendships between pupils from adjacent districts are not considered. Hence, the granularity of available information limits the effectiveness of the bootstrapping approach. To overcome these restrictions, the combined approach aims at joining the generality of the social circles approach with the specificity of the bootstrapping method. For this purpose, the social circles approach is incorporated in the first phase of the bootstrapping algorithm.

In **Phase I** the calculation of $P_{link_1}(u,v)$ is slightly modified according to Equation 3.7. Other than in the social circles implementation, newly estimated links are not treated as existing links immediately, but receive a probability as in the Bootstrapping approach. Hence k may not be calculated as the degree of existing links of a node, but as the sum of probabilities of potentially existing links.

$$P_{link}(u,v) = e^{\frac{-d(u,v)}{c}} \frac{1}{k+1} \tag{3.7}$$

Hereby personal information (group membership) is included into the definition of social distance according to Equation 3.8 where x and y describe the geographical vectors and s describes the social position of the individual. The social position is defined randomly on an interval $[0, \frac{\alpha}{2^n}]$, where n is the number of common social activities of an individual and α is an adjustable parameter. For experiments the social activities described in Table 3.1 were applied. Moreover, the social distance between two individuals may always decline with the number of common social activities.

$$d(u,v) = \sqrt{(x_u - x_v)^2 + (y_u - y_v)^2 + (s_u - s_v)^2} \tag{3.8}$$

Phase II implements the "*Common_Neighbors*" measure as explained in the bootstrapping approach. Subsequently a random threshold $t \in [0,1]$ is generated for each possible link. If $P_{link}(u,v)$ surpasses the threshold, the new connection is effectively created.

3.4 Experimental Results - Missing Data

This Section gives examples of how the quality of the generated global networks may be measured and presents the experimental results for the network extrapolation techniques *Social Circle Approach*, *Cold Start Link-Prediction* and *Combined Approach*. In a first Subsection, individual network properties are assessed while the second Subsection evaluates overall network features and compares them to known real world networks.

3.4.1 Quality Assessment of Individual Network Features

In order to assess the quality of the implemented extrapolation techniques, a set of reference values is introduced in a first step. For assessing how well the generated global network represents properties of real world social networks, the generated values for a number of network measures are being compared. The first measure is the *average degree* of the network, calculated as the average number of friends the individuals within the network have. Further the network *density*, given by the ratio of links to the number of possible links within the network may be consulted. *Average Shortest Path* indicates the average number of steps needed to reach any node v when starting from any node u in the network, if a path exists between them. The *Clustering Coefficient* can be calculated as the ratio of the triangles and connected triplets in the graph. Moreover, *Assortativity* indicates the correlation between the degree of connected vertices as proposed in [70]. Eventually, the average number of links a student has outside her school environment *Out-links* may be treated as a reference value. Reference values may be drawn from distinct sources in the literature. The respective benchmarks and sources can respectively be reviewed in Table 3.2.

Further, the performance of the distinct parameter settings may be assessed, analyzing the degree distributions of the generated networks in comparison with real friendship networks, as well as the relation between probability of friendship and physical distance between any two individuals.

TABLE 3.2: **Reference Values for Quality Assessment of Network Extrapolation**

reference value	**objective range or limit**	Description
Average Degree	$\in [5, 10]$ (objective range)	We consider networks of close friendships, hence the average degree may be limited [61].
Density	≤ 0.014	This density value has been calculated for a larger but similar study with adolescents in the United States [76]. However, as we interconnect the isolated schools we expect a lower density value.
Average Shortest Path	$\in [5, 7]$ (objective range)	We expect "small-world" features [62].
Clustering Coefficient	≤ 0.252	This clustering value has been calculated for a larger but similar study with adolescents in the United States [76]. As we interconnect the isolated schools, where we expect to have less connections between schools than within schools, we expect a lower Clustering Coefficient.
Assortativity	*positive*	The positive assortativity indicates the existence of preferential attachment [62].
# *Out-links*	$\in [1, 2]$ (objective range)	Average number of links outside the school in AddHealth study [76].
# *Isolated vertices*	*min*	We aim to connect isolated vertices and hence desire a minimum number of disconnected vertices.
# *Components*	*min*	We aim to connect components and hence desire a minimum number of disconnected components.

The table contains the measures chosen to evaluate the proximity of the generated networks to real world social networks.

Social Circles Approach: Experiments with the social circles approach were run using different values for the parameter c. As the parameter c reduces the exponent of the exponential decay function in Equation 3.1 and hereby increases the probability for a new link to be formed, it may be expected to generate more highly connected networks with increasing values for c. Due to the spacial emphasis of this approach, very distant schools led to problems in the implementation, that is why these experiments have been run on a reduced data-set, where outlier schools have been removed.

Experimental results for the reference values pointed out in Table 3.2 are illustrated in Figure 3.3. One may observe that *density* indicated by the green solid line in the upper Figure takes values between 0.001 and 0.003 for all networks generated with variations of parameter c. It hereby lies steadily under the maximum value indicated in Table 3.2.

The blue solid line in the second sub-figure representing the networks assortativity-coefficient fluctuates for different c values but remains positive for all parameter settings.

The *Clustering-Coefficient* as represented by the yellow solid line in the second sub-figure decreases with increasing c and crosses the upper limit defined in Table 3.2 at a c value of approximately 500.

The number of *Out-links* increases with growing c as presented by the solid green line in the third sub-figure, ranging from 0 for low c values to 8 for very high c values.

Average Degree is being illustrated by the solid red line in the third sub-figure. It increases also with increasing c and reaches the objective range as indicated in Table 3.2 between the two dashed red lines for c values between 500 and 1800.

Average Shortest Path, represented by the blue solid line in the fourth sub-figure, raises for low c but subsequently decreases with increasing c. This seems logical, since shortest path is calculated as the average of the shortest paths of all components in the graph. As the components become better interconnected with growing c, shortest path initially increases. The predefined objective interval for *Average Shortest Path* as indicated by the blue dashed lines can be reached for c values below 400.

The last sub-figure illustrates the results for the number of components as a percentage of original number of components with a red solid line and the total number of isolated vertices as a percentage of original number of components with a green solid line. It can be observed that the percentage of isolated nodes can be kept close to zero for all c values, while the percentage of components decreases with increasing c and reaches values close to zero when applying c values of 500 or higher. The Figure presents percentages above 100% for very few values of c. This is possible as components are defined as a set of at least two connected nodes that are only linked to each other. Hence, for those very low c values many former isolated nodes connect to each other and form new components.

Figure 3.4 presents the generated network with a c value of 500, where colored points represent pupils (each school is indicated by a different color) and edges represent friendships between the pupils. This Figure illustrates that a globally interconnected network has been generated where social network typical patterns can be observed.

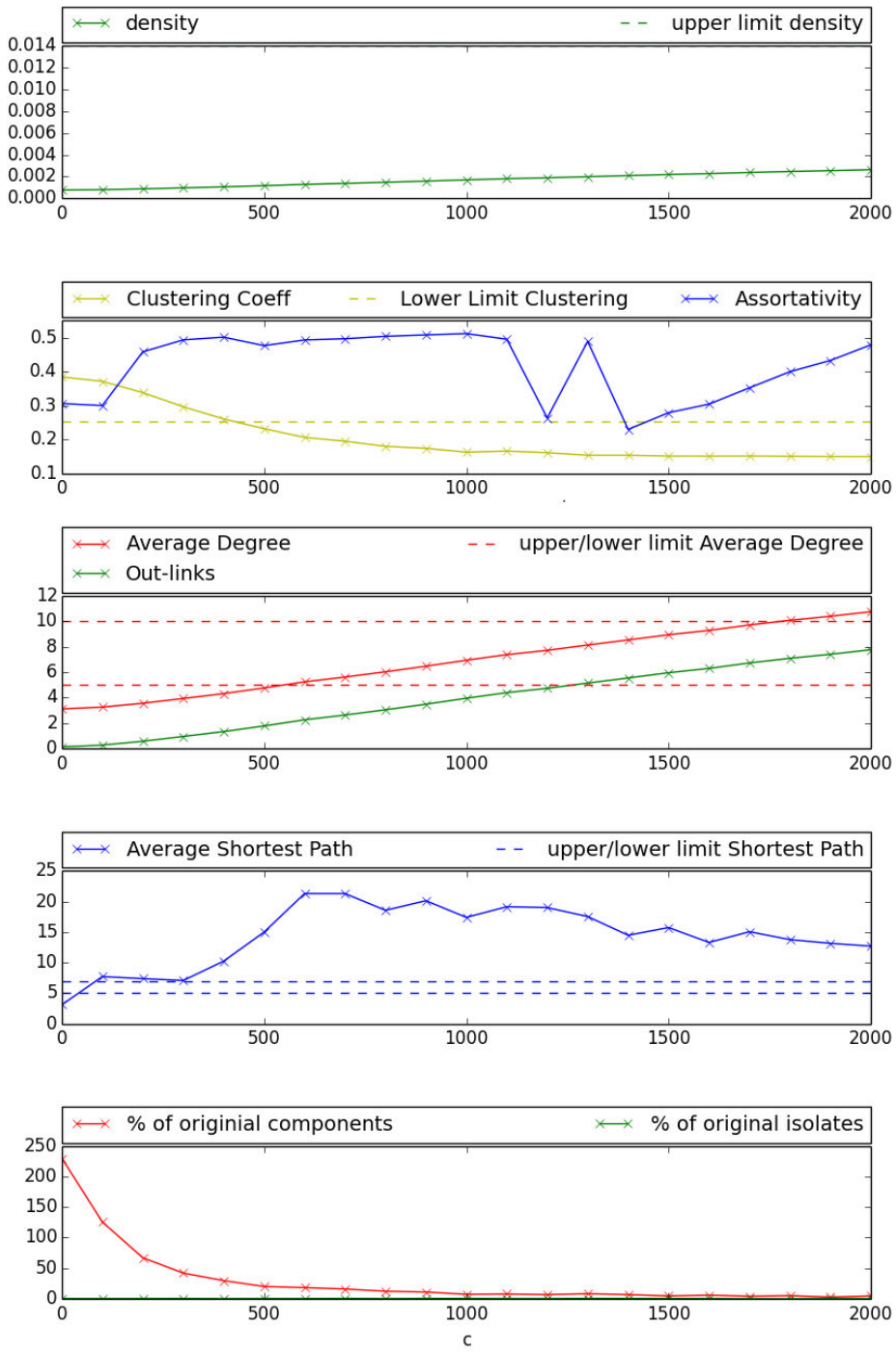


FIGURE 3.3: **Social Circles - Objective Values.** The abscissa scales the different values of the parameter c that controls the exponent of the exponential decay function in Equation 3.1; objective ranges and upper-/lower limits are indicated by dashed lines. Most of the desired objective values can be reached with differing values of c. The results indicate that a desirably small number of components may be reached using c values higher than or equal to 500.

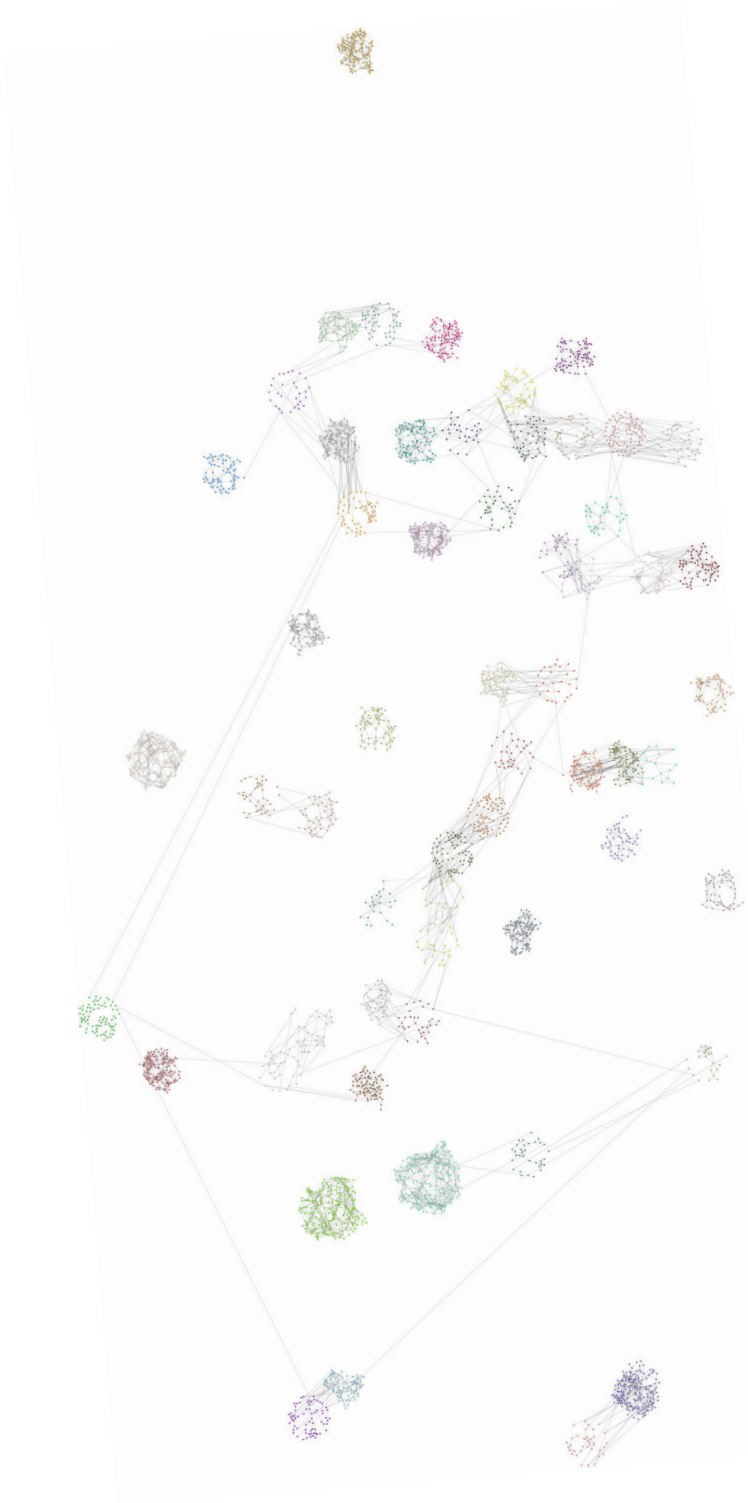


FIGURE 3.4: **Network Created by Social Circles Approach.** The Figure presents the individual pupils that participated in the FUNDAJ-Survey. Pupils are colored according to their school. Location of pupils is assigned by a Fruchtermann-Reingold algorithm within a radius of n^2 around the location of their school within the city of Recife. n indicates the number of pupils of the respective school. Grey lines indicate friendships between pupils as registered by the survey, as well as friendships estimated by the social circles method applying $c = 500$. It can be observed that a fairly well connected network has been generated.

Cold Start Link-Prediction - Bootstrapping Approach: The bootstrapping approach has been implemented according to the description in the previous Section. The respective experiments were carried out with several values for the threshold $r \in [0, 1]$. The threshold r controls the probability that a link exists, thus more densely connected networks are created with decreasing r.

Figure 3.5 reveals in the first and second sub-figure that *Assortativity* indicated by the blue solid line remains positive for all settings, *Clustering Coefficient* as indicated by the yellow solid line remains far above the maximum value and *density* represented by the green solid line remains below the maximum value for most settings. However, the third sub-figure shows that *Average Degree* and *Out-links* reach undesirably high values for low settings of r and can only reach the objective area for experiments with $r \geq 0.89$.

As indicated by sub-figure four, *Average Shortest Path* reaches desired values for $r \geq 0.89$. Analysis of sub-figure five yields that low and therefore desirable percentage values for the number of isolated components and the number of isolated nodes can be reached for low r. However, those indicators still yield a reduction of more than 30% for settings with high r.

Figure 3.6 illustrates the network created with the bootstrapping technique using a threshold of 0.91. One may observe within this Figure that the created network still contains a considerable amount of isolated components and that many schools remain unconnected even if they are very close to other schools. Furthermore, connections seem to be established between very few individuals of the different schools.

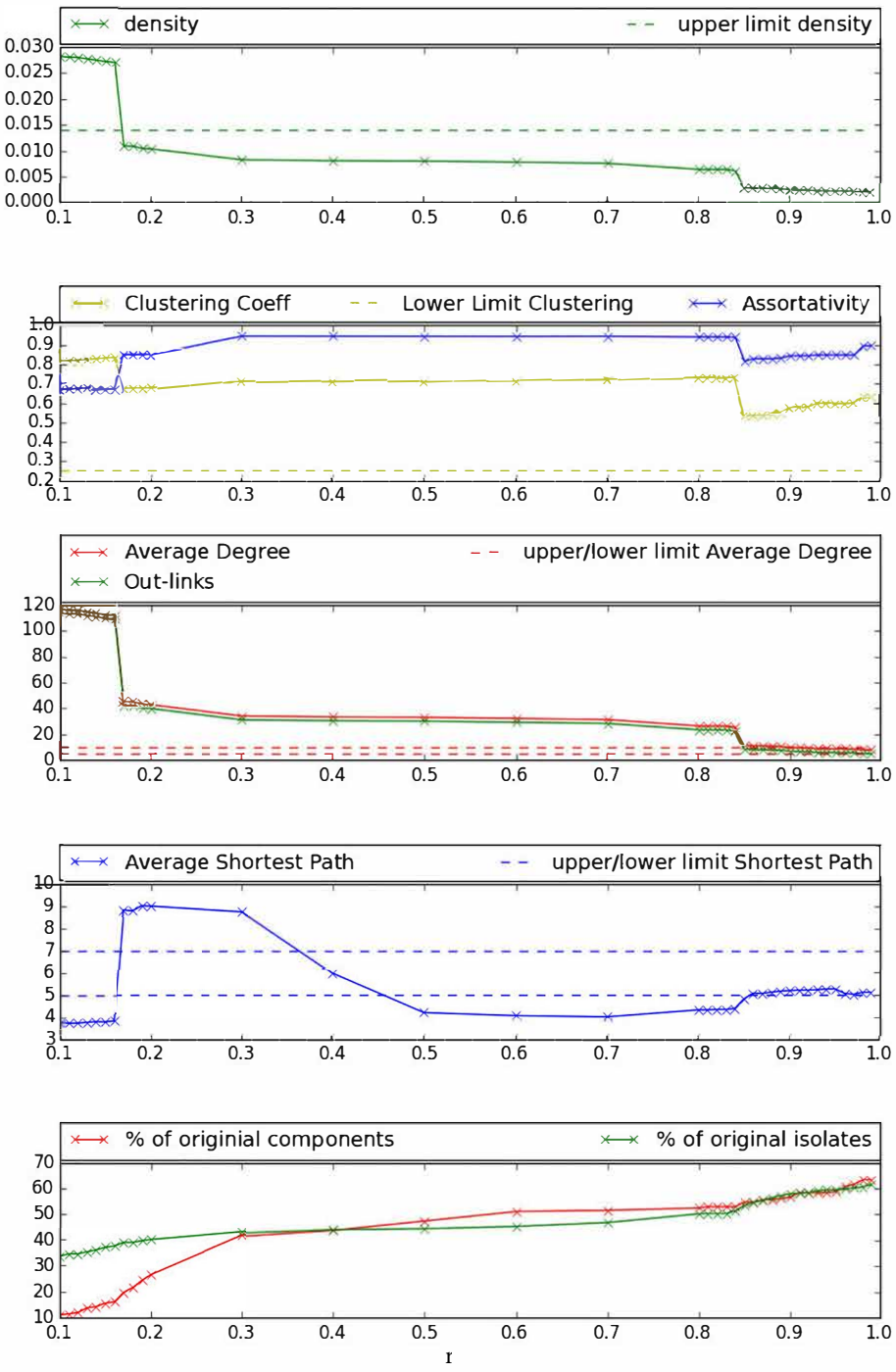


FIGURE 3.5: **Bootstrapping - Objective Values.** The abscissa scales the different values of the threshold parameter r; objective ranges and upper-/lower limits are indicated by dashed lines. Most objective values can be met if running the cold start link-prediction approach using a high final threshold ($r \geq 0.89$).

FIGURE 3.6: **Network Created by Bootstrapping.** The Figure presents the individual pupils that participated in the FUNDAJ-Survey. Pupils are colored according to their school. Location of pupils is assigned by a Fruchtermann-Reingold algorithm within a radius of n^2 around the location of their school within the city of Recife. n indicates the number of pupils of the respective school. Grey lines indicate friendships between pupils as registered by the survey, as well as friendships estimated by the Bootstrapping method applying $r = 0.91$. Note that a considerable number components remain disconnected.

Combined Approach: As explained in the previous Section, the combined approach incorporates elements from the Social Circles and Bootstrapping approaches. Experiments were run using varying values for the parameter c that controls the probability for a new connection to exist in the first phase of the algorithm. Since a random threshold is assigned to each connection finally, results can be illustrated by scaling the values for parameter c on the abscissa.

As Figure 3.7 shows in the first sub-figure, density can be kept on a desirable level for all c values. The second sub-figure reveals that also *Assortativity* remains positive for all c values, but *Clustering Coefficient* declines for growing c and reaches values under the objective upper limit for $c \geq 200$.

We observe in the third sub-figure that values for *Average Degree* reach the objective range for $c \in [300, 900]$, while this can be stated for *Out-links* for $c \in [0, 300]$.

Average Shortest Path as shown in the fourth sub-figure remains within the objective range for nearly all values of c, except $c \in [300, 900]$, where the objective range is slightly exceeded. As illustrated by the fifth sub-figure, the number of isolated components and isolated nodes decreases steeply with growing c, reaching an almost totally interconnected network for $c \geq 600$.

Figure 3.8 presents a network generated with the combined approach under a c value of 300. Obviously, inter-school links are better distributed between pupils and the graph seems better inter connected than the comparable graph generated with the bootstrapping approach. However, the network remains not fully connected for this setting and especially more distant schools remain isolated.

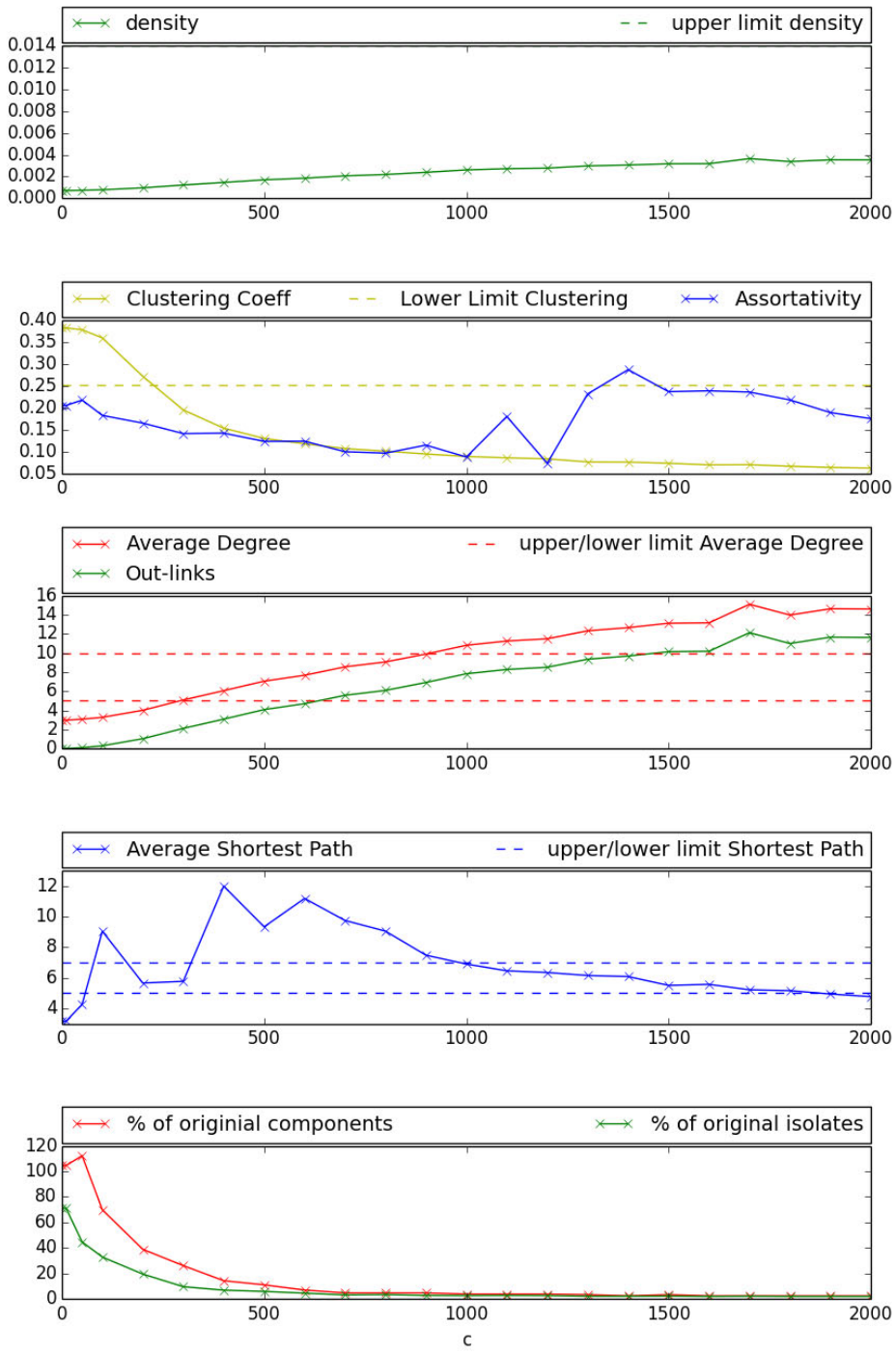


FIGURE 3.7: **Combined Approach - Objective Values.** The abscissa scales the different values of the parameter c that controls the exponent of the exponential decay function in Equation 3.7; objective ranges and upper-/lower limits are indicated by dashed lines. Global measures can be kept within or close to the objective intervals for parameter settings with $500 \leq c \leq 1000$.

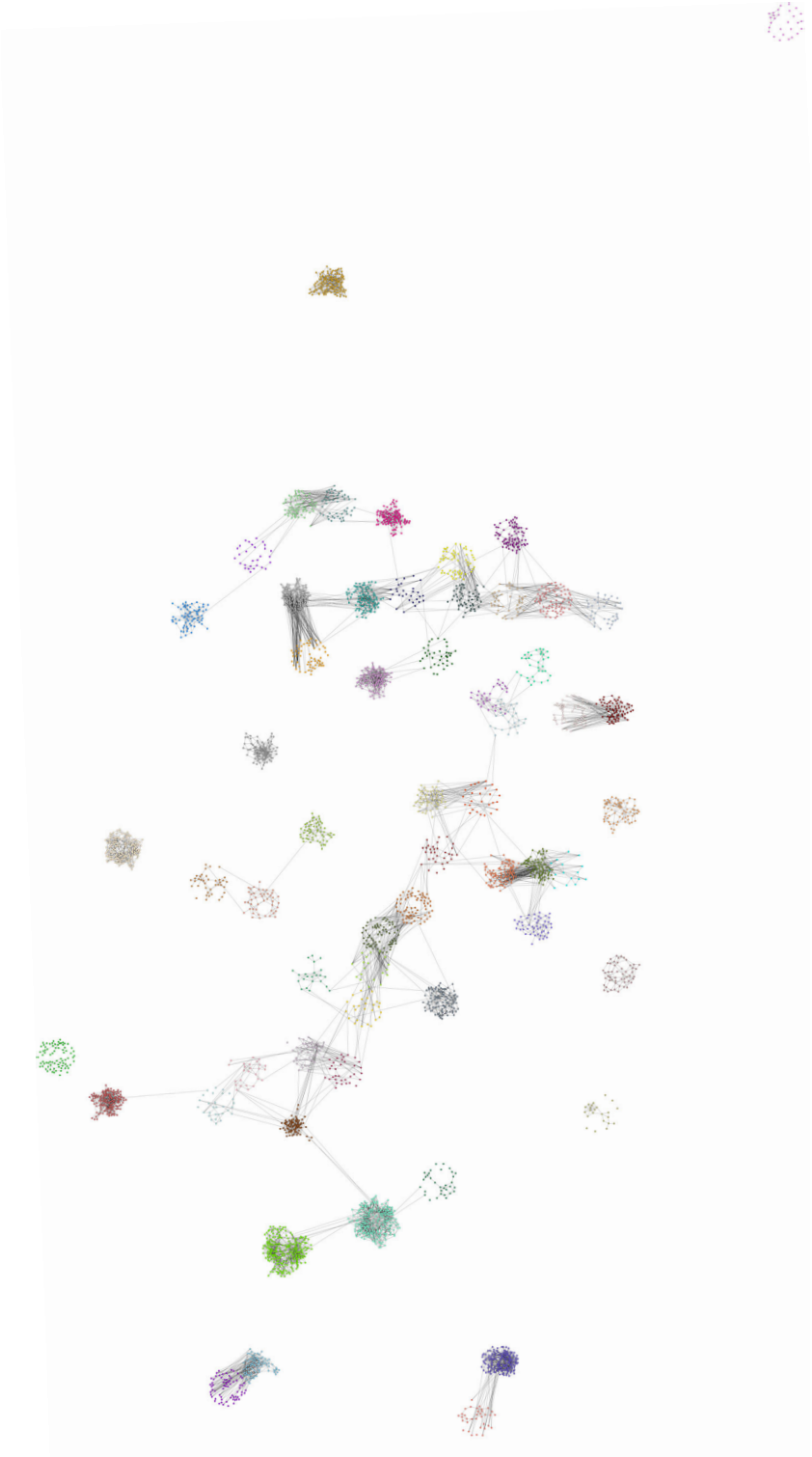


FIGURE 3.8: **Network created by Combined Approach.** The Figure presents the individual pupils that participated in the FUNDAJ-Survey. Pupils are colored according to their school. Location of pupils is assigned by a Fruchtermann-Reingold algorithm within a radius of n^2 around the location of their school within the city of Recife. n indicates the number of pupils of the respective school. Grey lines indicate friendships between pupils as registered by the survey, as well as friendships estimated by the combined method applying $c = 300$. The combined approach yields a better reduction of isolated components than the Cold-Start Link-prediction approach.

3.4.2 Quality Assessment of Overall Network Features

The measures introduced in the precedent Subsection indicate that all three approaches where capable of generating networks that exhibit real world social network properties.

In order to better understand the outcomes for the different approaches, this subsection analyzes overall network attributes and compares them to data from real-world networks.

Since the degree distribution has been found to be a feature that captures an important part of the network structure, we analyze the distributions of the networks generated with the extrapolation approaches in a first paragraph. Subsequently, an other robust characteristic of social networks, the relation between link probability and physical distance between vertices is being examined.

Analysis of Degree Distribution: This paragraph sets the degree distribution of the generated networks in relation to degree distributions of real world social networks. For comparison, original degree distributions from comparable networks are plotted. Those are: the original degree distribution of the FUNDAJ graph, the degree distribution of the AddHealth Network and the in- and out- degree distributions of the testimonial network of the Korean online social network “Cyworld".

The National Longitudinal Study of Adolescent to Adult Health (AddHealth) network was surveyed among a representative sample of adolescents in grades 7 to 12, it contains 90118 students from 145 schools in 80 communities in the United States of America. Students could not only nominate peers from their own school, but also peers from a “sister-school". AddHealth combines longitudinal survey data on respondents’ social, economic, psychological and physical well-being with contextual data on the family, neighborhood, community, school, friendships, peer groups, and romantic relationships [76].

Cyworld is an online social network similar to Facebook, where users may create a personal profile and create friendship ties to other users. Users are hereby enabled to write testimonials for other users. As the writing of a testimonial requires some effort and knowledge about the receiver of the testimonial, it has been shown that the network of testimonials in Cyworld resembles real friendship networks very closely [77]. In the remainder of this Chapter, we examine exclusively the Cyworld testimonial network, however in order to ensure readability, it is referred to as the “Cyworld network".

Figure 3.9 illustrates the degree distributions of the networks generated with the presented techniques along with log-log plots of degree distributions of Cyworld-, Add-Health- and the original FUNDAJ Network. Here the solid lines represent networks that were extrapolated based on the FUNDAJ graph. The dashed lines indicate the real networks.

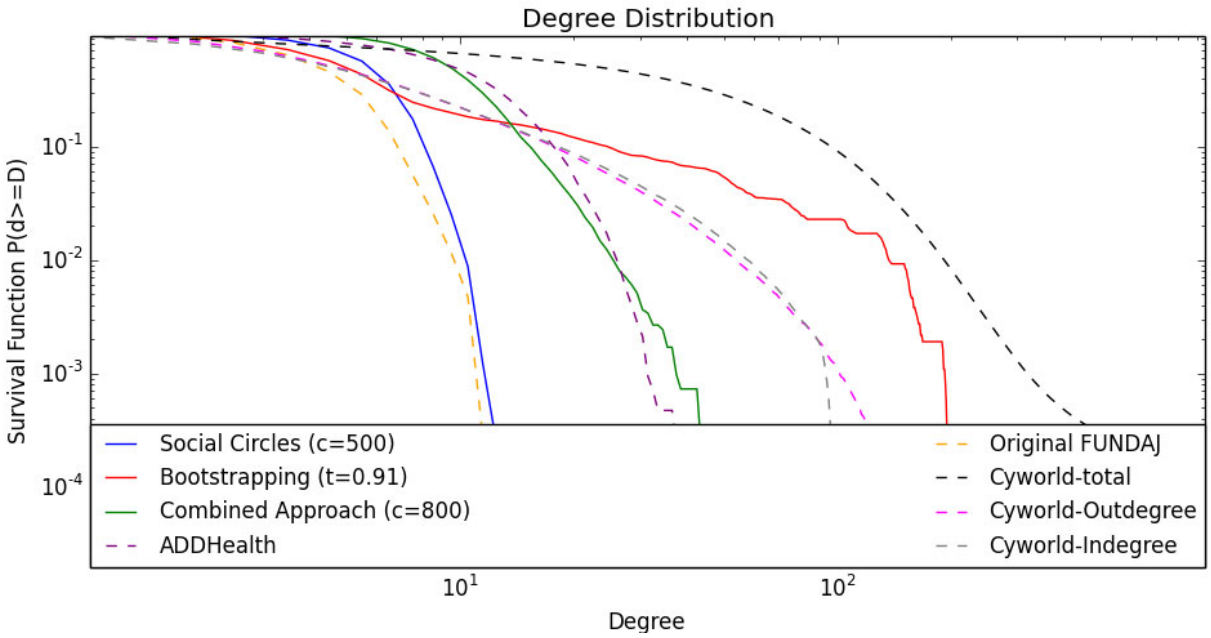


FIGURE 3.9: **Degree distributions for networks generated with Social Circles (c=500), Bootstrapping(r=0.91) and Combined approach (c=800).** Compared to degree distributions from real world social networks: Original FUNDAJ Graph, AddHealth Network and Cyworld in-and out degree distribution. Social Circles Approach maintains the original Degree Distribution, Combined Approach approximates the AddHealth Degree Distribution and Bootstrapping approach generates a network similar to Cyworld online social network

The Figure reveals that the different approaches generate degree distributions similar to the distributions observed in real friendship networks. It is striking that the degree distribution generated by the social circles approach, as indicated by the blue solid line, appears to very closely approximate the original FUNDAJ graph, represented by the dashed yellow line, featuring a similar shape and comparable maximum degree. On the other hand, the AddHealth-degree distribution (dashed purple line) seems to be well represented by the combined approach (solid green line) as their plots are almost congruent. Finally, the complementary cumulative distribution (CCDF) of Cyworld in- and out- degrees can be well approximated by the bootstrapping approach, as shown by the obvious similarity of the graph indicated by the red solid line (Bootstrapping approach) and the dashed purple and grey lines that represent Cyworld out- and respectively in- degree distributions.

Analysis of Friendship Probability in Relation to Physical Distance: Another characteristic property of real world friendship networks is the relation between

link-probability and distance between nodes [65]. It seems intuitive and is also universally agreed that the probability for the existence of a link decreases with increasing distance between the respective nodes. This property can also be found in recent networks, despite the progress made in information and transportation technology, while the exact relation varies depending on the studied data-set [65].

Therefore, if the generated networks are to represent real-world features, it may be expected to find a similar relationship for distance and node probability within the extrapolated data. Hence, in order to evaluate the suitability of the introduced approaches to produce real-world network properties, Figure 3.10 contrasts CCDF log-log plots for the probability of the existence of a link between nodes within a certain distance for the networks generated with the presented techniques, as well as from real world social networks. Here the solid lines represent either distributions from networks that were generated using the presented techniques, or distributions from the location-based online social networks Brightkite and Gowalla (purple and orange solid lines).

Both, Brightkite and Gowalla are location based online networking services, that enable the user not only to establish links to other users, but also to provide information about her current location. These characteristics make of Brightkite and Gowalla good references for the analysis of friendship probability in relation to physical distance in extrapolated social network data. Data from Gowalla and Brightkite where obtained from the Stanford Large Network Data-set Collection [78]. Since both data sources provide time series information about the places the users stayed at, the most frequent location for each user is considered as her respective domicile.

As the experiments presented in this Chapter all extrapolate the network data obtained by FUNDAJ in the city of Recife, comparison data was required from a locally restricted urban environment. Hence, also subsets of the Gowalla and Brightkite networks are plotted, containing solely individuals located within the city of New York (locations within the latitude interval [40.65,40.80)] and the longitude interval [-74.05,-73.90]). The solid yellow line represents the distribution for the Brightkite network from New York, while the black solid line indicates the Gowalla New York network (yellow and black lines are very close). In order to compare the different distributions, a total of 90 continuous probability distributions from the Numpy Library [79] were fitted to the data via least squares fitting. The distribution with the highest p-value and lowest Kolmogorov-Smirnov test statistic was chosen as the best fit.

Figure 3.10 reveals that all local networks could be reasonably well represented by folded Cauchy distributions as presented in Equation 3.9 with slightly different values for parameter a, while the distributions of the global Brightkite and Gowalla networks seem to be better approximated by a power law distribution with exponent between 0.15 and 0.7.

$$f(x,c) = \frac{1}{\pi(1+(x-a)^2)} + \frac{1}{\pi(1+(x+a)^2)} \tag{3.9}$$

Please note that the survival function of the power law plot does not appear as a straight line in this graph, as it is usually the case for power law relations on log-log plots. This is due to the finite nature of distances on earth. As distances between individuals on this planet are restricted, the power law plot bends down as the x_{max} is approached.

Goodness of fit is also checked, calculating the Pearson Correlation Coefficient for each pair of distribution and fit. The correlation coefficient supports the goodness of fit for all distributions and respective fitted distributions.

Visual analysis reveals that the network generated using the combined approach approximated the Gowalla and Brightkite networks of the city of New York best, while Social Circles approach seems to better maintain the original distance-link-probability as indicated by the pink solid line (original graph) and the purple dashed line (fitted folded Cauchy distribution). The red solid curve that indicates the relation of distance to link-probability for the Bootstrapping Approach appears to be differently shaped than the other distributions, as it shows a buckle for distances greater than $10^{2.5}$. The correlation coefficient still suggests the folded Cauchy distribution with parameter $a = 0.659$ as a good fit for the Bootstrapping curve. This disturbance stems from the nature of log-log plotting, where the deviations scale down for large values of the variables. Further observation of parameters c of the fitted folded Cauchy distributions supports the hypothesis that the combined approach most closely approximates the local Gowalla and Brightkite data, as the fitted folded Cauchy distribution for the combined approach has parameter $a = 0.504$ and hence lies between Gowalla New York ($a = 0.679$) and Brightkite New York ($a = 0.234$). The fitted folded Cauchy distribution for the Bootstrapping Approach network exhibits a parameter ($a = 0.659$) even closer to Gowalla New York. However, due to the different shape of the curves this similarity should be taken carefully. Moreover, the parameters of the fitted folded Cauchy distributions for the original network ($a = 1.854$) and the Social Circles Approach ($a = 1.067$) seem to be quite similar.

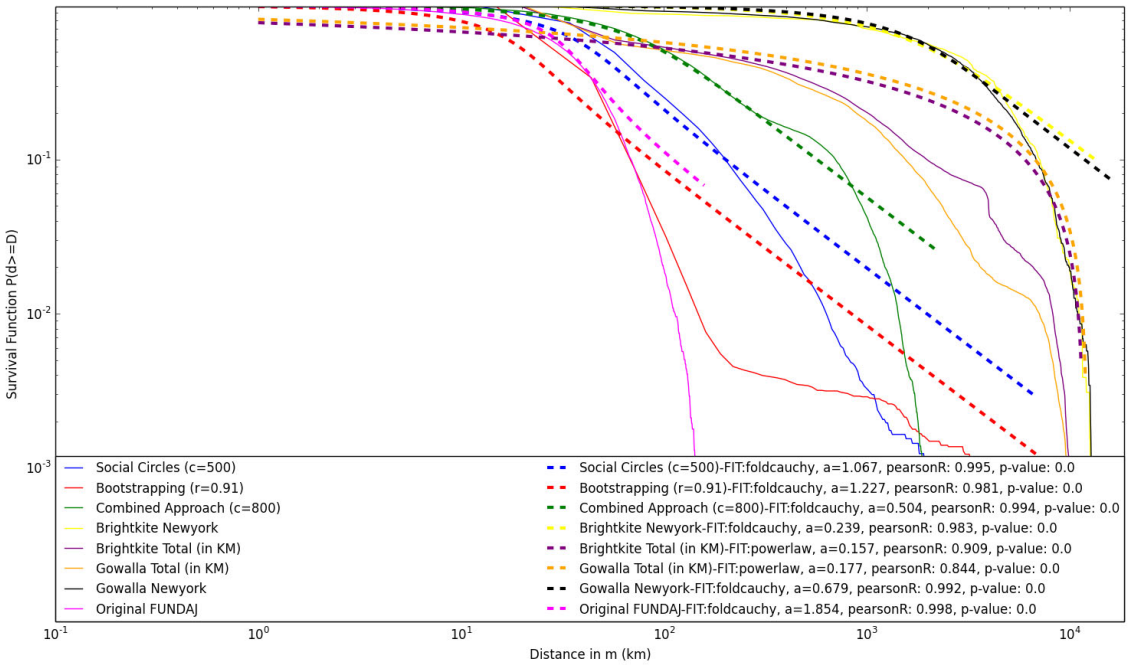


FIGURE 3.10: **Log-Log Plot of Survival Function (CCDF).** Link-probability related to physical distance between nodes for networks generated with Social Circles (c=500), Bootstrapping(r=0.91) and Combined approach (c=800). Compared to distance- link-probability distributions from real world social networks: Brightkite and Gowalla worldwide networks, as well as local sub networks for the city of New York. Visual analysis reveals that the network generated using the combined approach approximated the Gowalla and Brightkite networks of the city of New York best, while Social Circles approach seems to better maintain the original distance-link-probability

3.5 Discussion of the Three Approaches

This Section evaluates the results obtained from the experiments and sets them in the context of the initial problem, to extrapolate given data of social networks containing isolated components. The performance of the three presented approaches is discussed and promising applications for each of them are pointed out.

3.5.1 Social Circles Approach

The results presented in the previous Section indicate that most of the desired objective values presented in Table 3.2 can be reached with differing values of c. As the aim of this work is the interconnection of isolated sub-components of a given network, the number of components is a crucial value for demonstrating the suitability of the approach.

The results indicate that a desirably small number of components may be reached using c values higher than or equal to 500. However, the trade-off between two benchmarks with opposing trends has to be analyzed more deeply:

Clustering-Coefficient and *Average Shortest Path*. As shown in Figure 3.3, *Clustering Coefficient* reaches desirable values for $c \geq 500$, yet *Average Shortest Path* reaches the objective zone for c values ≤ 300 and also *Out-links* exceeds the maximum value for $c \geq 500$. Hence, one has to decide if it is more appropriate to create a network with "small world" properties but relatively high clustering and containing still a considerable number of isolated components or if a lower clustering, an interconnected network and relatively large values for *Average Shortest Path*, are desirable. In this case the researcher also assumes that the average number of links, an individual has with people from other entities is much higher than surveyed in the Add Health study [61].

The analysis of the degree distributions in Figure 3.9 shows that the Social Circles approach maintains the original degree distribution. This indicates it as a good model to generate an interconnected network that takes only very close friends in consideration, while it incorporates the hypothesis that micro-structures repeat themselves also on the macro level.

Figure 3.10 illustrates that the Social Circles approach generates the network exhibiting the highest similarity in distance-link-probability relation with the original graph. Even though maximum distance increases for the generated graph, the overall shape is kept and the overall structure of link-distances is maintained.

Therefore, it may be argued that the Social Circles Approach extrapolates the network data, maintaining the original characteristics of the isolated components best.

3.5.2 Bootstrapping Approach

The results show that most objective values (social network properties defined in Table 3.2) can be met if running the cold start link-prediction approach using a high final threshold $(r \geq 0.89)$. This setting also led to an improvement in the interconnection of the network by reducing the number of isolated nodes and components by approximately 30%.

However, it is not possible to create globally connected networks while reaching coherent values for average degree and the number of out-links of the pupils, because for those settings a considerable number of disconnected components remain. Moreover, clustering remains significantly above the lower limit defined in Table 3.2.

This may stem from the restriction of the prediction of friendships to pupils that live within the same district. This is a very strict restriction that does not necessarily represent reality, as pupils that live close to the district border may also have contacts from adjacent districts.

The plot of the degree distribution of the network created by the Bootstrapping technique, applying a final threshold of $(r \geq 0.89)$, seems to approximate the in- and out- degree distribution of the Cyworld testimonial network. As mentioned in previous work [77], the Cyworld testimonial network has a few, very active testimonial

writers with far above 100 written testimonials. The results produced here indicate that the restriction of activities to the four groups presented in Table 3.1 leads to a relatively small number of possibly connected individuals. Those individuals are assigned very high probabilities for link-creation in turn, and hence act similar to the “hubs" that appear in the Cyworld data.

From that point of view the bootstrapping approach seems to be well suited to generate networks with a usual degree distribution for close friendship networks, incorporating the special feature of the described “hubs".

However, in terms of distance - link-probability relation, the generated network is not comparable to other location-based social networks, as the ill-shaped curve for the relation of distance to link-probability of the Bootstrapping graph in Figure 3.10 indicates. Nevertheless, due to lacking access to the location of Cyworld users, it remains unclear if this ill-shaped curve could be a feature of online testimonial networks.

On the other hand, the applied groups are very broad. this may lead to a quite homogeneous and discrete distribution of probabilities in the first phase. Hence, in order to improve the performance, one could also try to narrow the definition of groups and thereby decrease overall probabilities for the existence of common groups. This may also lead to a better performance of low thresholds.

3.5.3 Combined Approach

The combined approach yields a better reduction of isolated components than the Cold-Start Link-prediction approach, while all other global measures can be kept within or close to the objective intervals. Those positive results can be obtained for parameter settings with $500 \leq c \leq 1000$. The combined approach further approximates the degree distribution of the AddHealth school network very closely for $c = 800$ as shown in Figure 3.9.

Although even the AddHealth study did not survey a completely interconnected network, but solely intra-school networks, it contains an interesting feature: the study design foresees a sister school for each participating school. The sister school is usually that junior-high school where the major share of students from a particular high school have studied before. The study design allowed the surveyed pupils to nominate also friends from this sister school, which leads to a network that is better interconnected than the network surveyed by FUNDAJ.

Moreover, the AddHealth questionnaire allowed the participants to nominate up to five best female friends and up to five best male friends, leading to a maximum of 10 nominated friends per pupil. Compared to the FUNDAJ survey, where only individual schools without sister schools have been surveyed and where a maximum number of five friends (female and male) could be nominated, this gives a hint on how the “fixed choice effect" and “boundary specification problem" may influence

degree distributions. The combined approach appears to be a good technique to extrapolate data in order to overcome those restrictions.

Additionally, as presented in Figure 3.10, the combined approach generates a link-probability versus distance curve that approximates those distributions from location-based online social networks such as Gowalla and Brightkite best. The approach shows therefore the most unambiguous results, indicating clearly the best setting for this algorithm.

However, the granularity of available information is still quite low. Recall that the social position of an individual is only defined by three activities, which is sports, participation in religious services and using the same public transport method. The use of more detailed information could probably further improve the results.

TABLE 3.3: **Recommendations for Application of the Proposed Network Extrapolation Approaches**

setting	**Social Circles**	**Bootstrapping**	**Combined**
$c \leq 500$	Social network extrapolation for intra-city networks, where acquaintances are locally restricted. Due to missing transport, social divide or decentralized city architecture, distant districts remain isolated.	-	-
$500 \leq c \leq 1000$	Extrapolation of social networks, where a globally connected network is to be created, sparse linking between dense local networks. "fixed choice effect" does not disturb the data e.g: close friendship network	-	Social network extrapolation within cities ($c = 800$ seems optimal). "fixed choice effect" and "boundary specification" are assumed to bias the original data.
$r \geq 0.89$	-	may resemble online testimonial network, yet to be investigated	-
$r \leq 0.89$	-	very high clustering. No conclusive social network.	

The table contains recommendations for the application of the proposed network extrapolation techniques and settings for distinct purposes.

3.6 Conclusion - Missing Data

This Chapter approached the suitability of three graph generation or respectively link-prediction techniques in order to impute [50] not at random missing (MNAR) about social ties within social network studies, and to interconnect disconnected components within the surveyed network. Hereby the target is to create a comprehensive model of the global network, stemming from original network data. The precision of the applied techniques had hereby secondary importance. Although a precise estimation of links would be desirable, a modeller may still settle with the creation of a reasonable model, especially because precision cannot be assessed with the available data-sets.

Firstly the Social Circles approach as proposed by Gilbert was modified such that it was suited to impute missing links between the locally isolated components of the original graph (**Approach 1**).

Secondly, a bootstrapping approach proposed by Leroy et. al [57] (**Approach 2**) was applied and thirdly a hybrid algorithm was created, combining features from 1 and 2 (**Approach 3**).

All algorithms were applied to the FUNDAJ data that serve as data-base for the use case model accompanying this work. The practical aim is to generate from the available student networks a global network representation for the city of Recife. Such globally interconnected network could then be applied to create a spreading model for the behavior "commitment at school".

The experiments show that all three approaches are able to impute data such that the number of isolated components within the graph decreases significantly. It was found that the three approaches create networks exhibiting degree distributions that resemble degree distributions of different real world networks.

While **Approach 1** appears to maintain the degree distributions of the original data-set, **Approach 3** leads to a degree distribution that can also be found in more interconnected networks in the real world. Further, **Approach 2** was able to reproduce degree distributions of close online friendship networks.

However, **Approach 2** was not able to create a completely interconnected network, when holding on to other restrictions such as a plausible average degree or clustering coefficient. Moreover, this approach led to an ill-shaped curve for the relation between link-probability and distance between two nodes. No evidence for the existence of a curve with such a shape in real world networks could be found.

In this regard, **Approach 1** and **3** performed better, as **Approach 1** produced link-probability distance curves very similar to the original data-set, while **Approach 3** was able to reproduce this relation very closely for the two locally restricted location-based online social networks Brightkite and Gowalla.

It is hence argued that the Social Circles **Approach 1** should be applied if the social scientist is willing to generate interconnected networks from unconnected components and acts upon the assumption that the macro network shall closely resemble micro structures. In other words, the approach put forward is able to predict links comprehensively if the assumption holds that missing data stems from "boundary specification" but that the "fixed choice effect" does not affect the data collection. An example for this could be the examination of best friends links between school children, where one assumes that most links are already established within the classroom and very few out-links are missing.

However, if one aims at creating interconnected networks, and does not expect that the network structure that can be observed in the original data approximates the macro network structure, the combined approach (**Approach 3**) seems to be even more adequate. In this case not only "Boundary specification" but also the "Fixed choice effect" causes missing data. Results indicate that the bootstrapping approach enables the scientist to reproduce close online relations. Yet, more than the other approaches, this one requires additional information about the individuals and hence suffers from data granularity issues. Also, the combined approach may solely be suited for data-sets that yield some information about the individuals that allows for calculating a social distance between them.

Moreover, it became clear that data granularity is a primary issue for the performance of the algorithms. The very general social circles approach seems to be very well suited for the task of completing a network of isolated components in a comprehensible way even if very little data about the nodes is available, or if this data is hard to structure for the underlying purpose.

Considering that even the large social network study that provided the data for this research only represents a relatively small sample of the whole population, it seems reasonable to settle for the good performance in generating networks that feature real world network characteristics. This is why for the purpose of this work, the Social Circles approach(**Approach 1**) is put forward to reach the aim of an interconnected network of students.

However, in combination with more information about the population (i.e, census data) the techniques that employ more personal information may contribute to an even more realistic network estimation.

Future work should deal with the application of the presented techniques to combine different data sources. In the underlying case, the survey data from FUNDAJ could hereby be connected to census data. In addition, further development of the social distance between a pair of nodes might improve performance of the combined approach and the social circles approach. Applied to the data-set used within this research, this may lead to an even more plausible globally connected network.

Further investigation of link-probability - distance curve of the bootstrapping network (**Approach 2**) seems to be interesting, especially when employing more individual information about the nodes. The additional groups may increase the quantity of generated links even for settings with a high threshold r and thereby heal the persisting problem of high clustering. In this case further comparison with data from online testimonial networks like Cyworld is recommended. The performance of the bootstrapping approach may be tested using more social groups with a more narrow definition. Additionally, testing the approaches on alternative data-sets may give further insights regarding the precision of the applied techniques. Although for our purpose, precision is not essential, other applications may require proof of it.

Chapter 4

Implementing the Model: Modeling Network Simulations

At this point, the thesis has addressed the selection of an adequate spreading model and the challenge that missing data represents for the modeler. Also, ways to generate a suitable network as a base for the social simulation have been described and evaluated. As a next step, the social scientist may specify in greater detail, how to simulate the spreading of behavior. In this regard, a first question is how sophisticated the individual agents should be designed and how autonomous they may act.

The research presented in Chapter 2 is based on a simple threshold model without any autonomous decision making. However, avoiding individual decision making and the concept of bounded rationality [41] disregards the human factor in a simulation which is subject to mistakes and not completely rational decisions. In order to overcome this lack of reality, the following Chapter presents first a possible solution for agent decision making and the incorporation of that decision making approach into a complete Agent-based simulation model. Subsequently, a proof of concept is delivered, demonstrating that the designed model is capable of mimicking human decision making in the given situation.

An Agent-based simulation model is proposed, where the modeling approach is tested observing only two agents of a complex simulation model in a micro scale. The Chapter describes and evaluates the implementation of a Learning Classifier System (LCS) as a decision making module for Agent-based models that incorporate social influence and heterogeneous interconnected agents. The aim is to develop a decision mechanism that resembles bounded rational human decision making (in the sense of H. A. Simon's approach to a more realistic theory of human economic decision making [80]) well and that incorporates imperfect information as a feature from real decision making situations. The use case of the simulation model is, in accordance with the foregoing and following Chapters, the decision about engagement at school of individuals, measured via the achieved mark of those individuals. Experiments with two interconnected agents are conducted in three distinct scenario settings. The simulation study shows that the proposed LCS performs well in achieving good solutions for both agents.

4.1 Introduction - a Justification for the Use of Agent-based Network Models

Currently, General Equilibrium Models [7] represent the most popular paradigm for macroeconomic simulation and thereby the most popular measure for political decision support. However, those models are based on strong neo-classical assumptions like rational decision making and perfect information for all actors. These assumptions do obviously not hold in the real world and lead to a stereotype average consumer, that is the rational individual or Homo Oeoconomicus. Critics on Homo Oeconomicus became louder during the last decade due to the unrealistic assumptions of the underlying model and the recent failure of rational individual based models [8]. These assumptions also suppose that our highly heterogeneous societies can be understood by investigating the behavior of rational average individuals and their communication and group behavior. It is argued against that irrationality does not exist, or at least not affect the crowd's behavior [9].

In order to better understand and predict human behavior, the concept of Agent-based modeling came up as an alternative for economists. Agent-based models use autonomous acting, communicating computer programs, the so called agents that are able to decide in a bounded rational way [10]. Agents within these models may resemble individuals, consumers or juristic persons. Agent-based models thereby are enabled, to better model human heterogeneity and thus create a more sophisticated image of reality. A social effect on educational choices has been confirmed and also successfully modeled with the Agent-based methodology [81].

Complementary, the research area of Social Network Science and Complex Networks suggests that human decisions are not entirely autonomous, but influenced by peers, siblings or parents [11]. This influence may occur through spread of information or contagion of behavior via social networks. The former foils the assumption of perfect information, the latter challenges fully rational decisions.

This motivates the attempt to join findings from Social Network Science and Agent-based modeling in order to create models that better represent reality, facilitating simulation of societies and prediction of policy effects.

In order to set-up a simulation model that addresses the stated shortcomings of state of the art General Equilibrium Models and copes with opinion dynamics in social networks, the agents within the models need to be equipped with an adequate decision making mechanism. Such a mechanism may approximate human decision making in the situation under investigation, enhancing the credibility and accuracy of the model. Moreover, the mechanism must be capable of coping with a dynamic environment. The presented research proposes such a decision mechanism for Agent-based models, incorporating network diffusion processes.

In an early work, Holland proposes Learning Classifier Systems (LCS) as a good option to mimic human decision making in Agent-based models. Principally he

argues in favor of LCS because they enable the agent to allocate environmental situations to broad categories which are progressively refined by the experience made. This in turn enables the agent to build internal models of the world, while non of the models is immutable, but always provisional and subject to change [82].

Further, Classifier Systems have been shown to be able to learn to play nash-markov equilibria both with and without the presence of imitation [83], [84]. Therefore, a LCS is implemented in order to make allowance for the often posited characterization of the human mind as a system to classify things and situations. The model is set up using the *NetLogo* [85] environment.

This Chapter presents a proof of concept for the utilization of LCS as an agent learning representation in Agent-based social simulations.

On the basis of the FUNDAJ data-set presented in Chapter 2 and 3, the use case for the modeling approach is the schooling decision of children. Important determinants of schooling success are the motivation of parents to support their children at school and the commitment of children to study, as well as the quality of schools. Children have to decide to which degree they commit themselves to their education. As a motivation for this commitment serves the question if education pays off or not (expected utility). As schooling success depends on a large number of influence factors, such as socioeconomic status, peer influence and current economic activity, it might be assumed that children cannot assess that expected utility but rather base their decision upon experience and peer information. Moreover, subjective perception, limited processing capacities and incomplete information may influence expected utility calculation of individuals.

4.2 Background - Implementation

This Section gives a general overview on recent advances in the fields important to the presented research, namely diffusion processes in social networks, Agent-based Computational Economics and Learning Classifier Systems.

4.2.1 Diffusion in Social Networks

Social influence and contagion, as well as spread of behavior and information through social networks has been documented in a wide range of cases [11]. This indicates the existence of those effects on the schooling decision of individuals. Marques [86] reveals the great differences between social networks of the poor and those of more wealthy people, which further encourages the considering of social network effects while studying social phenomena.

Consequently, scientists aimed at developing models to understand those spreading and contagion processes. Thus, econometric approaches have been developed in order to capture peer effects on schooling behavior of pupils [87]. However, even though this approaches incorporate empirical peer effects, they do not consider

the very mechanism of behavior spreading, nor bounded rational individuals. One simple approach to capture diffusion processes is to model them as Coordination-Games [15] (see Chapter 2) or employing group decision making approaches [30].

4.2.2 Agent-based Models - Agent-based Economics

According to Holland [88], Agent-based modeling (ABM) describes the study of systems consisting of autonomous computational agents. The agents may be designed heterogeneously and are able to interact, which enables the ABMs to reproduce macro phenomena that emerge from micro level behavior. The sub-fields of Agent-based Social Simulation (ABSS) and Agent-based Computational Economics (ABC) join the fields of Agent-based Computing, Computational Simulation and respectively Social Sciences [89] or Economics [90], where applications reach from demography [91] to tax compliance [92] or school effectiveness [93]. Using ABM to simulate social or economic contexts forces the researcher to debug and understand macro phenomena better, while large experimental studies may be conducted without numerical or ethical concerns arising in real world experimental setups. Contrary to traditional economic models, ABM enables the researcher to incorporate the imperfection of human rationality as well as limited information availability to the model. In addition, the iterative interaction of agents triggers insights that may be overseen in general equilibrium approaches. A detailed summary of sociology in ABSS can be found in [94], while [95] summarizes applications in Agent-based Computational Economics.

Literature on Agent-based Computational Economics suggest very distinct approaches to model agent decision making. Approaches employ unconscious techniques like reinforcement learning, routine-learning approaches like replicator dynamics, belief learning methods as classifier systems or Bayesian approaches [96]. many of them have been proven to produce outcomes that coincide with findings from experimental economics and even econometrics [97].

4.2.3 Learning Classifier Systems

Learning Classifier Systems (LCS) are rule based programs. They usually contain a Genetic Algorithm to manipulate the set of rules they operate on and a Reinforcement Learning part that aims at choosing the best performing rules [98]. Holland proposed LCS first as a model of the emergence of cognition [88]. Classifier Systems are regarded as an approximation to human decision making, given a perceived situation [96] although they are not belief based, which means that agents are not conscious about the existence of other agents within their environment [97].

According to Brenner [96], Classifier Systems consist of a set of condition-action rules, where the conditions $\bar{c}$ describing the perceived state and the actions $\bar{a}$, representing the respective action to be taken are stored as feature strings of the form $\{c_1, c_2, ..., c_n\}$ or respectively $\{a_1, a_2, ..., a_n\}$. The set of condition - action rules $R_i (i =$

$1, 2, ..., n$) combines then a condition string with an action string. Whereas c_{ij} or a_{ij} may be represented as a wild-card #, indicating that this feature applies independently from the given situation. For each iteration, the current signal $s = \{s_1, s_2, ..., s_n\}$ is compared to the condition strings $\bar{c}$. The most adequate of those rules with corresponding $\bar{c}$ is being chosen for execution. For the purpose of choice, each rule is being assigned a *Specificity* value and a *Strength* value. The *Specificity* determines the number of wild-cards within the rule, while the *Strength* is defined by the pay-off, the rule generated in preceding iterations. The value $B(R_i)$ is calculated according to Equation 4.1, where α, β and γ are parameters. Accordingly, the corresponding rule with the maximum value of $B(R_i)$ is regarded the most adequate rule.

$$B(R_i) = [\alpha]([\beta] + [\gamma]Specificity(R_i))Strength(t, R_i) \tag{4.1}$$

The *Strength* of each rule R_i at time t is hereby calculated according to Equation 4.2.

$$Strength(t+1, R_i) = Strength(t, R_i) + Payoff(t) - B(R_i) \tag{4.2}$$

Subsequently, the Classifier System employs a genetic operator that allows for creating new rules from the existing best performing rules and forgetting rules that did not perform well in the past.

4.3 Problem Environment

The agents within the presented simulation model are embedded in an environment consisting of their peers[1] and an individual socio-economic environment represented by individual variables. The aim is to model the behavior "commitment at school" which cannot be observed easily. Hence the mark in mathematics of the respective pupil is employed as a proxy for the engagement at school. The agents within the model iteratively decide what mark to achieve in the next iteration. In accordance with the findings from Chapter 2, it is assumed that agents benefit from aligning their behavior with peer behavior. Thus, an agent's utility is affected by the behavior her peers exhibit. Both, individual socio-economic status and peer social-economic status hereby affect the utility. Moreover, the agents are unaware of their own utility function and hence have to learn which action pleases them most.

Perceptions are represented as condition strings E of the form $\{s, p_1, p_2, ...p_n\}$, where s stands for the mark of the current individual and p_i stands for the mark of peer i. Subsequently, it is explained, how those perceived condition strings are processed in the decision module set up as a Classifier System. In every case, the agent decides on a set of actions, that may include all possible marks within the range $[0, 100]$.

[1] for the use case of the simulation, peers are thought of as friends within the friendship network of pupils according to (extrapolated) network data surveyed by FUNDAJ

4.4 The Learning Classifier System (LCS) Decision Mechanism

As stated above, the proposed LCS implementation is based upon a population of conditions, linked to corresponding actions. After setting up the system, those condition-action strings are selected according to their suitability for the currently perceived situation and their performance in the past. After each iteration, the selected condition-action string is being evaluated with the achieved fitness according to a given Fitness Function. An evolutionary process ensures that the decisions made continuously improve. The following subsections describe those features of the LCS in greater detail.

4.4.1 Condition-Action Rules

The classifier is based on a set of condition-action rules R of the form $\bar{c} -> \bar{a}$, where each $\bar{c}$ represents a condition string $\{c_1, c_2, ...c_n\}$. Respectively, $\bar{a}$ represents the action to be taken if the rule is selected. In the given scenario $\bar{c}$ contains the mark in mathematics of the respective agent as well as the current mark of her peer. Accordingly, the action $\bar{a}$ may be any mark between 0 and 100 that the agent will achieve in the subsequent iteration. The length n of $\bar{c}$ is given by the formula $n = d + 1$, where d denotes the degree of the respective agent. c_i stands for the interval $[x_i, y_i]$ with $x_i, y_i \in [0, 100], y_i \geq x_i$ but can adopt the # symbol also, indicating that this digit of the condition string matches all possible values of s or p_i respectively. The first digit of $\bar{c}$ narrows the mark of the respective agent, while the remaining digits narrow the mark of her peers. For example, one $\bar{c}$ may be $\{[0, 10], [80, 100]\}$. This condition would for instance match a situation where agent 1 achieves a mark of 7 and agent 2 achieves a mark of 90. with a corresponding $\bar{a} = 56$, agent 1 would change her mark for the next iteration to 56. At each time step, the algorithm creates the list of matching condition action strings M_i. M_i contains those strings for which the condition $\forall x \in E, x_i \in c_i$ holds.

4.4.2 Setup and Selection Mechanism

To setup the system, a number of condition-action-rules is created randomly. Here for each rule to be created, a random interval is set for each digit of the condition-string. The respective action of the condition-action-string is then drawn from a normal distribution with variance $VAR(x)_1$, while the mean is set to the initial mark of the respective agent. This approach for the set-up procedure has been chosen to avoid an unrealistic initial disturbance of the system.

Calculation of $Strength$ and $B(R_i)$ occurs according to Equation 4.2 and Equation 4.1 respectively for all $R_i \in M_i$. Subsequently, a roulette wheel mechanism ensures that the action of that R_i with the highest $B(R_i)$ is most likely to be taken, while the likelihood for the selection of $R_i \in M$ decreases with decreasing strength.

If R does not contain any rule that is compatible to the current perception string - meaning that $M_i = \emptyset$ -, that rule in R that is most similar to the current perception E mutates so that it matches E. Hereby the action of the mutated string is also drawn from a normal distribution where the mean is the currently performed mark of the agent and variance is $VAR(x)_3$.

4.4.3 Evolutionary Process

Furthermore, an evolutionary process is implemented, aiming at continuous improvement of the solutions found. Hereby a fraction of the weakest rules $(death - rate)$ in M_i is being deleted from R and new rules are created, recombining the n strongest rules in M via a cross-over operator until the original number of rules in R is reached. In order to ensure diversity, an additional mutation operator is introduced: A random mutation process starts with a probability of $mutation - rate$, altering random characters of the condition string of a randomly chosen rule $R_i \in M_i$ that is not the currently best performing rule. The character that indicates the action of the condition-action-string to be mutated is drawn from a normal distribution with variance $VAR(x)_2$ while the mean is set to the currently adopted mark of the respective agent.

Figure 4.1 illustrates this Classifier System for the simple case of an agent with degree 2.

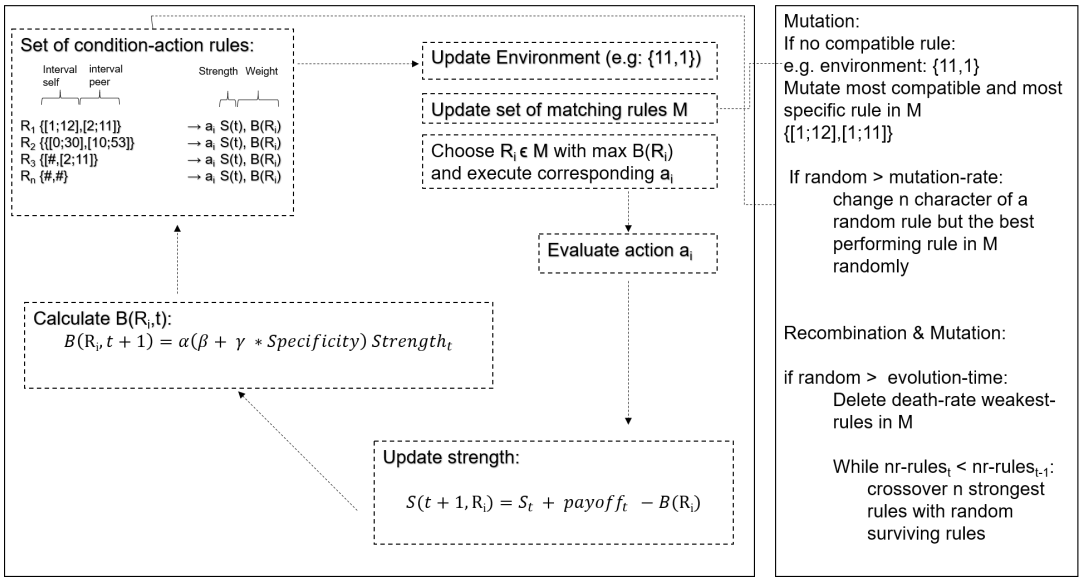


FIGURE 4.1: **Classifier System.** The Figure is an illustration of the rule based decision making mechanism.

4.4.4 Evaluate Action

The evaluation of the fitness or utility, an action taken by the agent causes, is being measured by a utility function. The utility function proposed in [87] is implemented

as presented in Equation 4.3. In this case $\theta_i(y)$ is a component that introduces exogenous heterogeneity to the model and δ is the imitation-factor of the model, controlling the peer influence. Moreover, x_i represents the mark achieved by the respective agent i and g_i stands for the binary peer matrix of the agent.

$$U_i(x_i, g_i) = [\mu g_i + \theta_i(y_i)]x_i - \frac{1}{2}x_i^2 + \delta \sum_{j=1}^{n} g_{ij}x_i x_j \tag{4.3}$$

The exogenous heterogeneity component $\theta_i(y_a)$ is computed according to Equation 4.4. y is a vector of variables that resemble observable differences between individuals, such as race, age, and other socio-economic variables. σ and ϕ are parameter vectors.

$$\theta_i(y) = \sum_{m=1}^{M} \sigma_m y_i^m + \frac{1}{g_i} \sum_{m=1}^{M} \sum_{j=1}^{n} \phi_m g_{ij} y_j^m \tag{4.4}$$

This fitness function not only introduces wide individual heterogeneity, but also accounts for a strategic complementarity in efforts [87]. This means that if the peer of agent i, agent j increases her behavior level, then agent i will receive increasing marginal utility, if she also increases her behavior level. Hence, we incorporate imitation to the model in line with the findings from Chapter 2.
Table 4.1 summarizes the model parameters and contains a brief explanation for each parameter.

TABLE 4.1: **Explanation of Model Parameters**

Model modules	Parameters	Explanation
Strength Calculation	α	controls the importance of past performance for the selection of a rule $R_i \in M_i$
	β	controls the importance of past performance for the selection of a rule $R_i \in M_i$
	γ	controls the importance of specificity of rules in the LCS
Genetic Operators	$mutation - rate$	controls how frequently rules in the LCS are replaced by randomly created rules
	$death - rate$	controls which share of the population of rules within the LCS is re -placed by newly created rules (cross-over recombination)
	$evolution - time$	controls how often an evolutionary process is triggered for all agents
LCS	$nr - action - rules$	controls how many condition-action-rules an agent possesses
	$VAR(x)_1$, $VAR(x)_2$, $VAR(x)_3$	Variance of the normal distributions in the generation and mutation of action rules. Control how different from the initial real mark (or respectively the current state of the simulation) the action of a newly created or mutated rule may be.
Utility Function	δ	Imitation Factor, controls the weight of peer behavior within the utility function
	σ	Parameter vector, assigns weights to the individual variables of each agent
	ϕ	Parameter vector, assigns weights to the individual variables of peers

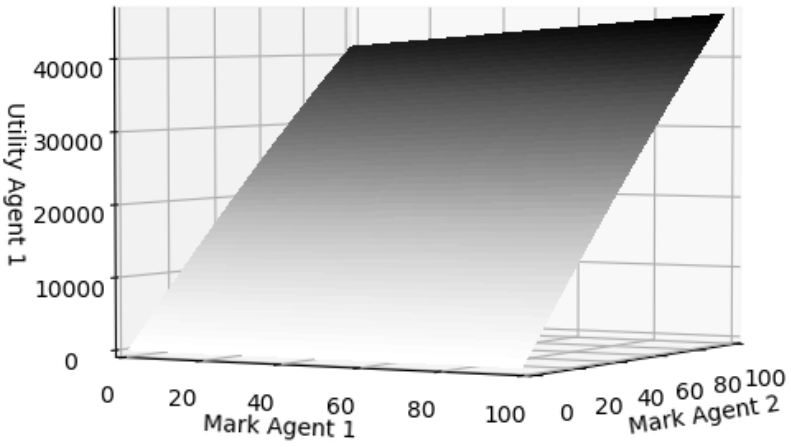
4.5 Experiments With a Simple Set-up

Seeking to verify, if the implemented decision making algorithm is capable of mimicking human decision making in the situation of interest, the most simple model set-up is chosen, containing two interconnected agents. The parameter vectors σ and ϕ of the utility function $U_i(x_i, g_i)$ are chosen so that clear strategies emerge for each agent. For the purpose of experimentation, the three distinct strategy settings are defined and listed below.

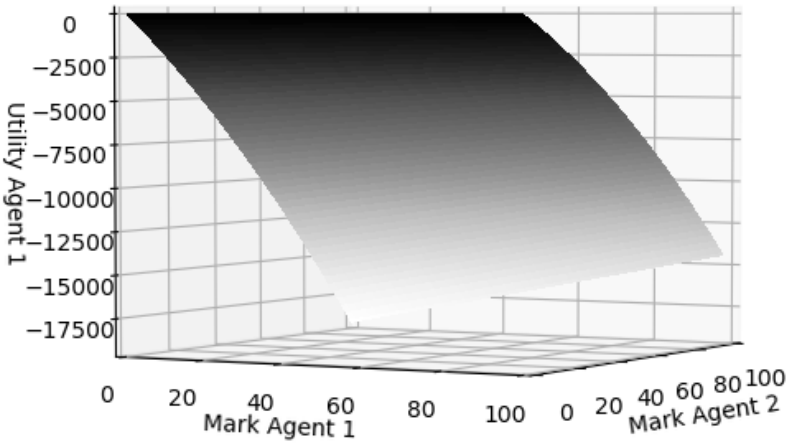
(i) "Good mark": both agents may always prefer to achieve the better mark, this is achieved by setting σ and δ so that $\frac{dU}{dx} > 0$.

(ii) "Bad mark": both agents may always prefer to achieve the worse mark, this is achieved by setting σ and δ so that $\frac{dU}{dx} < 0$.

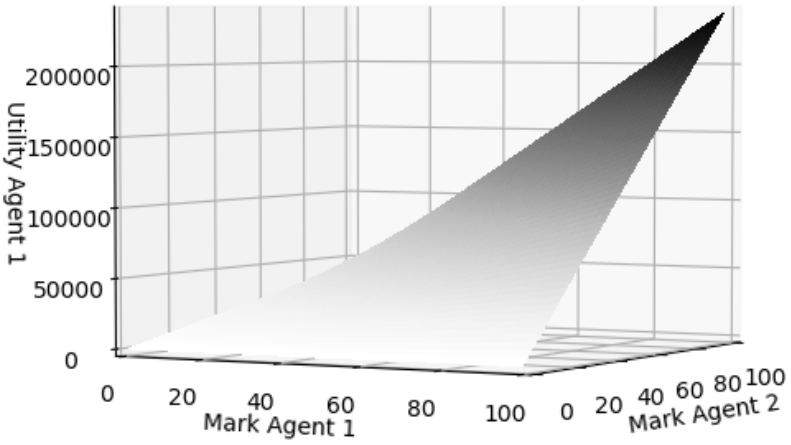
(iii) "Good mark imitation": achieving a good mark is a dominant strategy for both agents. However, peer behavior heavily influences the utility outcome. The parameter vectors are set as in (i) and the imitation factor γ is set to 20. For each scenario, the vector of variables resembling observable differences between individuals, y_a is set randomly in order to create two random agents. Figure 4.2 illustrates the respective utility for agent 1 as a function of her achieved mark $mark1$ and the achieved mark of her peer $mark2$.



(a) Utility Function for Dominant Strategy Agent 1: Good mark (i)



(b) Utility Function for Dominant Strategy Agent 1: Bad mark (ii)



(c) Utility Function for Dominant Strategy Agent 1: Good mark & factor imitation = 20 (iii)

FIGURE 4.2: **Utility Functions.** The Figure illustrates the utility function of agent 1 for the three strategy settings (i) "Good mark", (ii) "Bad mark" and (iii) "Good mark imitation".

The parameters are set as presented in Table 4.2. The model parameters have been chosen manually, analyzing the model behavior. As this Chapter serves as a proof of concept, it is not the purpose to find the best performing parameter setting, but merely one that performs sufficiently well. If more elaborated methods for parameter search where applied, measures should be taken to make sure the parameters are not over fitted.

TABLE 4.2: **Model Parameters for Experiments**

Model modules	Parameters	Values
Strength Calculation	α	0.74
	β	0.83
	γ	0.42
Genetic Operators	$mutation - rate$	0.3
	$death - rate$	0.75
	$evolution - time$	5
LCS	$nr - action - rules$	200
	$VAR(x)_1$	4
	$VAR(x)_2$**)**	40
	$VAR(x)_3$**)**	10
Utility Function	δ	$(i)(ii) : 0.5; (iii) : 20$
	σ	*
	ϕ	*

*set to create the respective strategy (i), (ii) or (iii).

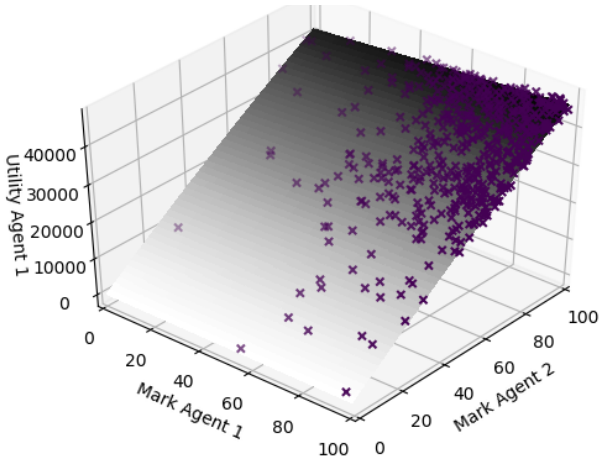
In order to assess, if the model behavior fulfills the expectations, it is measured, if the algorithm is capable of finding good solutions for each scenario. As the aims is to mimic human behavior, fully accurate and rational decision making is explicitly not expected. Though, the agents are expected to demonstrate a tendency towards the optimal solution while sporadic not optimal solutions are tolerated. Moreover, a learning process should be observable throughout run-time. Ultimately a human-like agent is expected to react on changes in her environment, namely the change of behavior of her peers and the alteration of her own situation. The degree of target achievement is measured here examining the probability for an agent to change the current action subject to recent alterations of the environmental variables, peer behavior and self-behavior.

The models are run 500 times with a run-time of 500 iterations.

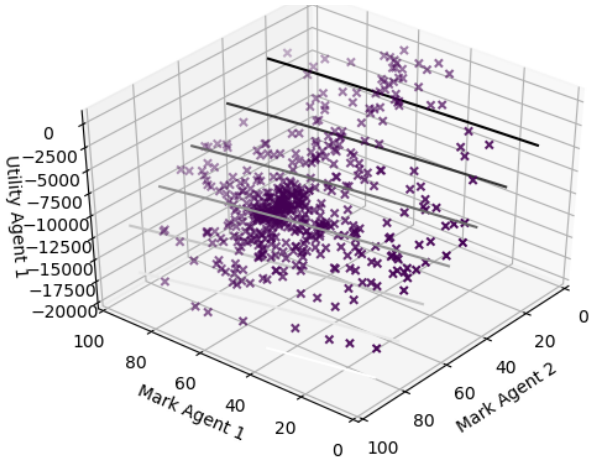
4.5.1 Overall Performance - Learning Process

The finally achieved mark of the agents after each run may be revised in Figure 4.3 for each scenario. Here each cross indicates the final mark of agent1 and agent2

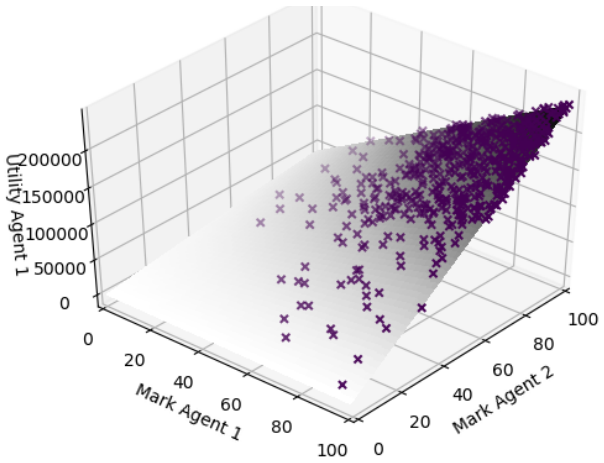
and the respective utility derived by agent1 after 500 iterations. One may observe that for scenarios (i) and (iii) both agents achieved final marks close to the function optimum. Also, for the majority of simulations, marks for both agents can be found in the upper half of the scale. The best possible solution in scenario (ii) would be a mark of 0 for both agents. however, as Figure 4.3(b) reveals, the agents did not achieve this optimal solution frequently. Nevertheless, a tendency towards lower marks is observable.



(a) Simulation Results for 500 Simulations After 200 Iterations for Dominant Strategy Agent 1: Good mark (i)



(b) Simulation Results for 500 Simulations after 200 Iterations for Dominant Strategy Agent 1: Bad Mark (ii



(c) Simulation Results for 500 Simulations after 200 Iterations for Dominant Strategy Agent 1: Good Mark & Factor Imitation = 20 (iii)

FIGURE 4.3: **Results for Experiments with LCS Decision Making Mechanism.** The Figure presents results obtained after 200 iterations of the two-agent model. Although the agents present non-optimum decisions, a tendency towards the maximum is observable.

4.5.2 Run-time Performance

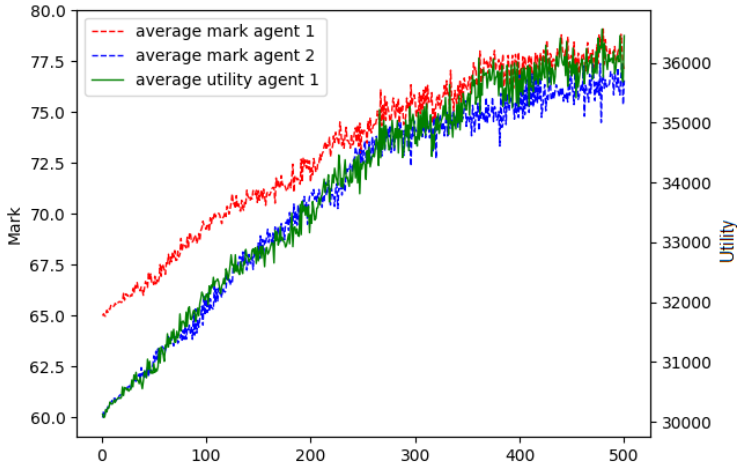
To investigate the model behavior for each iteration, an analysis was conducted for the marks achieved by both agents, as well as the utility for agent 1.

Figure 4.4 illustrates the average outcome for each iteration in 500 simulations. The solid green line indicates the averagely achieved utility of agent 1 for each iteration, while the dashed red line and the dashed blue line indicate the averagely achieved mark of agent 1 and agent 2 respectively.

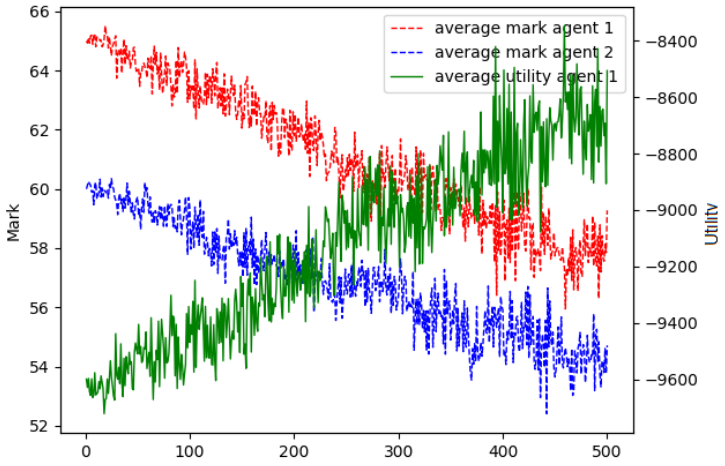
The plot for scenario (i) reveals that all indicators develop positively until the end of the run-time. An average final mark just below 80 is achieved.

Plotting the average outcomes for scenario (ii) indicates a negative development of marks throughout the run-time and respectively increasing average utility values. Finally achieved average mark for both agents lies below 60 while the achieved average utility amounts above -8800. Recall that the best possible decision for this scenario for both agents would be a final mark of 0 and respectively a utility of 0.

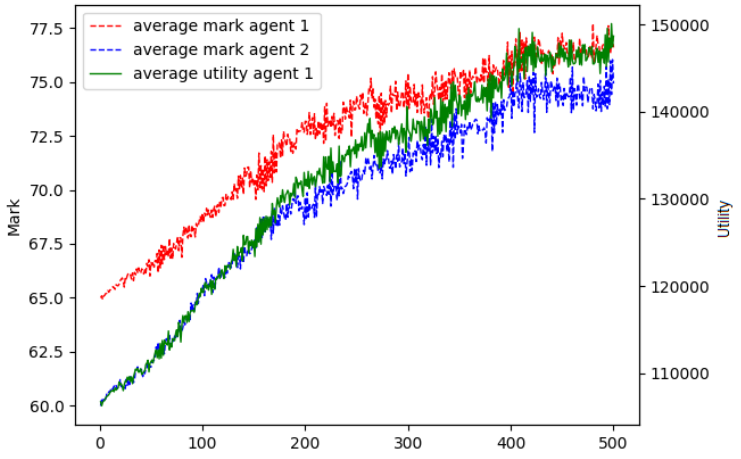
Also, utility is understood as an abstract value for the comparison of decisions. Hence, negative utility values do not have a special meaning. Scenario (iii) yields average mark and -utility development comparable to scenario (i).



(a) Agent 1: Good Mark (i)



(b) Agent 1: Bad Mark (ii)



(c) Agent 1: Good Mark & Factor Imitation = 20 (iii))

FIGURE 4.4: **Average Results per Iteration for Experiments with the LCS Decision Making Mechanism.** The Figure presents results for 500 experiments with the two-agent model for three scenarios. A continuous improvement of utility can be observed for all scenarios. Hence, the agents possess the ability to learn.

Moreover, the run-time analysis encompasses examination of agent behavior over time. In order to observe how repeatedly chosen actions affect the disposition of agents to try out different behavioral patterns, the frequency of occurrences of behavioral change have been related to the number of iterations with unchanged behavior preceding that alteration.

Figure 4.5 illustrates the respective outcomes. Here the green dashed line indicates how often a change of behavior was observed throughout all experiments after x iterations. The red dashed line represents the probability density function of the distribution of x. It becomes clear that the vast majority of action changes occurs after few repetitions of the same behavior. Very low frequencies are observed for more than 10 iterations. In order to ensure the validity of the calculated frequencies, x that occurred less than 20 times have not been considered for this analysis.

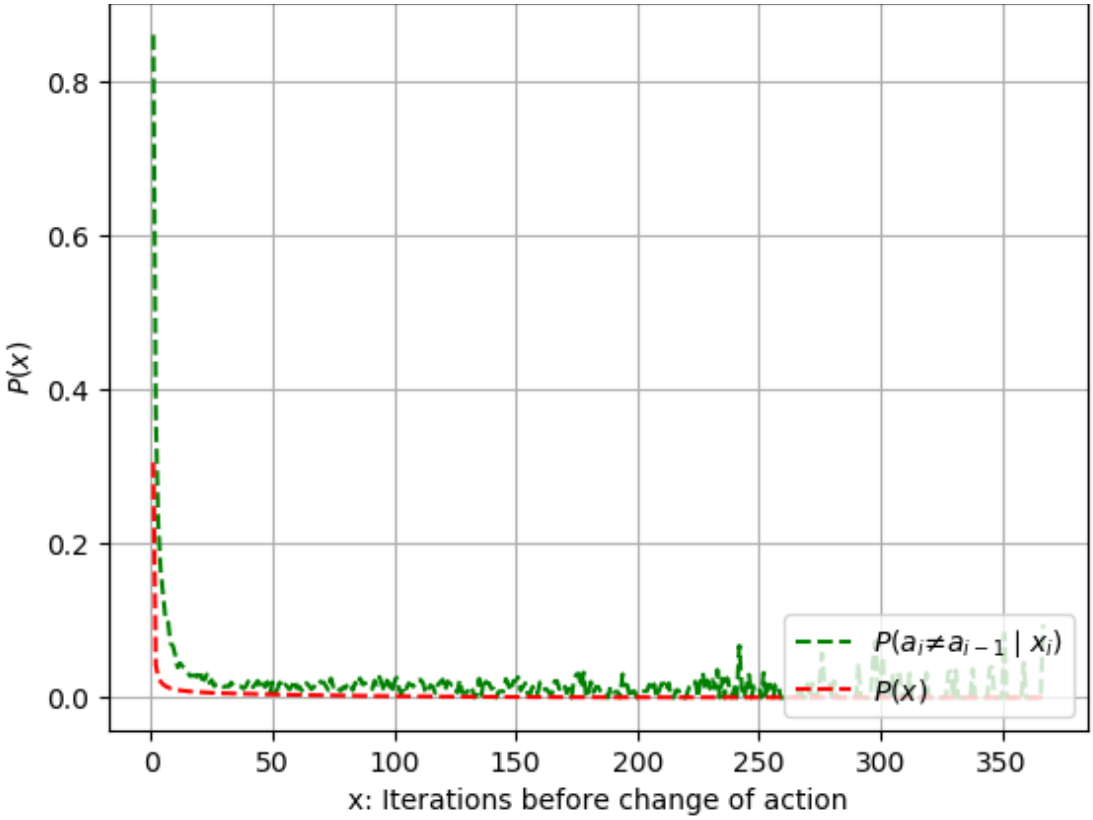


FIGURE 4.5: **Frequency of Action Change in Experiments with the LCS Decision Making Mechanism.** The Figure illustrates the frequency of a change of action of an agent in relation to the preceding number of repetitions of the same behavior. It becomes clear that the vast majority of action changes occurs after few repetitions of the same behavior.

4.5.3 Reaction to Variation of Peer Behavior

Finally it is investigated how the agent responds to changes in peer behavior and in own behavior. To this purpose the variable Δ is calculated according to Equation 4.5, where a_k^1 indicates the action of agent 1 taken in iteration k and a_k^2 is respectively the

action of agent 2 in iteration k. x_i indicates the mark of agent 1 at iteration i and y_i the mark of agent 2 at iteration i.

$$\Delta_i = \sqrt{(\sum_{i=k}^{j}(x_{i-1} - x_i))^2} + \sqrt{(\sum_{i=k}^{j}(y_{i-1} - y_i))^2} \tag{4.5}$$

Thus, Δ describes the degree of change in the environment between the current and the preceding iteration.

Figure 4.6 plots the cumulative frequency of Δ in the 2.5×10^5 iterations of the 500 experiments as a red solid line. It thereby describes how the degrees of environmental change are distributed among the whole simulation. The green line however, indicates the cumulative frequency of Δ in the subset of iterations that actually triggered a change of action for the observed agent $(a_k \neq a_{k-1})$. In other words: the green line indicates how often the environment changed to certain degree and agent1 changed her action. As the relations presented in this Figure are very similar for all three scenarios, The outcomes for scenario (i) are demonstrated.

For $\Delta > 10$, the green line appears to grow much steeper than the red line. Also, the red plot appears to be much more concave than the green plot. The more concave shape of the red plot indicates that Δ is represented less than proportional within the set of Δ that actually triggered an action change for low Δ, while the opposite holds as Δ grows. Thus, it becomes clear that the probability for an agent to change the current behavior is substantially higher if the environment, respectively the peer behavior, changes significantly.

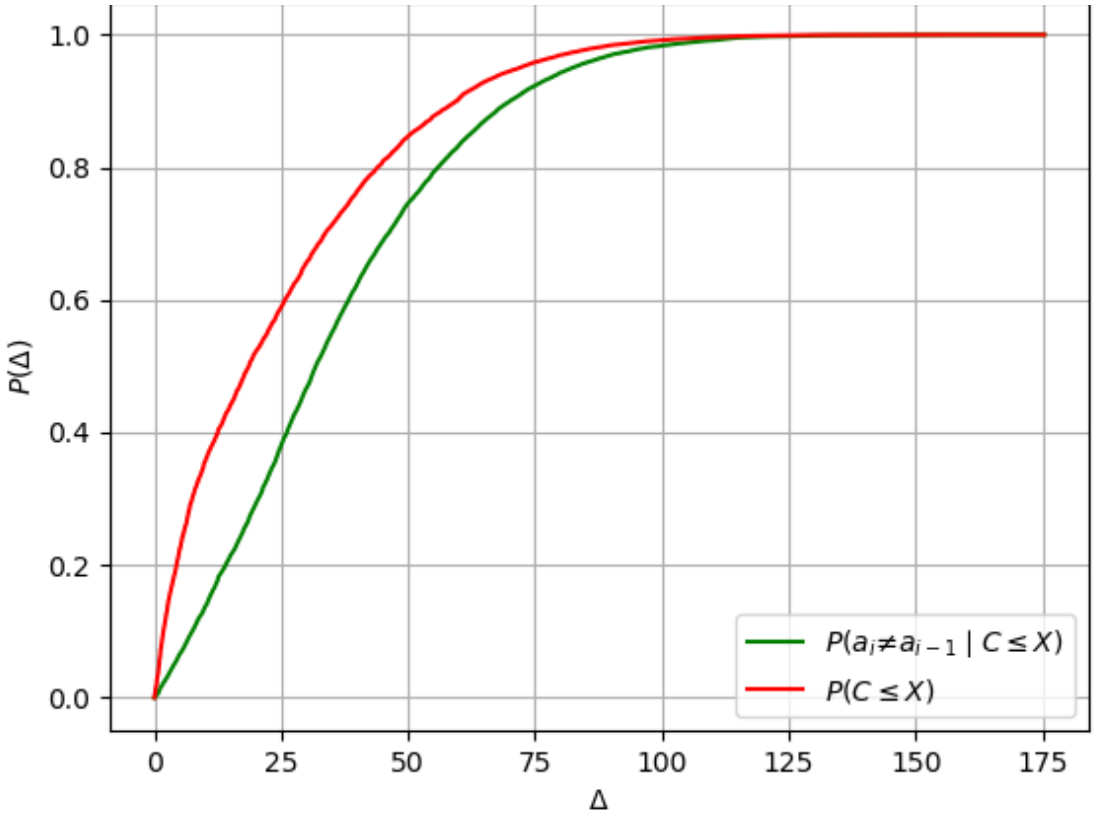


FIGURE 4.6: **Frequencies of Cumulative Environmental Change in Experiments with the LCS Decision Making Mechanism.** The Figure shows the cumulative frequency of Δ in the simulation. It becomes clear that the probability for an agent to change the current behavior is substantially higher if the environment, respectively the peer behavior, changes significantly.

4.6 Discussion - Can the Agents Mimic Human Decision Making?

As stated above, this Section seeks to present a solution for human alike agent decision making. Hence the decision making algorithm shall account for bounded rational decisions that may not be optimal in all cases but demonstrate a tendency towards good decisions. The results presented in Section 4.5.1 indicate that the proposed LCS is capable of delivering good solutions for differently shaped utility functions. In the examined simple settings with only two interacting agents, solutions yielding high utility were encountered in the majority of simulations. However, the algorithm also exhibited misjudgment and biased decisions that may also be expected from human decision makers. Difficulties were particularly encountered in situations with negative pay-offs. It could be argued that humans particularly struggle with situations where the outcome is always negative. However, there may be also alternative parameter settings that help the agents to better perform in negative utility functions. Moreover, it is not clear yet, if the implemented LCS also performs

well in more complicated settings with a larger number of heterogeneous peers and high imitation utility.

Furthermore, the realistic agents are expected to exhibit the ability to learn from past experiences. Section 4.5.2 illustrates that on average, the agent´s decision improves with increasing run-time specifically for the scenarios (i) and (iii). The decisions in scenario (iii) also improve, yet on a rather low pace. This may indicate that the LCS-implementation is more sensible to negative pay-offs. However, the continuously positive developing average utility is a strong signal that the agents exhibit learning behavior.

Finally, it was posited that agents should react sensibly to changes in peer behavior. In Section 4.5.3 it is found that the probability for an agent to change her current action is significantly lower, when the cumulative difference of her mark and of the mark of her peer to the respective marks after the preceding action change is close to zero. Shorter: an agent is more likely to change behavior, if the environment changes. This analysis also revealed that probability of action change increases with increasing cumulative difference of the environment. Hence, it can be argued that the agents do react on change in peer behavior and self behavior.

The run-time analysis further revealed that agents are significantly less likely to change their course of action, once a certain action has been executed repeatedly (see Figure 4.5). Most alterations in behavior have been observed in a short period after experimenting a new behavior. This resembles habituation in human beings, a behavioral feature that frequently occurs in reality. Once one created a habit like for example drinking a cup of coffee after lunch, it becomes quite difficult to change that behavior even if the environment changes.

4.7 Conclusion - Agent Decision Making

Within this Chapter the implementation of a Learning Classifier System is proposed as a decision making module for Agent-based models that incorporate social influence and heterogeneous interconnected agents. The aim is to develop a decision mechanism that resembles bounded rational human decision making well and that incorporates imperfect information as a feature from real decision making situations. The use case of the simulation model is the decision about engagement at school of individuals, measured via the achieved mark of those individuals.

Experiments with two interconnected agents are conducted in three distinct scenario settings:

(i) Firstly, a scenario is set-up, where the dominant strategy for both agents is to achieve the best possible mark.

(ii) Secondly, the environment is set so that the best possible decision for both observed agents would be not to engage at school at all and consequently achieve the worst possible mark.

(iii) Finally, a scenario with high utility derived from imitation of peer behavior is investigated.

The simulation study shows that the proposed LCS performs well in achieving good solutions for both agents for the respective scenarios. Still, optimization is not accurate but biased by peer decisions and habit and thus well resembles human decision making. Moreover, a learning effect could be identified which is essential when mimicking human decision making. Finally it could be shown that the agents react to environmental change while exhibiting a tendency to create habits which are not changed even if the environment changes.

Summarizing, it could be shown that the application of LCS can in fact be an adequate approach to mimic human decision making in Agent-based simulations. However, further study is required in order to verify if the LCS performs well also in more complicated settings, incorporating larger numbers of heterogeneous interconnected agents and settings incorporating exclusively negative pay-offs. Within this Chapter, only one well performing calibration of the simulation model was tested. More detailed analysis of model behavior under different parameter settings would most certainly contribute to further develop the decision module.

Chapter 5

Estimation and Calibration of Large Scale Network Simulations: Fitting the Model to Reality

So far, the selection of an adequate spreading model and the generation of a coherent network have been addressed. Building on this, a simulation model has been set up in the foregoing Chapter, that comprises bounded rational agents that interact and influence each others decision making. With the aim of conducting useful what-if-simulations and thereby derive insights for decision makers, the set-up model must be applied at a larger scale.

This Chapter addresses at the hand of the use case of educational commitment of pupils, how a simulation for a large scale society of agents may be calibrated to real data. To this purpose, the simulation model presented in Chapter 4 is set-up at a larger scale, incorporating a population of more than 4000 interconnected agents. The extrapolation of the exemplary model set-up from Chapter 4 is described in 5.1. Subsequently, the difficulties that arise when calibrating and estimating complex network models that incorporate bounded rational agents and a wide set of parameters are addressed in 5.2.

When extrapolating simulation models on a large scale, the choice of the parameters of those models becomes a challenge. Therefore the process of parameter selection including upcoming challenges is described and potential approaches to improve the process are presented in 5.3.

5.1 Extrapolation of the Model to a Large Scale

Pure extrapolation of the Agent-based simulation model from Chapter 4 is trivial. Instead of generating a two-agent model according to the described set-up, the setup procedure is conducted based upon the 4191 students from the the study *'Determinantes do desempenho escolar na rede de ensino fundamental do Recife'* [19]. Those data contain among others the social network of the pupils and their performance in the subject maths at the beginning and at the end of the year. Children were asked to

nominate their 5 best friends. Moreover, a large set of individual socioeconomic variables is available for each student within this data-set.

To generate a globally interconnected network, the social circles approach is applied according to Chapter 3, generating the interconnected network data-set illustrated in Figure 5.1.

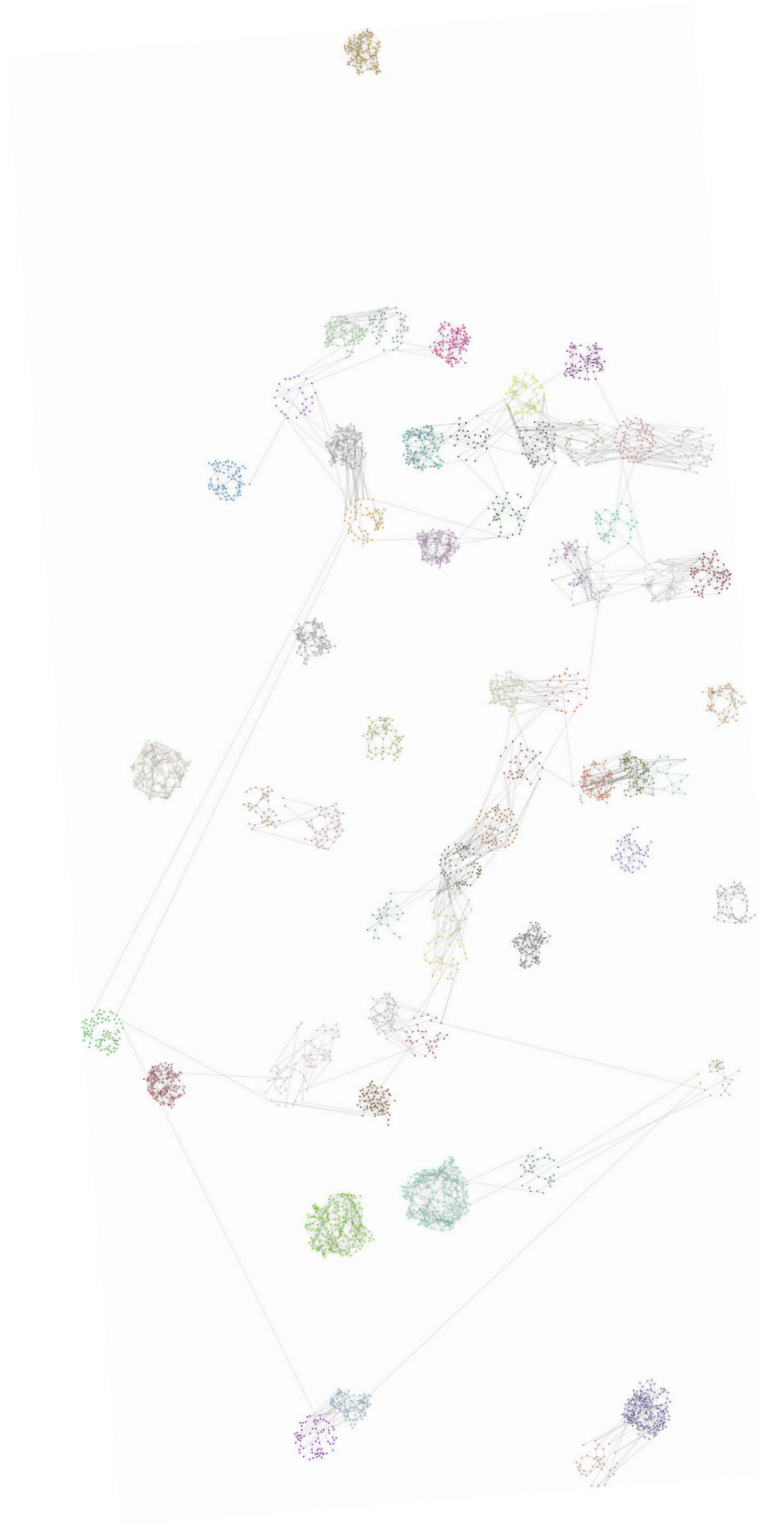


FIGURE 5.1: **Network Created by Social Circles Approach.** The Figure presents the individual pupils that participated in the FUNDAJ-Survey. Pupils are colored according to their school. Location of pupils is assigned by a Fruchtermann-Reingold algorithm within a radius of n^2 around the location of their school within the city of Recife. n indicates the number of pupils of the respective school. Grey lines indicate friendships between pupils as registered by the survey, as well as friendships estimated by the social circles method applying $c = 500$.

The model is set up, creating one computational agent for each of the 4191 students in the data-set, and assigning respectively the individual variables from the survey to each agent. Applied individual variables stem from the literature [87] and are described in greater detail in Appendix A.

As specified in Chapter 4, the agents play a sophisticated version of the Coordination-Game, pointed out in Chapter 2, aiming at maximizing their utility. Focus of the simulation is the behavior "commitment at school", represented by the mark the students achieve in mathematics. Utility is hereby calculated according to 4.4.4, and depends on individual characteristics of the agents as well as peer characteristics and peer behavior. The decision mechanism is initially implemented according to Chapter 4.4 and relies basically on a Learning Classifier System.

Having a set-up model however, does not provide meaningful insights, yet. Foremost, the model needs to be calibrated to real data. In the given use-case this means that the simulation model should be able to reproduce the mark of the students, surveyed after one school year adequately, when starting from the initially surveyed mark. In order to drive the model towards such reasonable output, the input parameters pointed out in Table 5.1 need to be set to proper values.

TABLE 5.1: **Explanation of Model Parameters**

Strength Calculation	α	controls the importance of past performance for the selection of a rule $R_i \in M_i$
	β	controls the importance of past performance for the selection of a rule $R_i \in M_i$
	γ	controls the importance of specificity of rules in the LCS
Genetic Operators	$mutation - rate$	controls how frequently rules in the LCS are replaced by randomly created rules
	$death - rate$	controls which share of the population of rules within the LCS is re -placed by newly created rules (cross-over recombination)
	$evolution - time$	controls how often an evolutionary process is triggered for all agents
LCS	$nr - action - rules$	controls how many condition-action-rules an agent possesses
	$VAR(x)_1$, $VAR(x)_2$, $VAR(x)_3$	Variance of the normal distributions in the generation and mutation of action rules. Control how different from the initial real mark (or respectively the current state of the simulation) the action of a newly created or mutated rule may be.
Utility Function	δ	Imitation Factor, controls the weight of peer behavior within the utility function
	σ	Parameter vector, assigns weights to the individual variables of each agent
	ϕ	Parameter vector, assigns weights to the individual variables of peers
Utility Function	δ	Imitation Factor, controls the weight of peer behavior within the utility function
	σ	Parameter Vector, assigns weights to the individual variables of each agent
	ϕ	Parameter Vector, assigns weights to the individual variables of peers

The process of finding those proper values for the presented input parameters is described by the terms "calibration" or "estimation". Thus, the following subsections are concerned with the question, how such a complex large scale model may be calibrated or estimated.

5.2 Background - Calibration and Estimation

When it comes to the selection of the right parameter values for simulation models, approaches may be divided into two broad categories: "Calibration" and "Estimation".

Hereby "Calibration" incorporates all attempts that employ external or expert knowledge, theoretical concept or empirical findings to define the model parameters. In contrast, "Estimation" is used when applying formal procedures [99], [100]. "Calibration" might be a promising approach when the model consists of few, clearly defined parameters with clear relation to reality.

However, when models become more complex and parameters may not have a clear relation to concepts observable in reality, "Calibration" becomes impossible and approaches for "Estimation" are required. Nevertheless, complex models with a large number of agents represent a challenge for both, "Calibration" and "Estimation" approaches.

Another intuitive approach for identification of parameters of Agent-based models (ABM) is an experimental one, comparing descriptive output with real data. However, this requires a significant number of simulations and has no promise for success. Especially in the case of multiple input parameters that arises frequently when working with ABM, this experimental approach meets its limits.

As long as the considered simulation model is analytically traceable, estimation can be successful, applying standard estimation approaches such as maximum likelihood estimation. However, due to their very nature as an alternative to analytical approaches, ABM are frequently analytically untraceable. Thus, they cannot be calibrated to real data using standard likelihood maximization approaches. Moreover, the calibration of ABM often appears to be non-convex and dynamic. This caused the issue of parameter setting for ABM to become increasingly central in ABM research during the last years.

A large part of the applied solutions may be subsumed under the term "Simulated Minimum Distance" (SMD). Hereby functions are applied to represent the real data and the simulated data. Afterwards, the distance between the function outcomes is minimized. The method of simulated moments [101] and Indirect Inference techniques [102] may be named as examples.

The latter is a common approach from econometrics in the case of intractability of the likelihood function. Here, an auxiliary parameter is linked via an invertible function to the original model via simulation data. Afterwards, the function estimated as link between simulation data and auxiliary parameter is applied to real observations in order to obtain an estimate of the auxiliary parameter for the real data. Inverting of the auxiliary function leads then to an optimal choice for the model input parameter. A first approach to employ Indirect Inference to ABM has been adopted by Gilli and Winker [103] and Winker [104]. Other ABM researchers also applied and recommended this method for the calibration of ABM [105], [106].

Subsequently, Campaglia came up with the use of Gaussian Mixed Models [107] to estimate the auxiliary parameters for the Indirect Inference method and Gaussian Processes to establish the invertible auxiliary function that links the auxiliary parameter to the real input parameters [106]. Within this work, he proposes to approximate Gaussian Mixture Models to the probability distribution of output data, generated via a set of simulations with the ABM to be calibrated. Subsequently a Gaussian Process approximation is used to fit an invertible function to the parameters of the Gaussian Mixed Models and the input parameters used in the simulation. Fitting a Gaussian Mixed Model to the real observation's probability distribution delivers then the estimated auxiliary parameters for the real observations. Subsequently, inverting of the auxiliary function that emerged from the Gaussian Process approximation provides good input parameters for the ABM.

Another approach that may be subsumed under the term SMD is heuristic [108]. Here as with the approaches presented above, the parameter setting for the ABM is conducted by defining an objective function or fitness function as distance between a representation of simulated and real data. The objective function is then minimized using heuristic approaches such as Genetic Algorithms (GA) [109], Particle Swarm Optimization (PSO) [110] or Simulated Annealing (SA) [111].

The advantage of those techniques is that especially for computationally expensive models the number of simulations needed for parameter estimation can be reduced and the estimation process therefore be accelerated. Heuristic approaches are probably better suited to estimate the input parameters of such Agent-based models because they generally present non-linear relations between input and output parameters [112].

5.3 Calibration: Challenges and Approaches

Since the use case model for this thesis is both: large and therefore computationally expensive and complex in the number of parameters, a mere calibration of parameters is impossible, and a brute force experimental approach was precluded due to limited time and computational resources. Therefore, the estimation of the presented model is approached twofold: First, applying a state of the art Indirect Inference approach using Gaussian Mixed Models in Section 5.4 and second, following the heuristic idea, applying a Genetic Algorithm in Section 5.5. Subsequently, it is explained using an example model, how the model itself may be altered in order to achieve better estimation results.

5.4 Estimation via Indirect Inference

As the ABM that is core of this research also contains a variety of parameters, an experimental parameter estimation approach would very likely not be successful. Hence, an Indirect Inference approach proposed by Ciampaglia [106] is adopted in

order to estimate input parameters that lead to reasonable output. This Subsection presents the steps to be taken and the challenges that occurred at this part of the work. The calibration method is hereby applied to the ABM presented in Section 5.1. For preliminary testing of the approach, the experiments presented below were conducted with a subset of pupils, representing 10% of the schools that participated in the FUNDAJ Survey (see 2.2 and 3.3.1), amounting to a number of 271 pupils. Moreover, for the experimental calibration, the original network presented in 3.1 was used. Gaussian Mixed Model fitting and Gaussian process approximation were conducted using the Sklearn tool for Python [113]. Please note that this calibration procedure is a recursive process. Thus, if calibration of the current model fails, the model itself may be revised.

5.4.1 Latin Hyper Cube Sampling and Simulation of Data

Following [106], a number of n input parameter samples $\theta_j \in \Theta$ are chosen using a "latin hyper cube" sampling method [114] in order to generate a space-filling design. The latin hypercube sampling method maximizes the minimum distance between any two sample points. Using this method, an input sample of $n = 500$ parameter sets was generated. Subsequently, the ABM is used to simulate output data $\omega_j \in \Omega$ of the form $\{x_{11}, x_{21}, ..., x_{nn}\}$, where x_{im} denotes the simulated mark of agent (pupil) i at iteration m. As the ABM is non deterministic, each parameter sample is simulated 10 times and the average output mark after 50 iterations is taken as model output.

5.4.2 Gaussian Mixed Model Approximation for Simulated Output Data

In order to obtain a set of auxiliary model parameters χ that reflect well the differences within Θ, Gaussian Mixed Models are fitted to the probability distribution of Ω. According to Equation 5.1 which represents the density function for a Gaussian Mixed Model with k components, the auxiliary parameters χ incorporate variance σ, mean μ and weighting parameter π for each component of the Gaussian Mixed Model and hence the parameter vector χ has length $3k - 1$ for a model with k components.

$$p(x) = \sum_{j=1}^{k} \pi_j p_j(x, \omega_j) \tag{5.1}$$

Figure 5.2 illustrates the fit of a Gaussian Mixed Model with four components to the output distribution of one exemplary input parameter set θ_j.

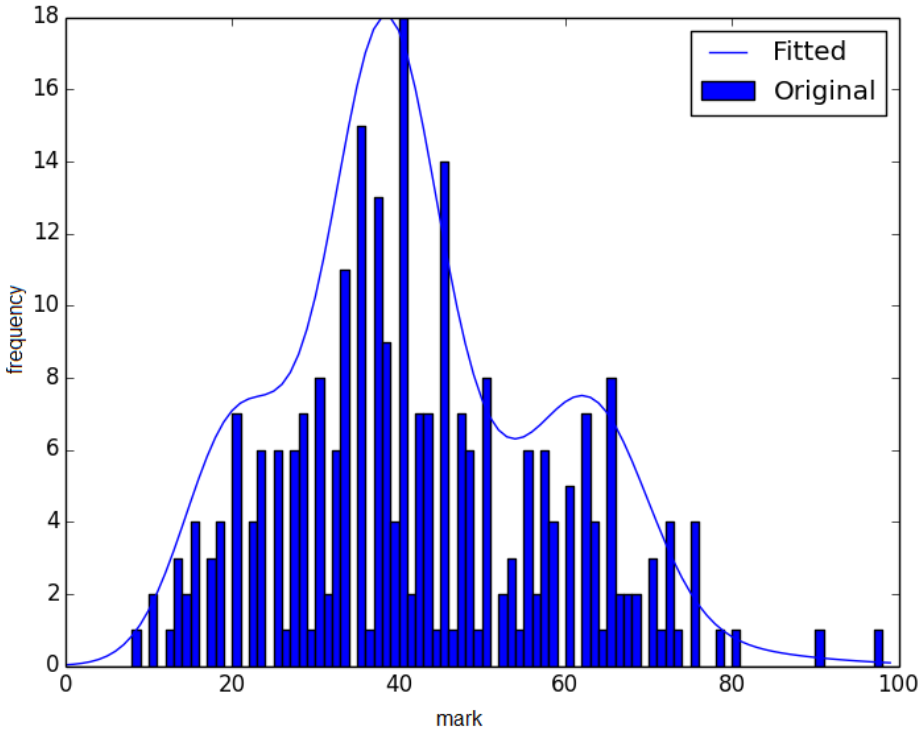


FIGURE 5.2: **GMM Fitted to Probability Distribution of Simulated Marks.** The Figure presents example results for the GMM that was fitted to the simulated marks of a subset of 271 pupils.

The number of components of the Mixed Gaussian Model is a sensible choice here, since the aim is to keep the fitted model unique for each θ_j. This aim may be compromised choosing either too few or too many components. Here the number of components is kept between 2 and 10. Moreover, a sensitivity analysis is to be conducted in order to ensure that auxiliary parameters χ account for variations in Θ.

5.4.3 Gaussian Process Approximation of Auxiliary Parameters and Input Parameters

Aiming at deriving an invertible function $\chi(\Theta)$ a Gaussian Process approximation is conducted using a training set τ containing Θ_τ and χ_τ and a test set Ψ with Θ_Ψ and χ_Ψ where $\tau, \Psi \in N < n$ and $\tau \cap \Psi = \emptyset$. The goodness of fit is hereby measured by the coefficient of determination R^2. The coefficient R^2 is defined as $(1 - \frac{u}{v})$, where u is the regression sum of squares as presented in Equation 5.2 and v is the residual sum of squares computed according to Equation 5.3, where $\overline{\Psi_{true}}$ denotes the mean of Ψ^{true} [113].

$$u = \sum_{j \in \Psi} (\Psi_j^{true} - \Psi_j^{pred})^2 \tag{5.2}$$

$$v = \sum_{j \in \Psi} (\Psi_j^{true} - \overline{\Psi^{true}})^2 \tag{5.3}$$

5.4.4 Gaussian Process Model for Multivariate Fitness Evaluation

The multivariate fitness evaluation contains a total of 47 parameters to be calibrated (see Table 5.1). Where 21 parameters compose each parameter vector σ and ϕ from Equation 4.4 and five parameters account for meta parameters that describe the decision making module of each agent. In this case, the training set could not be predicted reasonably, as R^2 remained negative for all predictions and nearly no variation was predicted for different test parameters Θ_Ψ.

σ and ϕ comprise a subset of parameters. In this case σ and ϕ are components that introduce exogenous heterogeneity to the model by weighting the individual input parameters describing the individual agents. Together with the meta parameters that describe the decision making module of each agent, in the original set-up a total of 47 parameters need to be estimated, turning the parameter estimation into a high dimensional optimization problem.

A further reflection of the Gaussian Process environment reveals, that in a situation with 47 dimensions, a training set τ of maximum length 499, which is the number of synthetic data points obtained via simulations minus at least one data point for the training set Ψ, is very few training data for such a high dimensional estimation problem. Hence, a larger number of simulations is to be conducted in order to obtain a larger training set τ.

5.4.5 Inverting of Auxiliary Function

Having obtained a reasonable function $\chi(\Theta)$, the Gaussian Mixed Model is fitted to the original data obtaining the auxiliary parameter vector $\chi_{original}$. In the use case of this work -the simulation of the emergence of marks of pupils from public schools in Recife- original data is the mark the pupils obtained after one year of study. $\chi_{original}$ is then used to solve the inverse of the obtained Gaussian Process Model $\chi(\Theta)$ in order to generate estimates for the input parameters of the ABM.

5.4.6 Challenges and Limitations

As pointed out before, the GMM approach requires an increasingly large set of simulated test data with increasing complexity of the model. In other words: The more parameters to estimate, the larger- and computationally costlier- the required test set becomes. This is why it seems reasonable, to evaluate the likelihood of success of the applied method before generating the large set of test data. In the underlying case, the complexity stems from the multivariate fitness evaluation within the ABM. Hence, in order to spare time and resources, the GMM method presented above has been applied to a reduced version of the ABM.

For this purpose, a simple heuristic fitness function $U(i,x)$ for agent i and behavior level x was created as presented in Equations 5.4- 5.7. Here, $U_I(i,x)$ denotes the utility, agent i receives from imitating the behavior of her peers, $C(i,x)$ denotes the effort, agent i has to make in order to achieve the respective level of behavior x and $U_P(i,x)$ represents the (private) utility the behavior itself generates for agent i. α, β, and γ are weighting parameters.

$$U(i,x) = \alpha U_I(i,x) - \beta C(i,x) + \gamma U_P(i,x) \tag{5.4}$$

The imitation utility is calculated according to Equation 5.5. Where N_i denotes the social neighborhood of agent i containing n agents, x denotes the behavior level (achieved mark) of agent i and x_j the respective behavior level of the peer of agent i, agent j.

$$U_I(i,x) = e^{\frac{\sum_{j=1,j\epsilon N_i}^{n} x - x_j}{n}} \tag{5.5}$$

Equation 5.6 explains how the costs of a certain behavior level $C(i,x)$ are computed. Δx denotes hereby the alteration of x since the previous iteration, η and δ are parameters.

$$C(i,x)) = \Delta x e^{\eta + \frac{x}{\delta}} \tag{5.6}$$

Private utility, derived from the behavior itself is represented according to Equation 5.7.

$$U_P(i,x)) = \Delta x e^{-\frac{x}{\delta}} \tag{5.7}$$

Even though this function is heuristic, it is set-up to be incentive compatible, creating marginally decreasing $U_P(i,x)$ and marginally increasing $C(i,x)$ for increasing x. This means the additional utility, derived from a better mark grows more slowly, the higher the original mark is, while any improvement comes with a higher cost for increasing marks. Lifting a students mark from a very low value like 5 to a medium value, e.g 50 derives therewith a higher additional utility to the agent than achieving a mark of 100 from originally 95. The contrary relation holds for the cost of improvement.

5.4.7 Gaussian Process Model for Simple Fitness Evaluation

The Gaussian Process approximation for the simple evaluation procedure, containing only three parameters α, β, γ from Equation 5.4 leads to reasonable predictions of the training set, where the coefficient of determination R^2 alternates between 0.3 and 0.5.

Figure 5.3 illustrates the prediction of one auxiliary parameter μ_j of the GMM by the Gaussian Process Model as a function of one input parameter. Here the predicted

values are represented by red crosses, while original values are represented by the blue solid line.

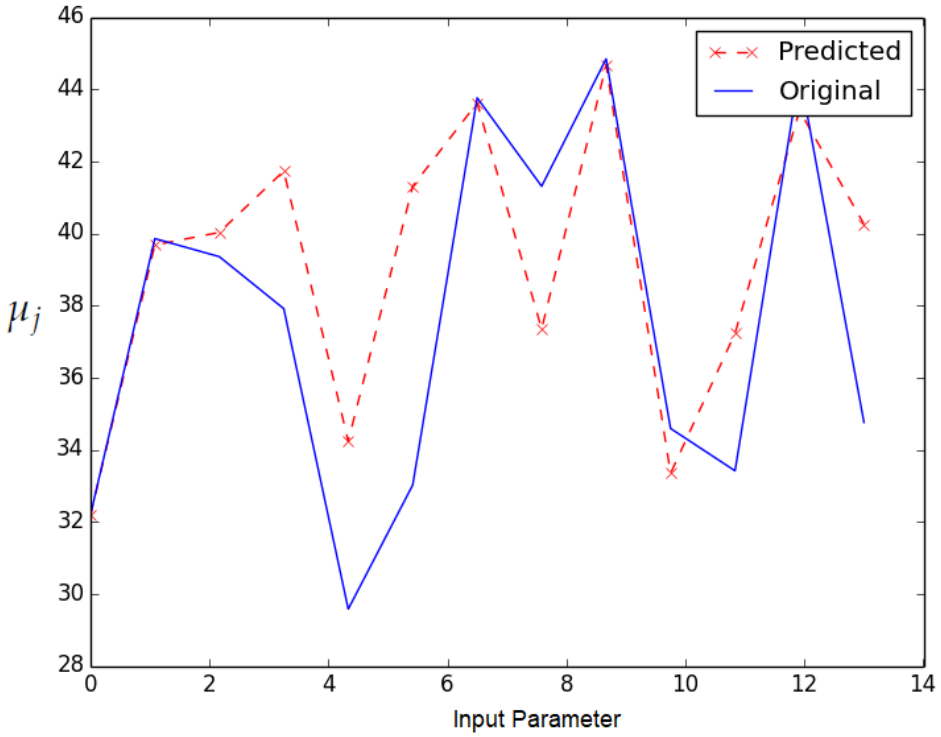


FIGURE 5.3: **Gaussian Process Prediction of Auxiliary Parameter.** The Figure illustrates the prediction based upon test data points derived from the reduced model with simple fitness evaluation. For the reduced model, the GP prediction seems reasonable.

The obtained parameters have been subsequently applied to the reduced ABM. Unfortunately, it turned out that the outcomes of the ABM could not be reproduced with the parameter set indicated by the Indirect Inference method. Instead, model behavior appeared to be quite volatile. This points out another problem of the application of estimation methods: The generation of the test data is generally strongly correlated with model behavior. When the model to estimate exhibits dynamic and unpredictable behavior over time, it is a big challenge to identify a suitable period for the test data.

The difficulties that have arisen with the presented estimation approach led to the decision to apply an alternative method that better coped with dynamic model behavior.

5.5 Heuristic Approach to Calibration

The advantage of heuristic estimation methods is that their success can be easily observed during the run-time and the test data must not be generated a priori, but

is being selected along the estimation process. This features enable the researcher to save time and computational resources. Also, Agent-based models such as the case study presented in this work feature generally a non-linear relation between input and output parameters [112]. This non-linearity is also imminent in heuristic approaches.

To set-up a heuristic estimation method, a proper understanding of the problem structure is needed. As stated above, the given simulation model comprises an extensive set of parameters to be calibrated. Table 5.1 presents the set of parameters, where σ and ϕ comprise a subset of parameters that introduce exogenous heterogeneity to the model. Together with the meta parameters that describe the decision making module of each agent, in the original set up a total of 47 parameters need to be estimated.

In such complex parameter estimation problems, literature frequently recommends Genetic Algorithms [88] as an estimation approach, since they can be classified as moderate approach in regard to the degree to which they make assumptions about the problem itself [115]. Following this classification, weak approaches would make very few assumptions about the problem and thus may require large computational resources. As a weak approach to the presented estimation problem one may think of a brute-force simulation approach. Strong approaches on the other hand may be too specific for large scale complex problems and tend to find wrong solutions when the assumptions made about the underlying problems do not fit well.

The strong estimation approach of Indirect Inference from Section 5.4 presented precisely the shortcomings of over specification and the scale and complexity of the model do not permit a weak approach due to restrictions in available time and computational resources. Hence, as a moderate approach a heuristic solution to the parameter estimation problem was applied. Among others, literature recommends Genetic Algorithms (GA) [109], Particle Swarm Optimization (PSO) [110] or Simulated Annealing (SA) [111] for this purpose.

The GA hereby imitates natural evolution, selecting, recombining and mutating a random set of initial solutions to the calibration problem.
The PSO approach is inspired by the searching behavior of natural swarms such as bees or birds. Hereby, the parameter search is initialized with a number of random solutions, the particles, where each particle searches for the optimal solution to the problem in the parameter space close to itself. Every particle searches iteratively around the individual maximum and is in parallel attracted to the best solution found by the entire swarm.

SA imitates the process of annealing, where materials are heated fast and cooled slowly afterwards in order to create harder materials. Translated to optimization, the algorithm defines a certain level of fitness of a random state of an ABM and subsequently disturbs this state and evaluates the derived fitness. If the fitness of the

new state of the model outperforms the current fitness, it is taken as a new starting point.

Since the encountered parameter search problem presents considerable complexity and is computationally expensive, an efficient optimization approach should be chosen. As the PSO tends to require higher computational costs due to necessary communication between particles [116] and the SA can be expected to need a higher number of simulations due to its undirected nature, they seem less adequate for the described purpose than the Genetic Algorithm. Also, the estimation problem incorporates a clear way to evaluate fitness (Residual Square Sum of Errors) and an extensive number of on continuous- or discrete scales measured input parameters to be calibrated. Thus, the problem set up is also well suited for the application of a Genetic Algorithm.

5.5.1 Description of the Genetic Algorithm

Figure 5.4 illustrates the setup of the basic Genetic Algorithm that is applied to the estimation problem as follows.

In a first step, the estimation process is initialized, hereby a set P of n candidate solutions p is created, where each solution represents a possible solution to the parameter estimation problem. The solutions are hereby represented as strings, where each digit stands for one parameter of the model. Meaningful ranges are given for all parameters as indicated in table 5.2.

TABLE 5.2: **Range of Model Parameters**

Model modules	Parameters	Range
Strength Calculation	α	[-1,1]
	β	[-1,1]
	γ	[-1,1]
Genetic Operators	$mutation - rate$	[0,1]
	$death - rate$	[0,1]
	$evolution - time$	[0,60]
LCS	$nr - action - rules$	[0,50]
	$VAR(x)_1, VAR(x)_2, VAR(x)_3$	24
Utility Function	δ	[-1,1]
	σ	$\forall v \in \sigma \rightarrow v \in [-1,1]$
	ϕ	$\forall y \in \phi \rightarrow y \in [-1,1]$

Subsequently, the Agent-based simulation model is setup for each candidate solution with the respective parameters and the model is executed for the number of iterations given in the respective candidate solution. Afterward, fitness is calculated for each candidate solution according to the fitness function of the Genetic Algorithm. A selection mechanism follows the initialization and fitness evaluation. Here

the share q of worst performing solutions are discarded before a crossover mechanism is applied, recombining the surviving solutions and creating a new breed of candidate solutions.

Finally, the new population of solutions undergoes a mutation process with a given probability m, ensuring a certain level of diversity in the population.

This process is repeated as long as the best fitness in the population does not converge. Convergence is hereby assumed when the fitness level of the population does not increase for a predefined number of iterations and denoted as $max - stagnation$.

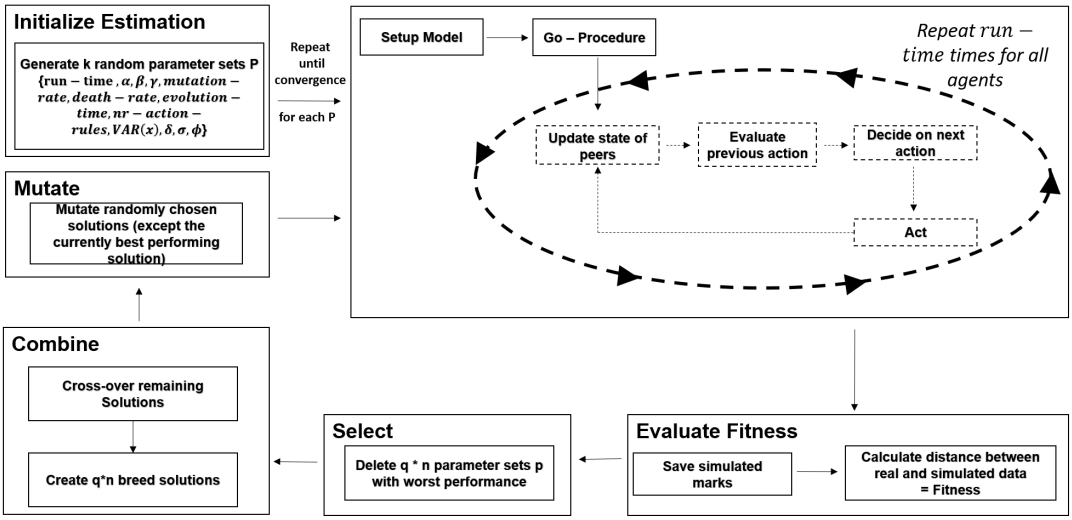


FIGURE 5.4: **Heuristic Estimation via Genetic Algorithm.** The Figure presents a schematic illustration of the GA approach for parameter estimation.

The crossover-mechanism and fitness evaluation are illustrated in greater detail in Figures 5.5 and 5.6. As indicated in Figure 5.5, the fitness of a set of parameters p is calculated as the Residual Square Sum of Errors (RSS) of the simulated marks x of all k agents in the model under the given parameter set p and the observed mark of all students in the real data-set $\tilde{x}$.

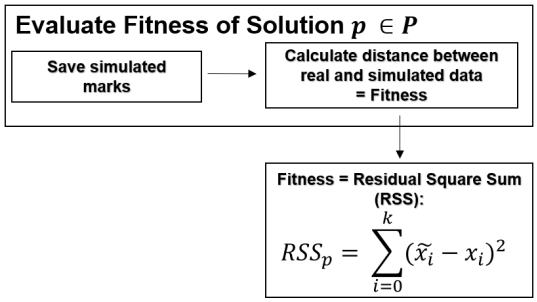


FIGURE 5.5: **Heuristic Estimation via Genetic Algorithm: Fitness Evaluation.** The Figure illustrates the fitness evaluation procedure of the GA approach for parameter estimation.

To create a new breed of possible solutions to the estimation problem, a recombination mechanism is employed according to Figure 5.6. After selecting the surviving

solutions with the highest fitness of the current population of solutions, recombination starts, ordering the survivors by their fitness value. Then, parent solutions P1 and P2 are extracted from the survivor list. Here P1 is the best performing solution of this list and P2 is any other solution. Each pair of parents breeds two child solutions, where the leading parent solution is switched. The child solutions are created as a weighted average of each digit of the parent solution strings. The weight factor is predefined for each calibration attempt and set to 0.7 for the presented experiments. Herewith, one child solution is predominantly influenced by P1 and the other child solution stems predominantly from P2. When the number of surviving solutions is uneven, the last remaining solution is crossed over again with the initially best performing one. This process is repeated until the original number of candidate solutions n is reached.

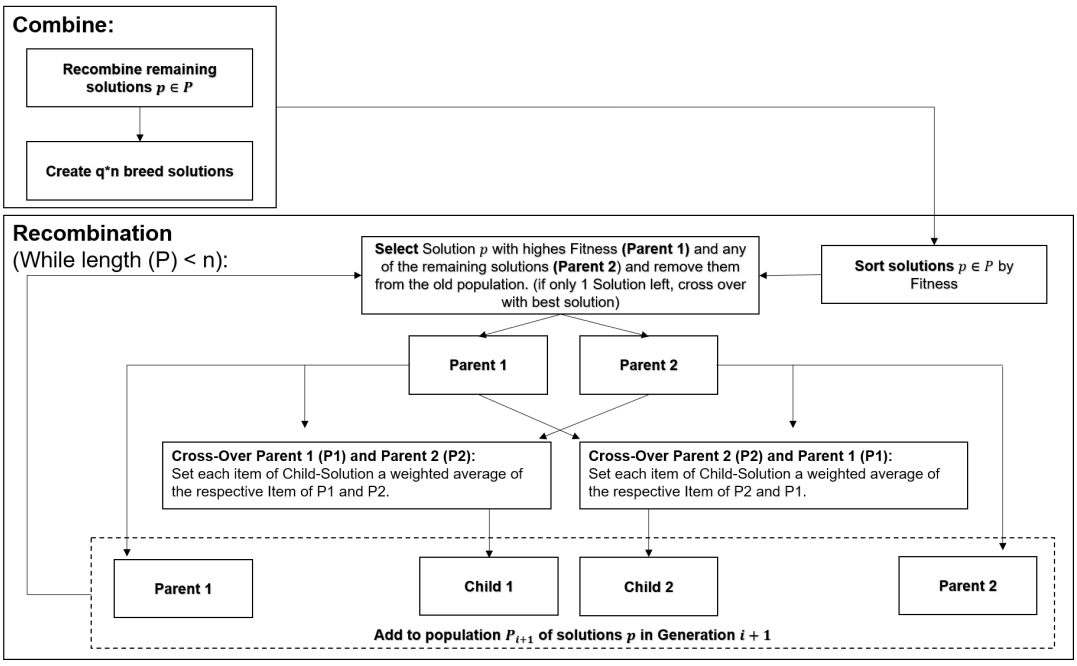


FIGURE 5.6: **Heuristic Estimation via Genetic Algorithm: Recombination.** The Figure illustrates the recombination procedure of the GA approach for parameter estimation.

Subsequently, a mutation process ensures diversity of the set of considered solutions P and herewith allows the algorithm to perform jumps on the search space. This hinders the GA to converge to local maximums. Figure 5.7 illustrates the respective workflow. The process relies twice on the generation of random floating point numbers: First, each candidate solution p but the currently best performing one enters the mutation process with probability $GA - Mutation - Rate\ m$. When chosen for mutation, another random number decides whether the entire solution is recreated from scratch or if single digits of the solution are changed at random. This provides the algorithm with the capability to perform jumps with varying step-size.

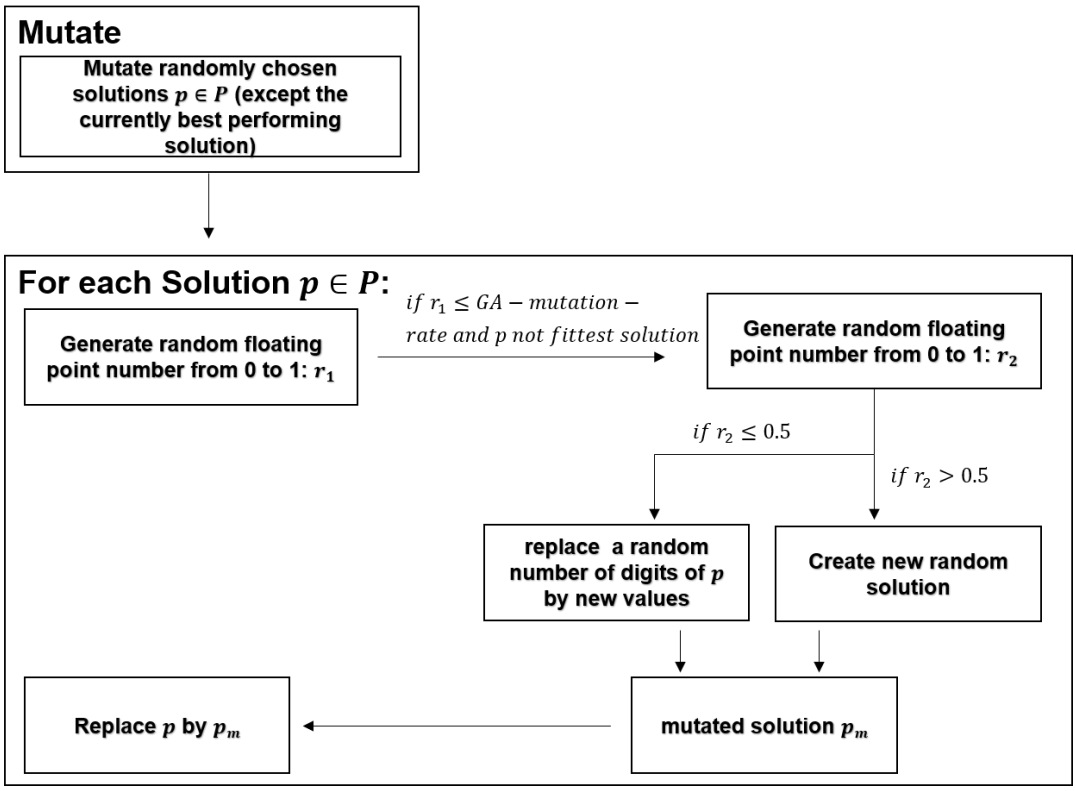


FIGURE 5.7: **Heuristic Estimation via Genetic Algorithm: Mutation Process.** The Figure illustrates the mutation procedure of the GA approach for parameter estimation.

As a resume, table 5.3 presents and describes the adjustable parameters of the applied Genetic Algorithm.

TABLE 5.3: **Adjustable Parameters of the Genetic Algorithm**

GA-Parameter	Description	Characteristic Values
Number-Solutions n	The number of Solutions within the population, $p \in P$	≥ 2
GA-Death-Rate q	Share of P that is discarded per iteration o the GA	$0 \leq q \leq 1$
GA-Mutation-Rate r_1	Probability that a Solution will enter the mutation process per iteration	$0 \leq w \leq 1$
Weight Recombination w	weight factor according to which the weighted average between P_1 and P_2 is calculated	$0 \leq w \leq 1$
Max Stagnation	maximum number of iterations without increase in fitness until the estimation stops	> 0

5.5.2 Step-wise Performance Enhancement

The foregoing subsection described the mechanics of the applied Genetic Algorithm. However, heuristic estimation requires a close understanding of the problem domain and may need to be adapted step-wise to deliver meaningful results. When

applied to the estimation problem, the presented GA setup did initially not provide convincing results. Figure 5.8 indicates the fitness (RSS) of the best performing solution within each generation of the GA. The estimation has been repeated 10 times, where the grey solid lines illustrate the respective fitness for each run per generation. The red solid line indicates the average fitness of all 10 repetitions per generation. To make the performance of the GA accessible, the original state of the model is expressed as a blue solid line as a baseline. Recall that the performance of the model is expressed as Residual Square Sum of Errors (RSS) between the simulated and the observed mark of the students within the data-set. Therefore, the original state of the model is represented as RSS between the observed mark of the students at the beginning of the year and the observed mark of the students at the end of the year. In other words: The blue solid line expresses the performance of the model if it was not run at all.

Hence, it can be said, that the model only delivers meaningful forecasts, if the red solid line crosses the blue solid line in the Figure.

Considering the above, one may conclude that the GA does not provide helpful solutions to the parameter estimation problem: Although demonstrating a visible performance enhancement with increasing number of generations, the GA does not provide parameter setups that equip the simulation model with significant predictive power.

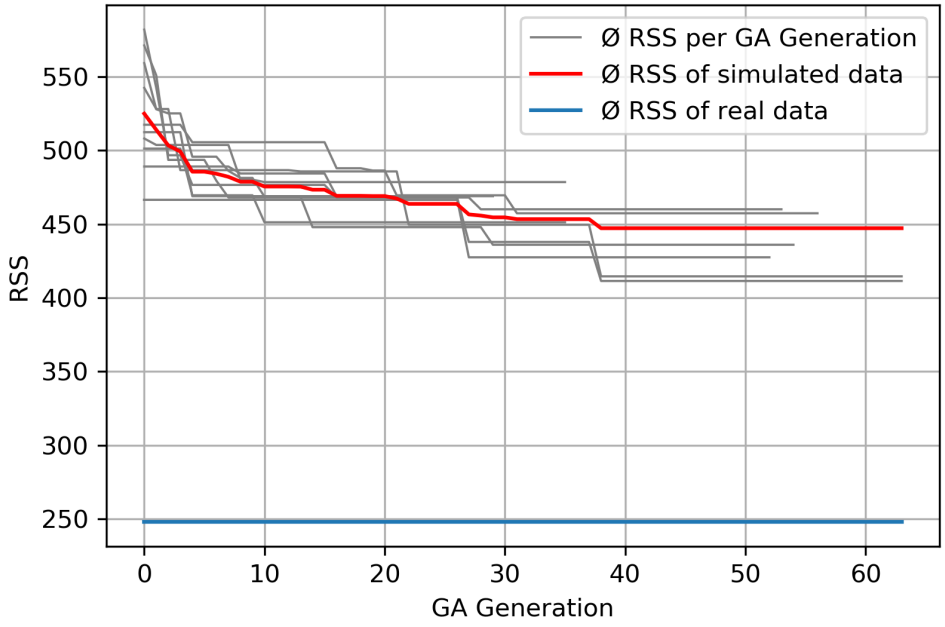


FIGURE 5.8: **Parameter Estimation Experiment 1.** The Figure presents the development of Fitness (RSS) of heuristic estimation via Genetic Algorithm. Global estimation, decision making by Classifier System, original set of parameters. The GA does not provide parameter setups that equip the simulation model with significant predictive power.

5.5.3 Analysis of the Set of Variables

When confronted with a poorly performing heuristic estimation method, the researcher may want to analyse whether the assumptions made about the general framework still hold for the given simulation environment.

The original overall set of variables, characterizing the individual agents, stems from the literature [87]. The originally applied variables are described in greater detail in Appendix A. To verify if the applied set of variables best suites the presented model and data-set, a statistical analysis is conducted. To this purpose, a correlation analysis (Pearson Correlation Coefficient) is run for the relation of the single variables in the data-set with the individual mark in mathematics after one school year. The correlations for each originally applied variable may be revised in Appendix A. As an outcome it is found, that the predictive power of the variables, suggested in the literature, is not sufficient. The originally applied parameter set is not the best suited for this purpose.

The survey that serves as a base for the FUNDAJ data-set [19] applies a methodology similar to the one applied in the AddHealth Study [76] which serves as a base for the reference work of [87]. However, the surveys have been conducted in differing countries and hence differing socioeconomic realities. While AddHealth collects data from adolescents in the United States, FUNDAJ-Study occurred in northeastern Brazil and focused o pupils from public schools that - in Brazil- tend do be frequented by children from poorer backgrounds. Surprisingly for the FUNDAJ data, socioeconomic background variables such as monthly gross income per capita do not have a significant effect on schooling success. Instead, variables that reflect socioeconomic status indirectly (like the question if the household is equipped with a washing machine and a refrigerator, or if the household employs a housekeeper) appear to have a much stronger correlation with the outcome variable. As a consequence, the applied parameter set is decreased and focused on those variables that correlate significantly with the outcome variable. The identified variables including the respective Pearson correlation with the outcome variable are listed in Appendix B.

Figure 5.9 illustrates the performance of the Genetic Algorithm in estimating the parameters for the Agent-based Model after the optimisation of applied parameters. Again, the estimation is repeated 10 times. Here the grey lines indicate the average RSS of all agents in the test set for each run per generation of the Genetic Algorithm. The solid red line indicates the average fitness per generation of all 10 estimation runs and the solid blue line marks the original state of the model, expressed in the average RSS of the surveyed mark of all simulated agents at the beginning of a school year and at the end of a school year.

It can be observed that although the RSS of the simulation is continuously lower than the RSS in Figure 5.8, RSS of simulated data never gets even close to the original RSS of the model. The GA converges to a RSS of approximately 410.

Thus, as a resume it can be pointed out that the alteration of input variables yields better estimation result, yet is not sufficient to equip the model with significant predictive power.

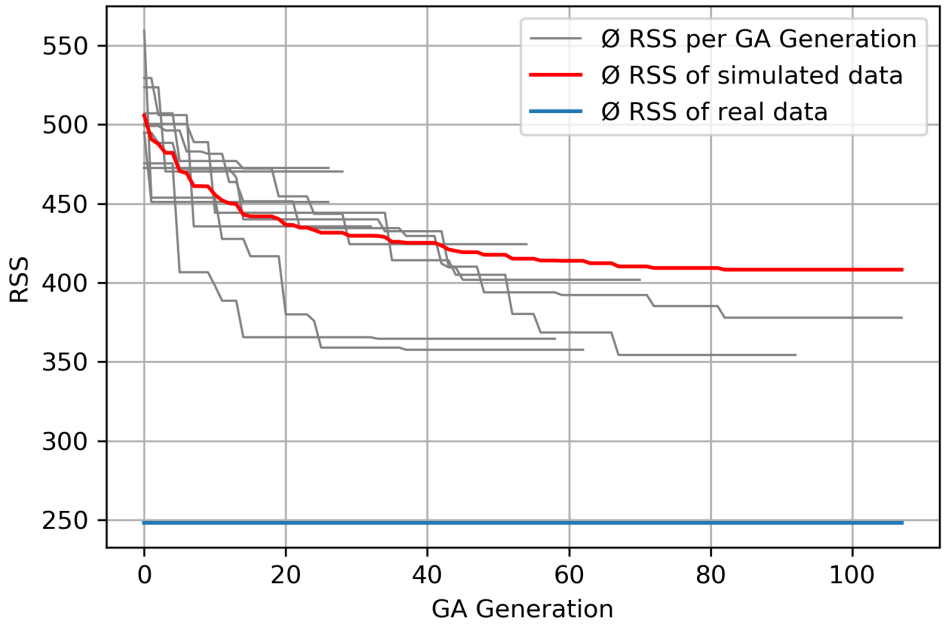


FIGURE 5.9: **Parameter Estimation Experiment 2.** The Figure presents the development of Fitness (RSS) of heuristic estimation via Genetic Algorithm. New set of parameters. Global estimation, decision making by Classifier System. The alteration of input variables yields better estimation results, yet is not sufficient to equip the model with significant predictive power.

5.5.4 Distributed Estimation

One problem that arises among the presented calibration approaches is certainly the complexity of the global agent model consisting of more than 4000 heterogeneous individuals from 122 schools from distinct geographic and socioeconomic backgrounds. Even within the reduced scope of the 10% subset of components, individual non-observable influences can be assumed to be very different among the individuals. According to this, the researcher may consider that a global estimation approach with the aim to find a single set of parameters that drives the simulation model towards the desired outcome cannot be found. When arriving at this conclusion, the global estimation approach should be abandoned and a decentral approach adopted instead.

This means that the parameter estimation no longer aims at finding a single parameter set for the entire model but that the model may be divided into reasonable components according to the network structure. Reasonable component may be though of as clusters within the network of agents that are strongly interconnected

but weakly connected to the rest of the network. Estimation then focuses on the search for component-specific parameter sets. This approach does not contradict the aim of creating a global network model since, after estimation of the component-specific parameter sets, the global simulation of the interconnected network is still possible. In this example, the calibration procedure is modified as follows:

The Genetic Algorithm (GA) presented in 5.4 is applied sequentially for each component of the model that contains more than one interconnected agent. The error is hereby defined as the difference between the simulated mark of a student and the observed mark after one year from the data-set, expressed as the Residual Square Sum of errors (RSS) per component. This approach allows for better fitting of the model to component-specific non-observable influences and can be found in the literature for similar research [87]. Figure 5.10 presents how the estimation process evolves with an increasing number of generations of the Genetic Algorithm. The RSS is respectively given as the average RSS of all agents within the model in each iteration of the GA. The estimation has been repeated five times. Here the grey solid lines represent the average RSS of all components of the model per run and generation of the GA. The red solid line indicates then the average performance of the GA for all five repetitions.

It is striking that the estimation success not only improves when shifting to decentral estimation, but even significantly outperforms the original state of the model. In other words: when applying decentral estimation, it is possible to find parameter sets that enable meaningful simulations with significant predictive power.

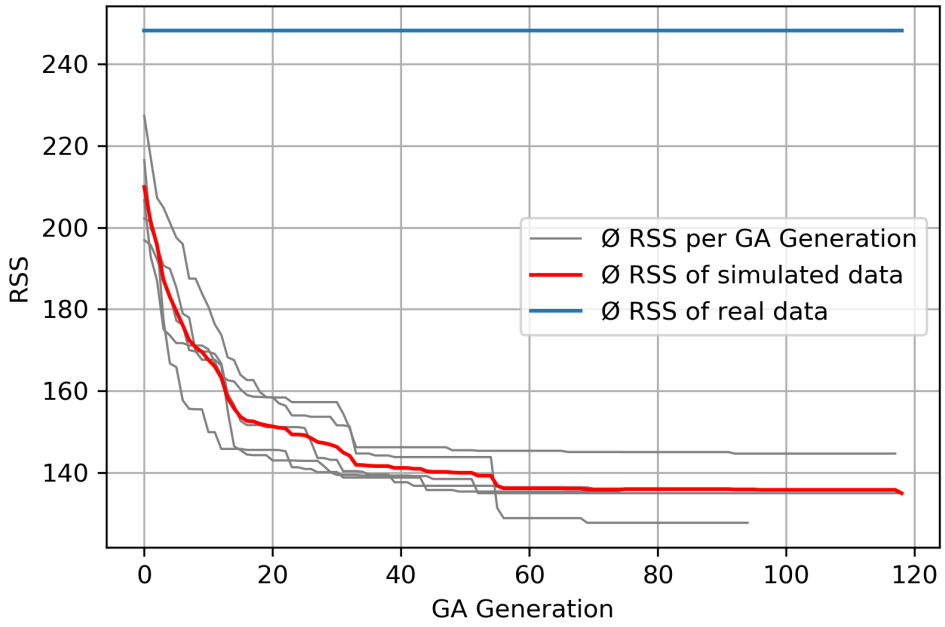


FIGURE 5.10: **Parameter Estimation Experiment 3.** The Figure presents the development of fitness (RSS) of heuristic estimation via Genetic Algorithm. New set of parameters. Decentral estimation, decision making by Classifier System. When applying decentral estimation, it is possible to find parameter sets that enable meaningful simulations with significant predictive power.

5.5.5 Considering Simplifications

As indicated in Figure 4.1 the original model architecture stipulates a decision-making module based on a Learning Classifier System (LCS). The architecture hereby aims at closely imitating human decision making under the given conditions. This decision-making approach is expected to be imperfect and guided by past experiences.

However, as specified in the above, the classifier-based decision mechanism implemented in a dynamic environment may cause the simulation model to be too volatile. This can be explained with the struggling of the LCS to cope with the dynamic environment. Moreover, the additional stochastic elements and the additional parameters of the learning approach negatively affect traceability of the model. This may motivate the attempt to simplyfy the model and assess if a comparable performance can be achieved.

Traceability and determinism of a simulation model can be enhanced with the elimination of stochastic elements within the decision process. Hence, to overcome the difficulties stated above the decision process may be replaced by a brute-force optimization method. To this purpose in the exemplary model setup from 5.1, the optimization method presented in Figure 5.11, replaces the LCS Decision Module of the model, referred to in Figure 4.1. Here, the current environment of the agents is assessed and subsequently each agent evaluates all possible actions and chooses the

one that yields maximum utility. This mechanism does not contain any probabilistic elements and hence drives the whole model towards a deterministic behavior.

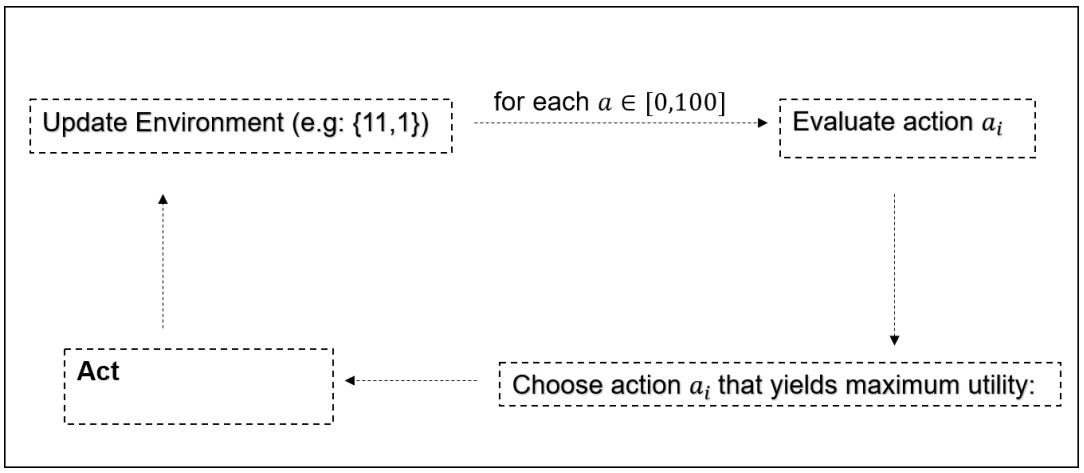


FIGURE 5.11: **Decision Module with Simplified Decision Mechanism.** The Figure illustrates the brute force approach for agent utility maximization.

Estimating the accordingly simplified model with the Genetic Algorithm presented above yields the results illustrated in Figure 5.12. Again, the parameter estimation was conducted five times and the GA performance for each run is indicated by the grey solid lines, while the average performance over all five runs can be assesses by the red solid line. The blue solid line represents the base-line respectively the original state of the model. It can be observed that the results are far less striking than under the more probabilistic approach in the foregoing paragraph 5.5.4. In fact, the model lacks predictive power when deleting the probabilistic elements inherent in the bounded rational decision making module. However, a positive development can be observed.

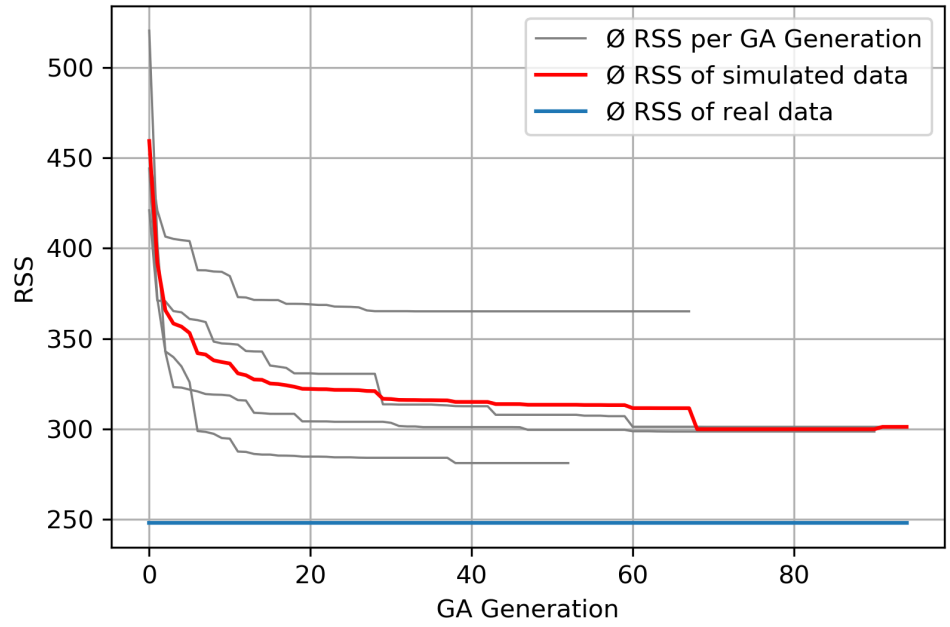


FIGURE 5.12: **Parameter Estimation Experiment 4.** The Figure presents the development of fitness (RSS) of heuristic estimation via Genetic Algorithm. Decentral estimation, decision making by brute-force utility maximization. It can be observed that the results are far less striking than under the more probabilistic approach in the foregoing paragraph.

5.6 Conclusion - Large Scale Model Calibration

This Chapter addresses the challenge to find proper input parameters for complex and large-scale Agent-based models. using the case-study of the exemplary model for simulating behavior of a large-scale student population in the Brazilian city Recife, it is shown, how this task can be approached and which difficulties may arise.

It is demonstrated that formal estimation approaches like the state of the art approach of Indirect Inference are applicable but require costly preliminary simulations and do not guarantee estimation success. Especially for large-scale models, the researcher shall therefore meticulously evaluate if such an approach promises meaningful outcomes. This holds especially if the model under consideration employs considerable probabilistic elements and volatile emergence over time.

As an alternative, heuristic procedures may be considered since heuristics provide the opportunity to observe the performance during the run-time, do not require foregoing, potentially costly simulations and mirror the non-linear relation between input parameters and model output inherent to Agent-based models.

By means of a Genetic Algorithm approach it is further shown that heuristic estimation requires close observation of the process and individual adaption to the respective simulation model. Hints are given on how to improve the performance

of the heuristic estimation method. It is especially found that for large network simulations a distributed approach to parameter estimation may be best suited.

Applying this distributed estimation approach, the Genetic Algorithm is capable of estimating input parameters that reproduce the real-world data from the applied data-set very closely. Thus it was demonstrated that the proposed Agent-based model for the simulation of complex contagions can be calibrated to real data.

Probabilistic elements continue to be the boon and bane of Agent-based models. On the one hand, those probabilistic elements provide the model with endogenous heterogeneity and enable the unpredictable macro-phenomena to emerge. On the other hand, they lead to volatile model behavior.

However, models that present volatile behavior are more difficult to interpret, calibrate and estimate. The behavior of very volatile models may only be reasonable analysed on average, which on the other hand means that the number of required simulations grows significantly. This is a difficulty especially when it comes to estimation of input parameters. Thus, as an approach to overcome this difficulties and verify the suitability of the chosen Agent-based approach, it is introduced the idea to simplify the model, reducing probabilistic elements and hence driving it towards a deterministic behavior.

Here, it falls to the researcher to decide whether the results are still realistic. In the underlying example it becomes clear that a complexity reduction reduces significantly the predictive power of the model. The fact that the complex ABM, incorporating bounded rational agents can be well calibrated to real data, while this attempt fails for the deterministic model that consists of fully rational agents, is a very strong argument to further pursue the initially presented complex approach.

Chapter 6

Conclusion

Starting with the aim of supporting policy making and particularly the design and evaluation of Conditional Cash Transfer (CCT)-Programs, this research aims at developing a modeling approach for complex contagion processes suited to simulate the complexity of the studied societies that is flexible to handle by policy makers.

Shortcomings of current DSGE models are overcome by the use of techniques from Agent-based Modeling and Network Science. Overall, the research objective was to develop a coherent methodology for the set-up of a simulation model, incorporating complex network spreading processes. The model enables the policy makers to conduct useful what-if-simulations when considering alterations of CCTs or related policies. A focus was hereby set on the spread of commitment to school education among adolescents.

This Chapter concludes the thesis, transforming the results above into a guideline for generating network simulations models. Further it points out valuable future research and applications of our work. These applications include what-if-simulations, addressing the question of restructuring of social networks and the testing of public policies. Particularly, applications concerning public policies for reduction of poverty and generation of human capital (like Conditional Cash Transfer Programs) were considered.

Our theoretical contributions are completed by a real-world example. This case-study stems from survey data from adolescents from the northeastern Brazilian city, Recife. A model for the simulation of diffusion patterns of the attitude towards education among those adolescents is developed an estimated to fit real world data.

6.1 Recapitulation of the Research Questions

As indicated above, this research had the overall objective of developing a coherent methodology for the set-up of a simulation model for complex contagion processes. Further, this methodology has been applied to the data-set surveyed within the study *"Determinantes do desempenho escolar na rede de ensino fundamental do Recife"* [19]. Eventually, policy makers will be enabled to conduct useful what-if-simulations concerning public policies in the areas of education and/or Conditional Cash Transfer-Programs and social welfare.

In order to achieve this overall objective, four major subjacent questions needed to be addressed.

Initially, a basic principle for the pattern that underlies the complex contagion process in question had to be found. Secondly, available data needed to be extrapolated to enable holistic network simulations. Further, a suitable decision mechanism for the individual agents was required and reasonable input parameters for the large scale model needed to be estimated. The following paragraphs explain the individual research goals in greater detail.

Modeling Complex Contagion of Behavior: First it was intended to identify and select the underlying spreading framework for the complex diffusion of behavior within the simulation model. Specific detail was given to the examination of the Coordination-Game mechanism [15] and its applicability for the underlying case. The Coordination-Game mechanism was chosen as a pattern since it resembles a simple but intuitive way of representing imitation among peers in networks and hence suits well when the aim is to verify if imitation plays a major role in a network spreading process. Therefore, the mechanism was to be tested on different data-sets and particularly the suitability for the practical example of contagion of educational attitudes among adolescents had to be examined.

Missing Data in Social Network Data-Sets: Secondly, the research had the objective to generate answers to the problem of missing data in network data-sets. Since the intended simulation model relies on a large interconnected network, it is essential for the successful implementation that the connections between the modeled individuals are represented well. (Social-) Network data usually stems either from surveys or from appropriate online sources such as online social network platforms. Both data sources suffer from missing data, while survey data is generally affected by the "Fixed Choice" effect and both sources usually present missing data due to the "Boundary Specification Problem".

The "Fixed Choice Effect" occurs when survey participants are asked to nominate up to a certain number of peers. When the number of possibly nominated peers is restricted, the network data is supposedly incomplete. "Boundary Specification" disturbs network data when links to peers are omitted because the survey is restricted to a certain location or to members of a certain entity such as pupils from a single classroom or employees from a certain company. The "Boundary Specification Problem" also occurs in online networks, for example due to age restrictions or the restriction of access to the platform to certain groups of people, or certain locations. This thesis therefore provides solutions for those missing data problems.

Decision Making: Moreover, as stated above, this approach stemmed from Agent-based modeling and network science and aimed at overcoming the restriction of

current DSGE-Models to the rational individual. To this purpose the thesis proposed an implementation of a Learning Classifier System (LCS)-based decision module for the agents within the simulation model on the basis of the extrapolated network data. This approach provides the agents with bounded rationality [10] and hence enables more realistic simulations. Here, the aim of the research was to examine whether the LCS is a good representation for the decision making of computational agents within the proposed framework. It was to be verified if the computational agents can therewith be equipped with bounded rationality and if their behavior subsequently mimics human behavior sufficiently.

Parameter Estimation: Finally, the research proposed a solution for the challenge of finding the right parameters for such a large-scale complex network simulation. The applicability of state-of the art inference approaches was to be examined and alternative heuristic approaches needed to be proposed, implemented and tested. Hereby a special focus was on which steps may be taken to improve heuristic estimation of large-scale Agent-based models. Due to the nature of the estimation problem that incorporated a clear way to evaluate fitness (Residual Square Sum of Errors) and an extensive number of input parameters to be calibrated, a Genetic Algorithm was chosen as an exemplary approach to parameter estimation.

6.2 Results

The research questions pointed out in the previous Section were addressed within this thesis. An overall methodology for a simulation model incorporating complex contagion on social networks was proposed and implemented using the example of adolescent pupils from public schools in Recife. Moreover, a methodology for the parameter estimation was developed and successfully implemented and the overall methodology was verified against the performance of a base-line model, incorporating fully rational agents. The results for the questions that were brought up above are described in the following paragraphs.

Modelling Complex Contagion of Behavior: The adoption of a Coordination-Game mechanism for simulating the spreading process of behavior throughout social networks was presented, accuracy was evaluated by several quality indicators. Strong indications were found that a Coordination-Game mechanism underlies the spread of the behaviors “commitment to school education” as well as “smoking” and “drug-use”. In contrast to similar studies from the literature [39] [40], comparable evidence for behavior “alcohol-use” could not be found in the applied data-sets. The results for behavior “practicing sports” were not clear.

Evidence was provided that there is an underlying game-environment for the agents within the social systems that can be modeled as a Coordination-Game. Therewith, it could be demonstrated that imitation plays a major role for the contagion

processes of interest. However the results also indicate that for a more accurate simulation, the players of this game, the bounded rational agents needed to be equipped with decision finding mechanisms that better approximate human decision making.

Missing Data in Social Network Data-Sets: Subsequently the suitability of three graph generation or respectively link-prediction techniques was evaluated in order to impute not at random missing data about social ties within social network studies, and to interconnect disconnected components within the surveyed network from the study *"Determinantes do desempenho escolar na rede de ensino fundamental do Recife"* [19]. Hereby the target was to create a comprehensive model of the global network, stemming from original network data.

Firstly the Social Circles approach as proposed by Gilbert was modified such that it was suited to impute missing links between the locally isolated components of the original graph (**Approach 1**). Secondly, a bootstrapping approach proposed by Leroy et.al (**Approach 2**) was applied and thirdly a hybrid algorithm was created, combining features from the former and the latter (**Approach 3**).

It was argued that the Social Circles **Approach 1** should be applied if the social scientist is willing to generate interconnected networks from unconnected components and acts upon the assumption that the macro network shall closely resemble micro structures. In other words, the approach put forward was able to predict links comprehensively if the assumption holds that missing data stems from “Boundary specification" but that the “Fixed choice effect" does not affect the data collection.

An example for this could be the examination of best friends links between school children, where it can be assumed that most links are already established within the classroom (where the survey has been conducted) and very few external links are missing in the survey-data.

However, if the researcher aims at creating interconnected networks, and does not expect that the network structure that can be observed in the original data approximates the macro network structure, the combined approach (**Approach 3**) seems to be even more adequate. In this case not only “Boundary specification" but also the “Fixed choice effect" causes missing data.

Our results indicated that the bootstrapping approach (**Approach 2**) enables the scientist to reproduce close online relations. However, more than the other approaches, this one required additional information about the individuals and hence suffered from data granularity issues. Also, the combined approach may solely be suited for data-sets that yield some information about the individuals that allow for calculating a social distance between them.

For the practical example that is developed alongside the theoretical contributions, the "Social Circles" **Approach 1** was chosen as extrapolation technique.

Decision Making: Further, the implementation of a Learning Classifier System (LCS) was proposed as a decision making module for Agent-based models that incorporate social influence (imitation) and heterogeneous interconnected agents.

The simulation study shows that the proposed LCS performed well since the experimental bounded rational agents presented a tendency towards optimal decisions. Optimization was hereby not accurate but biased by peer decisions and habit and thus well resembled human decision making.

Moreover, a learning effect could be identified which is essential when mimicking human decision making. Finally it could be shown that the bounded rational agents react to environmental change while exhibiting a tendency to create habits which are not changed even if the environment changes. Summarizing, it was shown, that the application of LCS is an adequate approach to mimic human decision making in Agent-based simulations and has therefore been put forward for application in the exemplary model.

Parameter Estimation: Ultimately, the challenge to find proper input parameters for complex and large-scale Agent-based models was addressed. Using the example model for simulating behavior of a large-scale student population in the Brazilian city Recife, it was shown, how this task can be approached and which difficulties may arise.

It was demonstrated that formal estimation approaches like the state of the art approach of indirect inference are applicable but require costly preliminary simulations and do not guarantee estimation success. Especially for large-scale models, the researcher should therefore meticulously evaluate if such an approach promises meaningful outcomes.

As an alternative, heuristic procedures were proposed since they provide the opportunity to observe performance during the run-time and do not require foregoing, potentially costly simulations. Most importantly, the heuristics seem to be better suited to mirror the generally non linear relation between input parameters and output values of Agent-based models.

A Genetic Algorithm [88] was implemented for parameter estimation. Hints were given on how to improve the performance of the heuristic estimation method. It was especially found that for large network simulations a distributed approach to parameter estimation may be best suited. Moreover it was shown how the sensitive choice of input variables can positively affect the estimation success and that over simplification of the model did not yield the expected improvements of the parameter estimation.

Contrarily, the simplified model, where the bounded rational decision module of the agents was replaced by a brute-force optimization algorithm could not be estimated reasonably from real data. The fact that the complex ABM, incorporating bounded rational agents could be well calibrated to real data, while this attempt failed for the

deterministic model that consists of fully rational agents, serves as a strong argument for the validity of our approach for LCS decision making.

6.3 Contribution

It has been demonstrated then that the proposed methodology enables the researcher to create a global network model consisting of bounded rational agents, suited to simulate complex contagion processes such as the spread of positive attitudes towards education. The agents within the model hereby mimic well human behavior in the case of educational commitment. The model was fitted to real-world data by our Meta-Heuristic process, based on a Genetic Algorithm. Within this Section, the single contributions to the literature made within this thesis are described.

According to the current state of the literature, several behaviors can be well modeled as Coordination-Games (e.g alcohol use). Moreover, a social effect on educational choices has been confirmed and also successfully modeled with the Agent-based methodology. However, the current approaches to modeling educational choices mainly focus on the decision of individuals about their educational path. This means that existing ABM address choices like "will I enroll for higher education or not?". This work however, contributes firstly with the finding that also contagion of educational commitment and not only the choice of an educational path of adolescents can be sufficiently well represented by a Coordination-Game mechanism. Hereby the stage is set for the incorporation of Coordination-Game like imitation processes to a more sophisticated model of the spread of educational commitment.

Further, the problem of missing information in social network data has been tackled from different disciplines. There are several fairly well performing methods to deal with both, (i) the total absence of information about links between individuals in the network, and (ii) the randomly missing information about links within a network. Traditional solutions to the missing data problems “Fixed Choice Effect" and “Boundary Specification Problem" are applied in survey planning, dealing with the careful definition of the survey group. The “Fixed Choice Effect" seems to disturb assortativity measures and degree distributions which may explain the frequent deviation of those measures when comparing surveyed social networks to other known social networks. The problem of missing data after completing data collection has been approached by social sciences mainly under the term imputation. The set of applied mechanisms incorporates for example the estimation of missing reciprocal ties in directed networks (reconstruction), the replacement of incomplete respondents by similar alters (hot deck imputation), or using the concept of preferential attachment (assortativity)

The missing data problem for social networks is also a recent issue for researchers dealing with large social networks from online sources. Here links may be omitted

due to privacy restrictions, or missing because observed networks tend to be dynamic. The task to predict links to be established in the future or links that have been omitted due to other reasons is here called the “Link Prediction Problem".

Unsupervised measures for link prediction build on the “similarity" of nodes in terms of network properties as for example the number of common neighbors or draw from common properties of social networks such as assortativity. Other approaches are for example supervised random walks, methods based on community structure or on mutual information. Furthermore, the problem to predict links between individuals that are not part of the same data-set or platform has been successfully tackled using machine learning techniques such as classifier systems.

A special case of the “Link-Prediction Problem" occurs when no previous data of the network is available, and the network structure is to be re-build based on other information about the nodes. This problem may arise in co-purchasing networks or recommendation networks where information about the nodes is available, but connections between them are omitted or simply not informed. This special situation requires a different approach for link-prediction, since network based measures are not applicable due to the total absence of links.

Nevertheless, to the best of our knowledge it has not been studied before, how systematically missing data between isolated components from social network surveys (MNAR) may be inferred (or imputed) in order to enable simulations on a global network model. The presented research contributes to the literature with the proposal of three approaches that are capable of generating interconnected networks with different features.

We have demonstrated that the adoption of a common network generation algorithm (“Social Circles Approach/ Waxman-Model”) can solve the given task, creating interconnected networks from unconnected components where the macro network closely resembles micro structures. Moreover, a method from the literature on link-prediction from scratch (Bootstrapping Approach) was successfully adopted to the task of network interconnection. Here a network of close online-relations could be reproduced. Finally, a solution for network interconnection is provided for the case that not only “Boundary specification" but also the “Fixed choice effect" disturb the available data (Combined Approach).

Moreover, literature on Agent-based Computational Economics suggests very distinct approaches to model agent decision making. Approaches employ unconscious techniques like reinforcement learning, routine-learning approaches like replicator dynamics, belief learning methods as classifier systems or Bayesian approaches. Many of them have been proven to produce outcomes that coincide with findings from experimental economics and even econometrics. Learning Classifier Systems (LCS) have been often recommended for the modelling of human decisions in Agent-based models but particularly in the case of diffusion of educational attitudes, they have not been tested so far. This marks another contribution of this thesis.

Finally, the work contributes to the literature on calibrating and estimating input parameters for complex and large-scale Agent-based models. In line with the current research, it is argued that heuristic approaches are better suited to estimate the input parameters of such Agent-based models. This arguments comes up due to a specific features of Agent-based models: non-linear relations between input parameters and model output. A Genetic Algorithm is proposed that successfully estimates the input parameters for the simulation model so that the model closely reproduces real data. Hereby, attention is drawn to the possibility of decentral estimation of ABM and it is demonstrated that decentral estimation approaches yield much better results than attempts to global parameter estimation.

The findings on the estimation of the large-scale Agent-based simulation further provide validity to the overall methodology since the complex ABM, incorporating bounded rational agents could be well calibrated to real data, while this attempt failed for the deterministic model that consists of fully rational agents.

As a summary, the presented research provides a step by step description of drafting and implementing an Agent-based model, overcoming not only issues with missing data but also restrictions of current DSGE-Approaches for economic decision support. The estimated model is delivered, enabling the realization of a wide field of simulations of what-if scenarios regarding educational policies and measures for poverty eradication.

6.4 Relevance

The simulation model that is proposed here may help to better understand complex spreading processes, respective multiplier- and spillover effects of public policies. The model can help to highlight the value of the conditions posed by CCTs such as "Auxilio Brasil". Analysis can be conducted of what would happen if the program would be expanded to a larger share of the society. Another interesting question that can be answered with the proposed tool is whether conditions of the program should be altered, discarded, or if new conditions should be introduced. One could for example examine what would happen to the overall attitude of the student-population towards education if the "Auxilio Brasil" receivers did not have the obligation to attend school.

Further, policy makers may consider to tighten the conditions so that school must not only be attended but certain success must be reached in order to receive the full benefit. With the proposed model, it could be studied if a multiplier effect can be expected from such measures.

On the other hand there may be the fear that pupils from poorer households that enter the public schooling system have a negative effect on attitude towards education and thereby on the schooling success of children of non eligible families. This may also be studied and if identified, protective measures could be taken to prevent

the spread of the undesirable behavior. One may think of preparatory classes for children from poor families or of a system of distribution of those pupils to different classes.

Moreover, for Brazil, as well as for many other Latin-American countries, a marked divide in social diversity within schools can be identified. Particularly in Brazil, this lack of social diversity can be ascribed to a large share to the social composition of public and private schools [117]. Although social segregation can also be detected within public schools e.g. due to geographic location of the school, private schools seem to be much less diverse in terms of socioeconomic status of the pupils. Recent studies further point out that income inequality correlates with network fragmentation in towns [118] and that poverty is strongly correlated with small and less diverse social networks. Moreover, literature agrees that the school has an important role as a social space when it comes to the establishment of diverse social networks.

Hence, to reduce inequality and enhance the social capital of the poor, public administration may consider to weaken the division between public and private schools or between schools from poorer and richer districts. For instance, exchange programs could be set-up that connect those pupils from diverse social backgrounds, another measure could be joint social activities such as competitions in sports, excursions or similar. The presented model can be of great help when estimating the positive effects that could be reached when reshaping adolescent networks in such ways.

6.5 Further Research

At the overall level, it is important to apply this research to several problem contexts. Possible areas of application have been named above. Those contain but are not restricted to the testing of the potential effect of new conditions within the Conditional Cash Transfer Program recently renamed to "Auxilio Brasil" or the alteration or release of given conditions. For instance it should be analysed whether a negative spillover effect to not eligible pupils can be observed if the CCT receivers would not be obligated to attend school in order to receive the benefit and hence would dropout from school in a higher percentage.

Another interesting application of the model is the investigation of effects of contrary measures such as the introduction of a new condition for the CCT. Such measures could, for example, bond the benefit not only to minimum school attendance but also to a certain level of schooling success such as not repeating of a school year. It seems to be valuable to study, if positive network spillover effects could be triggered by such a measure.

Beyond the conditions of the CCTs, other measures promoting the diversity of social networks of the pupils may be evaluated. As pointed out before, it can be of interest, if the schooling success of the pupils would increase if certain schools were better interconnected. For example if a school from a wealthier district was connected to a

school from a poorer region. It can be observed how the attitude towards education would spread among the reshaped network and hints could be given which schools should be better interconnected.

However, before conducting those what-if simulations, parameter estimation needs to be optimised. The presented GA-Approach was capable of generating promising input parameters for a subset of the model with bounded rational agents. Hence, the estimation of parameters should be extended to the entire model and continued either using the proposed Genetic Algorithm approach but larger computational resources or, implementing alternative estimation approaches.

Moreover, before departing for the application of the proposed work, further steps for validation should be taken. To this purpose, the methodology should be applied to more data sources. A promising data-set would be for instance the network gathered via AddHealth study [61]. AddHealth works with a methodology very similar to the FUNDAJ study and is therefore predestined as a further data-source. Also, it would be most interesting to observe the performance of the methodology in that context and investigate how the different living condition of the surveyed individuals (adolescents in the United States and pupils from public schools in northeastern Brazil) affect the simulation outcomes.

In this regard an application of the methodology and the simulation model to other data-sets from other countries may deliver insight for policy makers that contribute to some of the central United Nations Sustainable Development Goals such as the reduction of poverty, the eradication of hunger, good quality of education and reduced inequality.

An aspect, which is not addressed within this thesis are inter-temporal components, representing an "aging" of relations and behaviors, modifying the influence of neighbors according to the "age" of the friendship, as well as according to the past behavior of the neighbor.

The inter-temporal nature of connections in dynamic social networks have also not been regarded for the task of network interconnection and missing data. It would certainly be valuable to incorporate recent findings on inter-temporal link-prediction to the given scenarios.

Equally important but not addressed within this work is the idea that friendship weights may differ according to the position of friends within the individual networks since people may tend to follow "role models".

Future work should also deal with the application of the presented network extrapolation techniques to combine different data sources. In the underlying case, the survey data from FUNDAJ could hereby be connected to census data. In addition, further development of the social distance between a pair of nodes might improve performance of the combined approach and the social circles approach. Applied to the data-set used within this research, this may lead to an even more plausible globally connected network.

Further investigation of link-probability - distance curve of the bootstrapping network seems to be interesting, especially when employing more individual information about the nodes. The additional groups may increase the quantity of generated links even for settings with a high threshold r and thereby heal the persisting problem of high clustering. In this case further comparison with data from online testimonial networks like Cyworld is recommended. The performance of the bootstrapping approach may be tested using more social groups with a more narrow definition. Additionally, testing the approaches on alternative data-sets may give further insights regarding the precision of the applied techniques.

Resuming, it would be of great interest, to validate the overall model as well as the single components (underlying spreading pattern, network extrapolation, decision making module and parameter estimation) to similar data sources from different contexts and particularly different countries. Beyond the positive effects of those attempts for the presented research (validation and sharpening of the methodology), this could generate insights that contribute to the construction of human capital and the eradication of poverty and other United Nations Sustainable Development Goals such as the good quality of education and reduced inequality.

Appendix A

Variable list A

TABLE A.1: **Individual Socio-demographic Variables from the Literature**

Variable	Characteristic values	PCC with mark 2
monthly per capita income	in Brazilian Reais	-0.035
sex	1 (female), 0 (male)	.042*
race	1 (white), 2 (black), 3(Parda), 4(yellow), 5(Native Brazilian)	0.000
age	1 age in years	-.227**
how often did the pupil miss school due to health issues	numeric value	-0.029
do you attend any religious service	1 (always or almost always), 2 (some times) 3 (never)	-.038*
years of study	in years	0.079
at which frequency do you study for school	1(every day), 2 (only on school days), 3 (3 days per week), 4 (less than 3 days per week),5 (only when there is an exam ahead), 6 (never or almost never)	-0,068
self esteem	an index of self esteem according to [119]	
household-size	number of persons ≥ 2	-0.0270308
dummy variable if the household receives governmental benefit	0 (no), 1 (yes)	.053**
dummy variable if parents are married	0 (no), 1 (yes)	
dummy variable if single parent family	0 (no), 1 (yes)	
parental education	highest degree of parents 1 - 18	0.55**
parent age	in years	0.031
do you like your teacher?	1 (loves), 2 (likes a little), 3 (is indifferent),4 (does not like), 5 (hates)	-.047**
do you like to go to school?	1 (loves), 2 (likes a little), 3 (is indifferent),4 (does not like), 5 (hates)	-0.038
do you feel left aside at class?	1(always or almost always), 2(some times), 3(never)	0.043
at which scale violence is an issue in your neighborhood?	1 (a big problem), 2 (a common problem all over Recife), 3 (not a problem)	0.002
at which scale drugs are an issue in your neighborhood?	1 (a big problem), 2 (a common problem all over Recife), 3 (not a problem)	-0.002926973
at which scale dirt and environmental conditions are an issue in your neighborhood?	1 (a big problem), 2 (a common problem all over Recife), 3 (not a problem)	-0.005

**significant at the 0.01 level
*significant at the 0.05 level

Appendix B

Variable list B

TABLE B.1: **Adjusted Individual Socio-demographic Variables**

Variable	Characteristic values	PCC with mark 2
is there a computer at your domicile?	1 (yes, with internet access), 2 (yes, without internet access), 3 (no)	-0.170**
have you already repeated a class?	1 (No), 2 (yes, once),3 (yes, twice or more times)	-0.145**
do parents or responsible persons help pupil with the homework?	1 (mother), 2 (other women from the family), 3 (father), 4 (other man from the family), 5 (other female person not from the family), 6 (other male person not from the family), 7 (maid or housekeeper),8 (nobody)	0.125**
which measure of transportation do you usually use to get to school?	1 (own vehicle(car or motorcycle), 2 (car sharing), 3 (public transport),4 (school bus),5 (bicycle), 6 (walking), 7 (other)	-0.123**
do parents or responsible persons check if student does the homework?	1 (mother), 2 (other women from the family), 3 (father), 4 (other man from the family), 5 (other female person not from the family), 6 (other male person not from the family), 7 (maid or housekeeper),8 (nobody)	0.120**
does the household posses a car?	1 (yes), 2 (no)	-0.120**
how frequently do you eat the lunch offered at school?	1 (always or almost always), 2 (sometimes), 3 (never or almost never),4 (no lunch offered at school)	0.111
whom of the parents or responsible persons reacts when you do something wrong?	1 (mother), 2 (other women from the family), 3 (father), 4 (other man from the family), 5 (another female person not from the family), 6 (other male person not from the family), 7 (maid or housekeeper), 8 (nobody)	0.110
does your mother take you to cinema or theater?	1 (always or almost always), 2 (sometimes), 3 (never or almost never)	-0.104**
does your father take you to cinema or theater?	1 (always or almost always), 2 (sometimes), 3 (never or almost never)	-0.100**
is there a maid or house keeper at your house	1 (yes), 2 (no)	-0.099**
do you have a washing machine at home?	1 (yes), 2 (no)	-0.098**

**significant at the 0.01 level
*significant at the 0.05 level

Bibliography

[1] P. Freire, *Pedagogia do oprimido*. Paz e Terra, 2005, ISBN: 9788577530168. [Online]. Available: https://books.google.de/books?id=m9VzSQAACAAJ.

[2] K. Lindert, A. Linder, J. Hobbs, and B. De la Brière, "The nuts and bolts of brazil's bolsa família program: Implementing conditional cash transfers in a decentralized context," *World Bank social protection discussion paper*, vol. 709, 2007.

[3] J. L. Silvernale, "Do conditional cash transfers increase school enrollment? evidence from brazil," 2021. [Online]. Available: https://repository.usfca.edu/thes/1381.

[4] E. Draeger, "Do conditional cash transfers increase schooling among adolescents?" *International Economics and Economic Policy*, vol. 18, no. 4, pp. 743–766, 2021.

[5] G. J. Bobonis and F. Finan, "Neighborhood peer effects in secondary school enrollment decisions," *The Review of Economics and Statistics*, vol. 91, no. 4, pp. 695–716, 2009.

[6] F. Brollo, K. Maria Kaufmann, and E. La Ferrara, "Learning spillovers in conditional welfare programmes: Evidence from Brazil," *The Economic Journal*, vol. 130, no. 628, pp. 853–879, May 2020, ISSN: 0013-0133. DOI: 10.1093/ej/ueaa032. eprint: https://academic.oup.com/ej/article-pdf/130/628/853/33377464/ueaa032.pdf. [Online]. Available: https://doi.org/10.1093/ej/ueaa032.

[7] O. Blanchard, "The state of macro," *Annual Review of Economics*, vol. 1, no. 1, pp. 209–228, 2009.

[8] A. Kirman, "Can artificial economies help us understand real economies?" *Revue de l'OFCE*, vol. 124, no. 5, pp. 15–41, 2012.

[9] D. Colander, P. Howitt, A. Kirman, A. Leijonhufvud, and P. Mehrling, "Beyond dsge models: Toward an empirically based macroeconomics," *The American Economic Review*, pp. 236–240, 2008.

[10] G. N. Gilbert, *Agent-based models*. Sage, 2008.

[11] N. A. Christakis and J. H. Fowler, *Connected: The surprising power of our social networks and how they shape our lives*. Little, Brown, 2009.

[12] S. Lehmann and Y.-Y. Ahn, *Complex spreading phenomena in social systems*. Springer, 2018.

[13] D. Centola and M. Macy, "Complex contagions and the weakness of long ties," *American Journal of Sociology*, vol. 113, no. 3, pp. 702–734, 2007. DOI: 10.1086/521848. eprint: https://doi.org/10.1086/521848. [Online]. Available: https://doi.org/10.1086/521848.

[14] T. Jordan, P. De Wilde, and F. B. de Lima-Neto, "Modeling contagion of behavior in friendship networks as coordination games," in *Advances in Social Simulation 2015*, Springer, 2017, pp. 181–194.

[15] D. Easley and J. Kleinberg, *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[16] T. Jordan, O. C. Pinho Alves, P. De Wilde, and F. Buarque de Lima-Neto, "Link-prediction to tackle the boundary specification problem in social network surveys," *PloS one*, vol. 12, no. 4, e0176094, 2017.

[17] T. Jordan, P. de Wilde, and F. B. de Lima Neto, "Decision making for two learning agents acting like human agents : A proof of concept for the application of a learning classifier systems," in *2020 IEEE Congress on Evolutionary Computation (CEC)*, © 2020 IEEE. Reprinted, with permission, from Jordan, Tobias and de Wilde, Philippe and de Lima Neto, Fernando Buarque, Decision making for two learning agents acting like human agents : A proof of concept for the application of a Learning Classifier Systems, 2020 IEEE Congress on Evolutionary Computation (CEC), 2020, 2020, pp. 1–8.

[18] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[19] Coordenação de Estudos Econômicos e Populacionais, Fundação Joaquim Nabuco – Fundaj, *Determinantes do desempenho escolar na rede de ensino fundamental do Recife*, 2013. [Online]. Available: https://www.gov.br/fundaj/pt-br/destaques/observa-fundaj-itens/publicacoes-e-notas-tecnicas/banco-de-dados-da-dipes-1/acompanhamento-longitudinal-do-desempenho-escolar-de-alunos-da-rede-publica-de-ensino-fundamental-do-recife-2013.

[20] M. Lelarge, "Diffusion and cascading behavior in random networks," *Games and Economic Behavior*, vol. 75, no. 2, pp. 752–775, 2012.

[21] B. Latane, "The psychology of social impact.," *American Psychologist*, vol. 36, no. 4, p. 343, 1981.

[22] A. Elsisy, B. K. Szymanski, J. A. Plum, M. Qi, and A. Pentland, "A partial knowledge of friends of friends speeds social search," *PloS one*, vol. 16, no. 8, 2021, ISSN: 1932-6203. DOI: {10.1371/journal.pone.0255982}.

[23] N. A. Christakis and J. H. Fowler, "The spread of obesity in a large social network over 32 years," *New England Journal of Medicine*, vol. 357, no. 4, pp. 370–379, 2007.

[24] J. J. Jordan, D. G. Rand, S. Arbesman, J. H. Fowler, and N. A. Christakis, "Contagion of cooperation in static and fluid social networks," *PloS one*, vol. 8, no. 6, e66199, 2013.

[25] A. D. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, vol. 111, no. 24, pp. 8788–8790, 2014.

[26] T. A. Brakefield, S. C. Mednick, H. W. Wilson, J.-E. De Neve, N. A. Christakis, and J. H. Fowler, "Same-sex sexual attraction does not spread in adolescent social networks," *Archives of Sexual Behavior*, vol. 43, no. 2, pp. 335–344, 2014.

[27] E. Marques, *Redes sociais, segregação e pobreza*. Editora Unesp, 2010, [in Portuguese].

[28] J. Kratzer and C. Lettl, "Distinctive roles of lead users and opinion leaders in the social networks of schoolchildren," *Journal of Consumer Research*, vol. 36, no. 4, pp. 646–659, 2009.

[29] E. L. Paluck and H. Shepherd, "The salience of social referents: A field experiment on collective norms and harassment behavior in a school social network.," *Journal of Personality and Social Psychology*, vol. 103, no. 6, p. 899, 2012.

[30] N. E. Friedkin, *A Structural Theory of Social Influence*, ser. Structural Analysis in the Social Sciences. Cambridge University Press, 1998. DOI: 10.1017/CBO9780511527524.

[31] M. Granovetter, "Threshold models of collective behavior," *American Journal of Sociology*, pp. 1420–1443, 1978.

[32] M. Granovetter and R. Soong, "Threshold models of interpersonal effects in consumer demand," *Journal of Economic Behavior & Organization*, vol. 7, no. 1, pp. 83–99, 1986.

[33] T. C. Schelling, "Dynamic models of segregation†," *Journal of Mathematical Sociology*, vol. 1, no. 2, pp. 143–186, 1971.

[34] L. Michell and A. Amos, *Teenage friends and lifestyle study dataset*, 1997.

[35] A. Knecht, "Networks and actor attributes in early adolescence," *ICS Codebook*, vol. 61, 2006.

[36] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[37] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.

[38] W. D. Richards and R. E. Rice, "The negopy network analysis program," *Social Networks*, vol. 3, no. 3, pp. 215–223, 1981.

[39] R. Corten and A. Knecht, "Alcohol use among adolescents as a coordination problem in a dynamic network," *Rationality and Society*, vol. 25, no. 2, pp. 146–177, 2013.

[40] P. Ormerod and G. Wiltshire, "Binge drinking in the uk: A social network phenomenon," *Mind & Society*, vol. 8, no. 2, pp. 135–152, 2009.

[41] P. M. Todd and G. Gigerenzer, "Bounding rationality to the world," *Journal of Economic Psychology*, vol. 24, no. 2, pp. 143–165, 2003.

[42] E. O. Laumann, P. V. Marsden, and D. Prensky, "The boundary specification problem in network analysis," *Research Methods in Social Network Analysis*, vol. 61, p. 87, 1989.

[43] P. W. Holland and S. Leinhardt, "The structural implications of measurement error in sociometry†," *Journal of Mathematical Sociology*, vol. 3, no. 1, pp. 85–111, 1973.

[44] F. Pinheiro, J. Pacheco, and F. Santos, "From local to global dilemmas in social networks.," *PloS one*, vol. 7, no. 2, e32114–e32114, 2011.

[45] Z.-K. Zhang, C.-X. Zhang, X.-P. Han, and C. Liu, "Emergence of blind areas in information spreading," *PloS one*, vol. 9, no. 4, e95785, 2014.

[46] M. Newman, A.-L. Barabasi, and D. J. Watts, *The structure and dynamics of networks*. Princeton University Press, 2006.

[47] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[48] J. L. Schafer and J. W. Graham, "Missing data: Our view of the state of the art.," *Psychological Methods*, vol. 7, no. 2, p. 147, 2002.

[49] G. Kossinets, "Effects of missing data in social networks," *Social Networks*, vol. 28, no. 3, pp. 247–268, 2006.

[50] M. Huisman, "Imputation of missing network data: Some simple procedures," *Journal of Social Structure*, vol. 10, no. 1, pp. 1–29, 2009.

[51] W. D. Richards, "Nonrespondents in communication network studies," *Group & Organization Management*, 1992.

[52] M. Huisman and C. Steglich, "Treatment of non-response in longitudinal network studies," *Social Networks*, vol. 30, no. 4, pp. 297–308, 2008.

[53] L. Backstrom and J. Leskovec, "Supervised random walks: Predicting and recommending links in social networks," in *Proceedings of the fourth ACM international conference on Web search and data mining*, ACM, 2011, pp. 635–644.

[54] A. Clauset, C. Moore, and M. E. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.

[55] F Tan, Y Xia, and B Zhu, "Link prediction in complex networks: A mutual information perspective.," *PloS one*, vol. 9, no. 9, e107056–e107056, 2013.

[56] E.-Á. Horvát, M Hanselmann, F. Hamprecht, K. Zweig, and S. Gómez, "One plus one makes three (for social networks)," *PloS one*, vol. 7, no. 4, e34740, 2012.

[57] V. Leroy, B. B. Cambazoglu, and F. Bonchi, "Cold start link prediction," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, Association for Computing Machinery New York, 2010, pp. 393–402.

[58] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150–1170, 2011.

[59] X. Ma, S. Tan, X. Xie, X. Zhong, and J. Deng, "Joint multi-label learning and feature extraction for temporal link prediction," *Pattern Recognition*, vol. 121, Jan. 2022, ISSN: 0031-3203. DOI: {10.1016/j.patcog.2021.108216}.

[60] G. J. de Bruin, C. J. Veenman, H. J. van den Herik, and F. W. Takes, "Supervised temporal link prediction in large-scale real-world networks," *Social Network Analysis and Mining*, vol. 11, no. 1, Dec. 2021, ISSN: 1869-5450. DOI: {10.1007/s13278-021-00787-3}.

[61] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, "Classes of small-world networks," *Proceedings of the national academy of sciences*, vol. 97, no. 21, pp. 11 149–11 152, 2000.

[62] L. Hamill and N. Gilbert, "Simulating large social networks in agent-based models: A social circle model," *Emergence: Complexity and Organization*, vol. 12, no. 4, pp. 78–94, 2010.

[63] R. Huerta-Quintanilla, E. Canto-Lugo, and D. Viga-de Alva, "Modeling social network topologies in elementary schools," *PloS one*, vol. 8, no. 2, e55371, 2013.

[64] D. Mok, B. Wellman, and J. Carrasco, "Does distance matter in the age of the internet?" *Urban Studies*, vol. 47, no. 13, pp. 2747–2783, 2010.

[65] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, "Socio-spatial properties of online location-based social networks.," *ICWSM*, vol. 11, pp. 329–336, 2011.

[66] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins, "Geographic routing in social networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 33, pp. 11 623–11 628, 2005.

[67] B. Bollobas, "Random graphs," English, in *Modern Graph Theory*, ser. Graduate Texts in Mathematics, vol. 184, Springer New York, 1998, pp. 215–252, ISBN: 978-0-387-98488-9. DOI: 10.1007/978-1-4612-0619-4_7. [Online]. Available: http://dx.doi.org/10.1007/978-1-4612-0619-4_7.

[68] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[69] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, "Structure of growing networks with preferential linking," *Physical Review Letters*, vol. 85, no. 21, p. 4633, 2000.

[70] M. E. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, p. 208 701, 2002.

[71] E. zu Erbach-Schoenberg, S. Bullock, and S. Brailsford, "A model of spatially constrained social network dynamics," *Social Science Computer Review*, p. 0 894 439 313 511 934, 2013.

[72] R. De Caux, C. Smith, D. Kniveton, R. Black, and A. Philippides, "Dynamic, small-world social network generation through local agent interactions," *Complexity*, vol. 19, no. 6, pp. 44–53, 2014.

[73] B. M. Waxman, "Routing of multipoint connections," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 1617–1622, 1988.

[74] T. M. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software: Practice and Experience*, vol. 21, no. 11, pp. 1129–1164, 1991.

[75] M. J. B. Guimarães, N. M. Marques, D. A. Melo Filho, and C. L. Szwarcwald, "Condição de vida e mortalidade infantil: Diferenciais intra-urbanos no recife, pernambuco, brasil living conditions and infant mortality: Intra-urban differentials in recife, pernambuco state, brazil," *Cad. Saúde Pública*, vol. 19, no. 5, pp. 1413–1424, 2003.

[76] K. M. Harris, C. T. Halpern, E. A. Whitsel, *et al.*, "Cohort Profile: The National Longitudinal Study of Adolescent to Adult Health (Add Health)," *International Journal of Epidemiology*, vol. 48, no. 5, 1415–1415k, Jun. 2019, ISSN: 0300-5771. DOI: 10.1093/ije/dyz115. eprint: https://academic.oup.com/ije/article-pdf/48/5/1415/30801568/dyz115.pdf. [Online]. Available: https://doi.org/10.1093/ije/dyz115.

[77] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *Proceedings of the 16th international conference on World Wide Web*, ACM, 2007, pp. 835–844.

[78] J. Leskovec and A. Krevl, *SNAP Datasets: Stanford large network dataset collection*, http://snap.stanford.edu/data, Jun. 2014.

[79] T. E. Oliphant, *A guide to NumPy*. Trelgol Publishing USA, 2006, vol. 1.

[80] H. A. Simon, "Theories of bounded rationality," *Decision and Organization*, vol. 1, no. 1, pp. 161–176, 1972.

[81] S. Leoni, "An agent-based model for tertiary educational choices in italy," *Research in Higher Education*, pp. 1–28, 2021.

[82] J. H. Holland and J. H. Miller, "Artificial adaptive agents in economic theory," *The American Economic Review*, vol. 81, no. 2, pp. 365–370, 1991.

[83] R. Marimon, E. McGrattan, and T. J. Sargent, "Money as a medium of exchange in an economy with artificially intelligent agents," *Journal of Economic Dynamics and Control*, vol. 14, no. 2, pp. 329–373, 1990.

[84] E. Başçı, "Learning by imitation," *Journal of Economic Dynamics and Control*, vol. 23, no. 9, pp. 1569–1585, 1999.

[85] U. Wilensky, *Netlogo*, http://ccl.northwestern.edu/netlogo/, Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL.

[86] E. C. Marques and H. Torres, *São Paulo: segregação, pobreza e desigualdades sociais*. Senac, 2004.

[87] A. Calvó-Armengol, E. Patacchini, and Y. Zenou, "Peer effects and social networks in education," *The Review of Economic Studies*, vol. 76, no. 4, pp. 1239–1267, 2009.

[88] J. H. Holland, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.

[89] P. Davidsson, "Agent based social simulation: A computer science view," *Journal of Artificial Societies and Social Simulation*, vol. 5, no. 1, 2002.

[90] L. Tesfatsion and K. L. Judd, *Handbook of Computational Economics: agent-based computational economics*. Elsevier, 2006, vol. 2.

[91] M. A. Barbosa Jr and F. B. de Lima Neto, "Distributed agent-based social simulations: An architecture to simulate complex social phenomena on highly parallel computational environments," in *Intelligent Agent (IA), 2011 IEEE Symposium on*, IEEE, 2011, pp. 1–8.

[92] F. J. Miguel, J. A. Noguera, T. Llacer, and E. Tapia, "Exploring tax compliance: An agent-based simulation.," in *ECMS*, 2012, pp. 638–643.

[93] M. Salgado, E. Marchione, and N. Gilbert, "Analysing differential school effectiveness through multilevel and agent-based modelling," *Journal of Artificial Societies and Social Simulation*, vol. 17, no. 4, p. 3, 2014, ISSN: 1460-7425. DOI: 10.18564/jasss.2534. [Online]. Available: http://jasss.soc.surrey.ac.uk/17/4/3.html.

[94] M. Meyer, I. Lorscheid, and K. G. Troitzsch, "The development of social simulation as reflected in the first ten years of jasss: A citation and co-citation analysis," *Journal of Artificial Societies and Social Simulation*, vol. 12, no. 4, p. 12, 2009.

[95] L. Tesfatsion, "Agent-based computational economics: Growing economies from the bottom up," *Artificial Life*, vol. 8, no. 1, pp. 55–82, 2002.

[96] T. Brenner, "Agent learning representation: Advice on modelling economic learning," *Handbook of Computational Economics*, vol. 2, pp. 895–947, 2006.

[97] J. Duffy, "Agent-based models and human subject experiments," *Handbook of Computational Economics*, vol. 2, pp. 949–1011, 2006.

[98] O. Sigaud and S. W. Wilson, "Learning classifier systems: A survey," *Soft Computing*, vol. 11, no. 11, pp. 1065–1078, 2007.

[99] J. Grazzini and M. Richiardi, "Estimation of ergodic agent-based models by simulated minimum distance," *Journal of Economic Dynamics and Control*, vol. 51, pp. 148–165, 2015.

[100] S. R. Eliason, *Maximum likelihood estimation: Logic and practice*, 96. Sage, 1993.

[101] D. McFadden, "A method of simulated moments for estimation of discrete response models without numerical integration," *Econometrica*, vol. 57, no. 5, pp. 995–1026, 1989, ISSN: 00129682, 14680262. [Online]. Available: http://www.jstor.org/stable/1913621.

[102] C. Gourieroux, A. Monfort, and E. Renault, "Indirect inference," *Journal of Applied Econometrics*, vol. 8, S85–S85, 1993.

[103] M. Gilli and P. Winker, "A global optimization heuristic for estimating agent based models," *Computational Statistics & Data Analysis*, vol. 42, no. 3, pp. 299–312, 2003.

[104] P. Winker, M. Gilli, and V. Jeleskovic, "An objective function for simulation based inference on exchange rate data," *Journal of Economic Interaction and Coordination*, vol. 2, no. 2, pp. 125–145, 2007.

[105] C. Bianchi, P. Cirillo, M. Gallegati, and P. A. Vagliasindi, "Validating and calibrating agent-based models: A case study," *Computational Economics*, vol. 30, no. 3, pp. 245–264, 2007.

[106] G. L. Ciampaglia, "A framework for the calibration of social simulation models," *Advances in Complex Systems*, vol. 16, no. 04n05, p. 1 350 030, 2013.

[107] D. A. Reynolds, "Gaussian mixture models.," *Encyclopedia of Biometrics*, vol. 741, pp. 659–663, 2009.

[108] M. Gilli and P. Winker, "Heuristic optimization methods in econometrics," *Handbook of Computational Econometrics*, pp. 81–119, 2009.

[109] B. Calvez and G. Hutzler, "Automatic tuning of agent-based models using genetic algorithms," in *International Workshop on Multi-Agent Systems and Agent-Based Simulation*, Springer, 2005, pp. 41–57.

[110] S. Alaliyat, H. Yndestad, and P. I. Davidsen, "Optimal fish densities and farm locations in norwegian fjords: A framework to use a pso algorithm to optimize an agent-based model to simulate fish disease dynamics," *Aquaculture International*, vol. 27, no. 3, pp. 747–770, 2019.

[111] F. Pappalardo, M. Pennisi, F. Castiglione, and S. Motta, "Vaccine protocols optimization: In silico experiences," *Biotechnology Advances*, vol. 28, no. 1, pp. 82–93, 2010, ISSN: 0734-9750. DOI: https://doi.org/10.1016/j.biotechadv.2009.10.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0734975009001797.

[112] J. C. Thiele, W. Kurth, and V. Grimm, "Facilitating parameter estimation and sensitivity analysis of agent-based models: A cookbook using netlogo and 'r'," *Journal of Artificial Societies and Social Simulation*, vol. 17, no. 3, p. 11, 2014, ISSN: 1460-7425. DOI: 10.18564/jasss.2503. [Online]. Available: http://jasss.soc.surrey.ac.uk/17/3/11.html.

[113] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[114] M. D. Morris and T. J. Mitchell, "Exploratory designs for computational experiments," *Journal of Statistical Planning and Inference*, vol. 43, no. 3, pp. 381–402, 1995.

[115] C. Lucasius and G. Kateman, "Understanding and using genetic algorithms part 1. concepts, properties and context," *Chemometrics and Intelligent Laboratory Systems*, vol. 19, no. 1, pp. 1–33, 1993, ISSN: 0169-7439. DOI: https://doi.org/10.1016/0169-7439(93)80079-W. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016974399380079W.

[116] S. Panda and N. P. Padhy, "Comparison of particle swarm optimization and genetic algorithm for facts-based controller design," *Applied soft computing*, vol. 8, no. 4, pp. 1418–1427, 2008.

[117] OECD, *Balancing School Choice and Equity*. 2019, p. 108. DOI: https://doi.org/https://doi.org/10.1787/2592c974-en. [Online]. Available: https://www.oecd-ilibrary.org/content/publication/2592c974-en.

[118] G. Tóth, J. Wachs, R. Di Clemente, *et al.*, "Inequality is rising where social network segregation interacts with urban topology," *Nature Communications*, vol. 12, no. 1, pp. 1–9, 2021.

[119] M. B. C. Gonçalves, I. P. D. A. Raposo, S. M. F. P. O. Gomes, *et al.*, "A relação entre habilidades não-cognitivas e desempenho escolar," in *Anais do XLII Encontro Nacional de Economia [Proceedings of the 42nd Brazilian Economics Meeting]*, ANPEC-Associação Nacional dos Centros de Pos-Graduação em Economia, 2016.