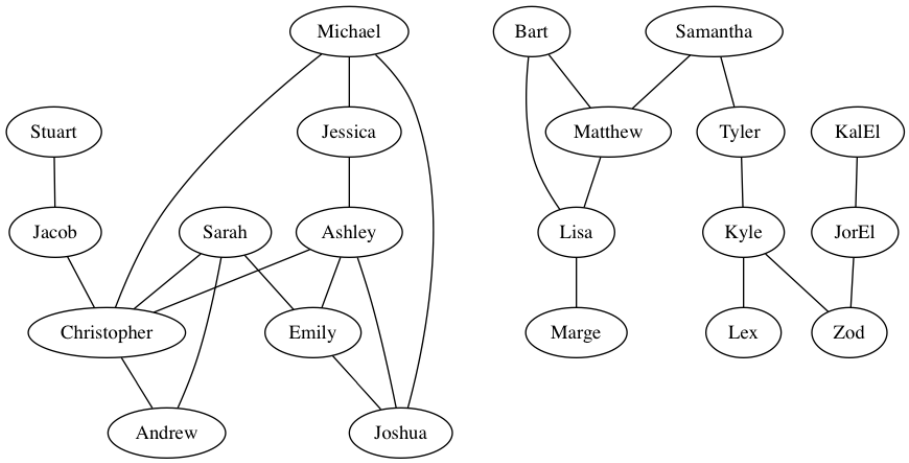
university of groningen / faculty of mathematics and natural sciences

A longitudinal analysis

Co-evolution of friendship and academic performance in an international setting



Lianne Jansen
Master thesis Mathematics (SBP-track)
Supervisor: Prof. dr. E.C. Wit
Second supervisor: Nynke Niezink, MSc
October 2014-September 2015

Picture on the front cover is taken from [1].

Executive summary

Did you ever wonder why you are friends with your friends? What makes it that you are friends with some people and not with others? Is it a common behaviour that attracts you in others, and if that is the case, can you influence that behaviour?

Questions like this are studied in literature. It is shown that friendships are indeed influenced by behaviour. For example at high schools the grades play a role in the friendship formation. There are clusters of students with high grades and clusters of students with low grades. Other research shows that adolescents form groups based on smoking behaviour or music taste. The existence of these clusters is not new, but the interesting question is how these clusters are formed. Is it the case that students with high grades choose friends with high grades - and smokers choose other smokers as friends? Or do the clusters with smarter students arise because these students influence each other's behaviour; by doing their homework together, friends obtain a higher grade - and smokers influence non-smoking friends to try a cigarette. The first scenario is called "selection"; people choose their friends based on a similar behaviour variable. The second scenario is "influence"; friends influence each other's behaviour. It can be useful to know which of the two mechanisms has the largest effect (for example for policy purposes). With the help of a mathematical model, the stochastic actor-based model, the distinction between "selection" and "influence" can be made.

The Faculty of Mathematics and Natural Sciences of the University of Groningen introduced the international bachelor in the academic year 2013-2014. This international bachelor ensures that courses are taught in English, what makes it possible for foreign students to attend the courses. This gives rise to a wider variety of different backgrounds that are present in the first year of study. The question now is if these different backgrounds influence the friendship formation and the academic performance. This question is studied in this report by distributing questionnaires to first year mathematics students three times during the academic year 2014-2015. From these questionnaires the development of the friendship networks and the academic performances of the students can be obtained. This provides insight in the cluster formation, for example if there are clusters of people with high grades or people with low grades (as was the case for high schools). If this is the case, it can be studied which mechanism is predominant. The stochastic actor-based model states that the actors, the friends, can choose if they start or end a friendship. This starting or ending of a friendship is only allowed in the model at certain small time intervals. In this way it will become clear which change in the network (the friendship network) or the behaviour (the academic performance) happens first. "Influence" and "selection" can be distinguished in this way. This report discusses the Monte Carlo algorithm that is used in order to see if the "influence" or the "selection" mechanism predominates.

This master project shows that there is no influence of the academic performance on the friendship networks or vice versa. That means that there are no groups of friends who only have high (or low) grades. This group formation based on performance is the case at high schools, but the difference in levels at university is probably smaller. People only start studying mathematics when they already obtained high math grades at high school. Furthermore, this research shows that the different nationalities mix very well. There are no separate groups of Dutch students or international students. This is good news for the set-up of the international bachelor; it really turned into an international bachelor. Remarkable is the fact that international students take the initiative to become friends with Dutch students. The nationality does not influence the academic performance. This means that differences in entry level are resolved in a good manner. Furthermore, there are no other effects that influence the academic performance. Neither the attendance of lectures and tutorials, the gender, the living situation nor the usage of English has an influence on the academic performance. The development of the friendship network is influenced by the gender and the living situation of the students, but not by the usage of English or the presence at lectures and tutorials. Female students are more often nominated to be a friend than male students. Furthermore, students who live with their parents are more popular than the students that do not. This is not self-evident, but it seems that female students and students who live with their parents try harder to be seen as nice persons. For the students who live with their parents, this could be caused by the fact that they mostly are not a member of a student association. This makes them more eager to make friends in the study environment.

A mathematical model that tries to predict social developments is a tough task that is seldom perfect. This report studied the co-evolution of friendship formation and academic performance, but there are many more possibilities to obtain insight in the selection and influence mechanisms in friendship networks. From the questionnaires a data set was obtained, which shows many opportunities for future research.

Contents

Chapter 1

Introduction

Some people are friends, some people are not. Friendships are started and ended again. Why are these relations formed and ended? What makes it that you choose your friends to be your friends? Is it a common behaviour factor, is it just because you meet each other often due to common friends, are there other influences that are important (for example a similar background or a similar sex)?

There are multiple papers about friendships that study questions like this. An interesting observation in friendships is often that people with a similar behaviour form clusters in the network. Jennifer Flashman studies this cluster formation at high schools [2]. The influence of the performance at school as a factor on the friendship formation is studied. This paper indeed showed that people with high grades are friends with people with high grades and also people with low grades are friends with each other. Then the question can be asked if people with high grades basically choose others with high grades to become friends and that these people end the friendships with the ones with lower grades. This mechanism is called selection and is based on the homophily principle. Shortly explained this means that it is easier or more rewarding for a person in a network to interact with similar persons than with dissimilar persons [3]. Therefore it is more likely that people select similar people to be friends, than dissimilar people. The other option is that friends stimulate each other to study hard and to do the homework together and in this way a group that obtains high grades is formed. For the people with low grades this can also work in this way, but they stimulate each other to go out and not to do the homework. This mechanism is called influence and this is based on the assimilation principle. This principle states that a person in a network adopts his own individual characteristics to match his social neighbourhood [4]. So that means that the behaviour of your friends also influences your own behaviour (in this case the amount of study hours of your friends influences your own study behaviour). [2] shows that the selection mechanism is more important than the influence mechanism for the academic performance. Friendships are formed between high grades-high grades and between low grades-low grades, and the bonds between high grades-low grades are broken.

A similar question is asked in [5], but now the academic performance is replaced by the question to what extend people like school. It is examined if the fact that people like or don't like school has effect on the friends that they have. This report also takes a look at the network of people that don't like each other. Besides these reports about the school-behaviour, also the effect of alcohol, smoking and music taste on the development of friendship networks is studied (for example in [6], [7] and [8]).

We are interested in a similar subject, namely how the academic performance at university and friendship networks evolve together over time. An interesting detail now is that the first year university students that we study face the newly formed international bachelor. This means that the bachelor is completely taught in English. Therefore more and more international students enter the first year. We would like to study if this international setting has influence on the evolution of friendship networks and on the academic performance. We ask if people with high grades also have friends with high grades, if international students mix with Dutch students, if international students get higher or lower grades and if the different attitudes towards the English bachelor have influence on the social network. This study of the friendship networks is unique by the international setting. It is also the first study where academic results are studied in a university setting, until now these kind of studies were performed at high schools.

In order to complete this study, three waves of data collection were done. First year mathematics students of the University of Groningen in the 2014-2015 cohort were asked to complete a questionnaire. More information about the set-up of the questionnaire can be found in chapter 2 of this thesis. After the data collection the data was studied with the help of a stochastic actor-based model. The theory behind this model will be discussed in chapter 3 and 4. The analysis of the data by the RSiena package in R (this is the software package of the stochastic actor-based model), is discussed in chapter 5 and 6 and the thesis ends with a conclusion in chapter 7.

Chapter 2

Research questions and method

In this chapter the background of this study is explained. The research questions are stated and it is explained what kind of information is needed. Also the sampling method and data collections are explained.

2.1 Research questions

In this research project we would like to study the co-evolution of the friendship network and the academic performance of first year mathematics students in the (new) international setting. We are curious to see if in this group of university students also forms clusters of people with high grades and clusters of people with low grades, as was the case for the high school students. Furthermore we are interested if the international students form clusters and if they have higher or lower grades than the Dutch students. We think that the attitude and behaviour of the Dutch students (if they talk English all the time for example and if they like the international character) influences the formation of friendships between Dutch and international students. A big difference between the university and high school is the voluntary versus compulsory character. The students at university have the choice to go to the lectures and tutorials. The question is if their presence at lectures and tutorials influences their performance and their friendship network. As a summary the following questions will be investigated in this thesis:

- Are there clusters of people with high grades and clusters of people with low grades? And if this is the case, are these clusters formed by selection or by influence?
- Do the international students get higher grades than the Dutch students and does this influence the number of friends that they have?
- Is there nationality homophily observed? In other words do the international students mix with the Dutch students or do they form a separate cluster?
- Does the attitude towards the international character of the bachelor have an influence in the friendship formation between Dutch and international students?
- Does the presence of the students at lectures and tutorials have an influence on the academic performance and/or on the friendship networks?

2.2 Method

In order to answer the research questions, data is needed. This data is sampled from the first year mathematics students (cohort 2014-2015) at the University of Groningen. We asked all these students three times during the academic year (in the end of November 2014, in the end of February and in the beginning of June 2015) to fill in a questionnaire. At the start of the academic year, each student has been assigned a mentor and in order to reach all the students we contacted their mentors. Mentors have meetings with their students that are compulsory. This way we are certain to be able to reach all first year mathematics students (although participation remains completely voluntary). We asked the mentors to distribute the questionnaires right after their mentor meeting. For the second and third measurement the mentor meetings were already ended and therefore the questionnaires were distributed during a compulsory (practical) course (in order to reach all students). Before the first measurement a consent letter was handed to all students. This letter described the purpose of the study and invited the students to participate. The students who would like to participate signed the letter and completed the questionnaire. The study was approved by the ethical committee of social sciences of the University of Groningen. In total 60 students participated in the study, representing 91% of the population of first year mathematics students of the University of Groningen. Not everybody completed all three measurements (for more details see section 5.1.3). The consent letter and the questionnaire can be found in the appendix.

In order to get the information that we need, the students have to complete a questionnaire. Which variables are obtained from the questionnaire and used in the study are discussed below.

- **Friendships**
  Together with the questionnaire there was a list with names and numbers. The students were asked to indicate which of their fellow students they see as their friends. This is done by filling in the numbers that correspond to their friends in the available spots on the questionnaire. There is no restriction about the number of friends they have to/may nominate. Some of the students nominated nobody and others nominated up to 15 friends. The students were also asked to circle their best friends. The study is performed on the network of the friends that they indicated, but a follow-up study might be done with the best friends network.

- **Variables needed for analysis**
  There are some background variables needed from the students in order to do the analysis and to answer the research questions. This background variables consist of the grades, the nationality, the attitude towards the international bachelor and the presence at lectures and tutorials. These information is asked for in the questionnaires as explained below for each variable.

  - *Grades*
    In each questionnaire the student was asked to indicate the grades that he/she obtained in the last exam period. The student can write down the course and the grade.

- *Nationality*
  The students were asked what their nationality is.

- *Attitude towards the international bachelor*
  In order to get an idea about the attitude towards the international bachelor, the students were asked to agree or disagree on a fivefold scale with these statements: "I think the international character of the bachelor is an advantage to the study", "I find it hard that everything is in English", "If it was possible to do a Dutch bachelor mathematics in Groningen, I would prefer that over the international bachelor". Furthermore, in order to get an idea about the behaviour of the students towards the international character of the bachelor, the students were asked to indicate on a five fold scale (from none of the time to all of the time) to what extend the following propositions apply to them: "I speak English during tutorials", "I speak English during breaks".

- *Presence*
  In order to get an idea about the presence of the students at lectures and tutorials, the students were asked to indicate how many hours of lectures and how many hours of tutorials they attend in a typical study week.

- **Control variables**
  In order to find the right effects of the variables that we are interested in, we need to control for the variables what we think have an effect on the friendship formation. The first control variable that is included is the effect of the gender. It can be the case that girls tend to form friendships with other girls and boys with boys. Therefore the students were asked in the questionnaire what their gender is. The other variable that we control for is the fact that the students live at their parents or that they live away from home (mostly somewhere in Groningen). This difference can have an effect on the friendship formation and therefore we asked the students to indicate what their living situation is.

Chapter 3

Stochastic actor-based model

For the analysis of the data that are obtained from the first year students, a stochastic actor-based model will be used. In this chapter the theoretical background of this model will be discussed. The parameter estimation and the inference of the model will be covered in the next chapter. This theoretical background is all incorporated in the software RSiena, which is an abbreviation for R-Simulation Investigation for Empirical Network Analysis. This software will be used to analyse the data.

3.1 General idea and definitions

Suppose that the group that we study (in our case the first year mathematics students), consists of N actors. Within this group there are relations, for example friendships. The actors together with all relationships will form a network. Furthermore all actors have different behaviours. For example how many hours they study at home, how many lectures they follow, if they smoke and more behaviours like this. Moreover each actor has its own characteristics, like his background, the place where he is raised and so on. These characteristics are called the actor covariates. The characteristics between two actors are called the dyadic covariates. For example the distance between the places of living of two actors, the fact that two actors have the same sex and things like this are included by the dyadic covariates. Covariates can be constant or variable over time (for example the covariate that measures the similarity of the sexes will be constant, whereas the distance between the actors can change over time). An important difference between the behaviour variables and the covariates lies in the assumption that the covariates can have influence on the behaviour, but the behaviour variables cannot have influence on the covariates. For the network there is a similar assumption: the covariates can have an influence on the network, but the network has no influence on the covariates [5]. That means that for the modelling we have to be very careful in determining the covariates and the behaviour variables.

With the help of the stochastic actor-based model, relationships between the network, the behaviour of the actors and the covariates of the actors can be studied. By collecting longitudinal data, information about the network, behaviour and covariates at different moments in time can be obtained. When there are at least two different measurements, the stochastic actor-based model analyses which factors influence the change of the network and behaviour. These factors may have influence on the network, the behaviour and the covariates.

3.2 Mathematical notation

The whole friendship network can be summarized as a graph. If we define the actors $i = 1, ..., N$, then we can represent the pattern of links between them in an adjacency matrix X. This is a binary network where $X_{ij} = 0$ if there is no tie from actor i to actor j. If there is a tie from i to j, $X_{ij} = 1$. This means that actor i calls actor j his relation partner at time t, where we define a continuous time parameter t, with the observation moments called $t_1, ..., t_M$. Matrix X is called the adjacency matrix or digraph. X_{ij} is called the tie indicator or the tie variable. In this representation there are directed relations $i \rightarrow j$, where there is a sender and a receiver. The sender i is called *ego*, the receiver j is called *alter*. The relationships are all unilateral, so it is possible that $X_{ij} = 1$, whereas $X_{ji} = 0$. Furthermore $X_{ii} = 0$ for all i. The dependence on the time is made more explicit by writing $X(t)$.

The behaviour variable is indicated by $Z(t)$. This variable is measured at the same times as the network is measured. It consists of H components ($H \geq 1$), which indicate the different behaviour parts that are under investigation. This can be for example the study hours, the smoking behaviour and so on. Therefore $Z(t)$ can be written as $Z(t) = (Z_1(t), Z_2(t), ..., Z_H(t))$. The value that indicates the value of behaviour h at time t for individual i is denoted by $Z_{hi}(t)$.

The actor-dependent covariates will be indicated by v. The value of covariate k for actor i at time t is denoted by $v_i^k(t)$. The dyadic covariates are indicated by w. The value of the dyadic covariate k of actor i and j at time t is denoted by $w_{ij}^k(t)$.

The stochastic process $(X(t), Z_1(t), ..., Z_H(t))$ together with the covariate data is represented by the symbol $Y(t)$. The available data is denoted by $y(t_1), ..., y(t_M)$.

Example
A small example to get used to the notation; a friendship network of five persons, connected in the way as is shown in figure 3.1.

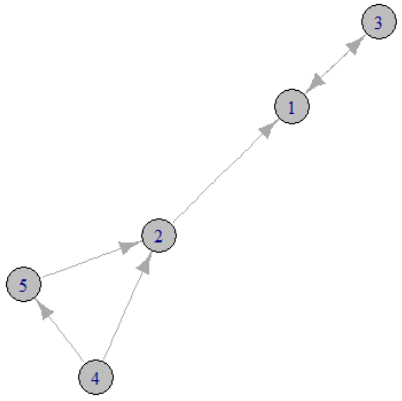


Figure 3.1: The friendship relations between five persons (1-5)

The arrows indicate who assigns who as his friend. The adjacency matrix X that belongs to this situation is given by

$$X = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

Now the behaviour can be modelled by the formation of Z. For example we ask if the people smoke or not. This behaviour component is then indicated by Z_1. $Z_{1i} = 1$ if actor i smokes and $Z_{1i} = 0$ if this is not the case. Another behaviour component can be the hours that the actors spend on studying. The answers to this question will be stored in Z_2. The total matrix Z for this measurement of these two behaviour components will then be

$$Z = (Z_1, Z_2) = \begin{pmatrix} 1 & 10 \\ 1 & 15 \\ 0 & 8 \\ 0 & 6 \\ 1 & 15 \end{pmatrix}.$$

In this representation we would like to include the information about the gender of the actor and the nationality of the actor. This can be done by the introduction of covariates. In this case the first covariate will be the information about the gender. This will be stored in the variable v_1, where $v_{1i} = 1$ if actor i is male, and it is 0 when actor i is female. v_1 in this case looks like this

$$v_1 = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

In for the second covariate v_2, $v_{2i} = 1$ if actor i is Dutch, and $v_{2i} = 0$ if the actor is not Dutch. In this example v_2 is given by

$$v_2 = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 \end{pmatrix}.$$

In this case both covariates are constant over time. We can also include variable things like age or place of residence. These covariates will vary over time, so for different measurements in time, different values of v will be obtained.

This is a very simple example, only meant to get used to the notation. For the real data set, there will be more measurements over time and there are more actors involved. When this data is obtained, the stochastic actor-based model will be applied. The theory behind this model will be explained in the upcoming sections.

3.3 Assumptions

The stochastic actor-based model needs some assumptions. These assumptions are described in [9], [7] and [5]. The assumptions are explained below.

1. The general idea of a stochastic actor based model (or also called stochastic actor-oriented model) is that the actor (the individual) can decide to add or to remove a connection to another person or to make changes in his behaviour. The assumption

is that the actors control their outgoing ties X_{ij} and their characteristics Z_{hi}. This assumption states that changes in ties are made by the actors who send the tie, based on their position in the network, their and others' behaviour, their perceptions about the network and so on. It does not directly mean that actors can change their ties at will. In this way the network evolves as a stochastic process that is "driven by the actors" [9].

2. The underlying time parameter is continuous. This means that the change process takes place in time steps of varying length, which can be very small. However, the parameter estimation is based on observations at discrete time points $t_1 < t_2 < ... < t_M$. At least two observations are needed to estimate the parameters.

3. The changing network is the outcome of a Markov process. This means that only the current state of the network determines the probability of change in the future; there are no effects from the past. All information is present in the current state. This assumption limits the applicability of the model, but it is hard to build a model without this assumption. The model now is meaningful if the network $X(t)$ and the behaviour $Z_h(t)$ together can be regarded as a state with, in a reasonable approximation, endogenous dynamics of these variables themselves [7]. Therefore the model cannot be applied to ephemeral phenomena or brief events for which a dependence on latent variables would be plausible [7]. So for events like going to a movie or email exchange this model will not give reliable results. For events that can be considered as states the model can be applied. For example for the dynamics of friendships and lifestyle-related behaviour or for strategic alliances between companies the model can be used.

4. At any given time only one actor gets the opportunity to change a tie. The actor will be probabilistically selected and he/she can change not more than one tie at the time. This implies that the changes cannot be coordinated and thus it is not the case that a reciprocal tie is formed at once. The actors act conditionally independent of each other. There must have been someone with the initiative and one who reciprocated it. This assumption excludes the networks that are coordinated, however for directed networks it is a reasonable simplifying assumption.

5. The changes in network and behaviour cannot be done at the same time. An actor can only change his network position or its behaviour at a given time t. The probability for simultaneous changes is zero [7].

6. At any given time only one edge or one behaviour component can be changed. For the behaviour components the change can only consist of one unit up or down.

After having a closer look at the assumptions, the model can be divided into two parts. At a single moment in time only one actor may make a change. At this time the selected actor i can choose to add a new tie, to remove a tie, to change his behaviour with one unit, or to do nothing. So the model consists of the waiting times until the next opportunity for a change made by actor i on one hand and the probabilities of changing X_{ij} and Z_{ih} conditional on the opportunity of change on the other hand.

With this decomposition between the timing model and the model for change, the development of the model can be depicted as follows: at randomly determined moments t, the actor i has the opportunity to change a tie or a behaviour variable X_{ij}. The smallest changes that are possible in the evolution of the network and the behaviour are called

micro steps. The time between these micro steps can be modelled, as will be explained in the next section.

The data at the measurement moment t_1 will be the beginning point of the stochastic process. Then the model will consist of two parts.

- The time between the micro steps. So we would like to model the moments that the actors get the opportunity to make a change.
- The types of changes. That means that we model which change an actor makes when he has the opportunity to make a change.

3.4 Waiting times

In the stochastic actor-based model it is assumed that the changing network is the outcome of a Markov process. That means that only the current state of the network determines the probability of change in the future. Therefore a distribution with a memoryless property is needed to derive the times between the micro steps. In order to model these waiting times before an actor gets the opportunity to make a change, the exponential distribution is taken. This distribution has the memoryless property, what states that $P(T > t+s|T > t) = P(T > s)$. For the exponential distribution it can be derived that this is indeed the case. For the exponential distribution we have $P(T = t) = \lambda e^{-\lambda t}$ and $P(T > t) = e^{-\lambda t}$. Then we also have $P(T > s + t) = e^{-\lambda(s+t)}$. Therefore

$$P(T > t + s|T > t) = \frac{P(T > s + t)}{P(T > t)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-\lambda s}.$$

This is exactly $P(T > s)$. So now we showed $P(T > t+s|T > t) = P(T > s)$ and thus the exponential distribution is memoryless. Furthermore, the exponential distribution is the only distribution that is memoryless (as can be shown with the use of a survival function). Therefore the exponential distribution is taken in order to get a distribution where the time between the micro steps t_m and t_{m+1} is independent of the micro steps in the past (t_k, where $k < m$).

We can take $T_i^{[X]}$ as the variable for the waiting time between changes in the network and we take $T_i^{[Z_h]}$ as the waiting time between changes in the behaviour. These two variables have an exponential distribution, so we say $T_i^{[X]} \sim \exp(\lambda_i^{[X]})$ and $T_i^{[Z_h]} \sim \exp(\lambda_i^{[Z_h]})$. From the exponential distribution it is known that the expected waiting times for the network will be $\frac{1}{\lambda_i^{[X]}}$ and for the behaviour it will be $\frac{1}{\lambda_i^{[Z_h]}}$. All waiting times will be independent.

Now the distribution of the waiting times until there is a micro step for the network or the behaviour will be investigated. We are interested in the time it takes before there is an actor i who is allowed to make a change in the network or the behaviour. That is the same as studying the distribution of the variable T^* that is defined as

$T^* = \min(T_1^{[X]}, ..., T_N^{[X]}, T_1^{[Z_h]}, ..., T_N^{[Z_h]})$. For T^* the following holds:

$$\begin{aligned} P(T^* > t) &= P(\min(T_1^{[X]}, ..., T_N^{[X]}, T_1^{[Z_h]}, ..., T_N^{[Z_h]}) > t) \\ &= P(T_1^{[X]} > t) \cdot ... \cdot P(T_N^{[X]} > t) \cdot P(T_1^{[Z_h]} > t) \cdot ... \cdot P(T_N^{[Z_h]} > t) \\ &= e^{-\lambda_1^{[X]} t} \cdot ... \cdot e^{-\lambda_N^{[X]} t} \cdot e^{-\lambda_1^{[Z_h]} t} \cdot ... \cdot e^{-\lambda_N^{[Z_h]} t} \\ &= e^{-\sum_{i=1}^{N} (\lambda_i^{[X]} + \lambda_i^{[Z_h]}) t} \end{aligned}$$

From this it can be seen that T^* is exponentially distributed with parameter $\lambda_{total} = \sum_{i=1}^{N} (\lambda_i^{[X]} + \lambda_i^{[Z_h]})$. That means that the waiting time before a randomly chosen actor i gets the opportunity to make a change to the network or to the behaviour is exponentially distributed with parameter λ_{total}. The probability that the micro step for actor i is a micro step for the network is $\frac{\lambda_i^{[X]}}{\lambda_{total}}$ and the probability that is for the behaviour is $\frac{\lambda_i^{[Z_h]}}{\lambda_{total}}$.

It is possible that there is heterogeneity in the activity of the actors. Some actors may change their network ties or their behaviour more often than others [6]. These differences might be present due to sex differences or by the existing network structure. This can be incorporated by introducing λ's that depend on actor attributes and network positions. However, throughout this thesis, the assumption will be made that the waiting times for the different actors are equally distributed. Therefore for each actor the waiting time between two micro steps in the network is exponentially distributed with parameter $\lambda_m^{[X]}$. The waiting time between two micro steps in the behaviour is exponentially distributed with parameter $\lambda_m^{[Z_h]}$ for each actor. Here the index m only indicates a time period with $m \in [1, M]$, not an individual any more. This is also done in [6] and [5].

3.5 Change determination model

After the first step where an actor is chosen that will make the changes, the way of changing needs to be determined. The selected actor may change one outgoing tie or he may change his behaviour. He may add a tie, remove a tie or do nothing in the network or he may move one level up, down or do nothing in his behaviour. In order to model which change will take place when actor i gets the opportunity to make a change, the so-called objective functions are important. First we will have a look at the objective function for the network and after that the objective function for the behaviour will be discussed.

3.5.1 Dynamics of the network

When an actor gets the opportunity to make a change to the network, he has to determine what change that will be. The probability of change for the network depends on the so-called objective function. This function measures how likely it is for the actor to change his network in a particular way. In practice the objective function will depend on the personal network of the actor, what means the network of the actor and the actors where there is a direct tie to, as well as the covariates of these actors. So in the end the probabilities of changing a tie will depend on the personal networks that would be formed when the possibly changes are made, together with their covariates.

The objective function specifies in which direction the change will take place. It indicates what the preference of the actors is to change the network. When actor i gets an

opportunity for network-change, this actor can make a change to the network; he can remove one tie, he can add one tie or he can do nothing. This gives in total n possibilities. Which of these events is the most likely depends on the objective function $f_i(\beta^{[X]}, y)$. The choice probabilities for the network changes are given by

$$\Pr(x(i \rightsquigarrow j)|x(t), z(t)) = \frac{\exp(f_i^{[X]}(\beta^{[X]}, x(i \rightsquigarrow j)(t), z(t)))}{\sum_k \exp(f_i^{[X]}(\beta^{[X]}, x(i \rightsquigarrow k)(t), z(t)))}. \tag{3.1}$$

Here $f_i^{[X]}$ again denotes the objective function, $x(i \rightsquigarrow j)$ means for $j \neq i$ the network resulting from a micro step in which actor i changes the tie variable to actor j (from 0 to 1, or vice versa) and $x(i \rightsquigarrow i)$ is defined to be x. That means that for $i \neq j$, $x(i \rightsquigarrow j)_{ij} = 1 - x_{ij}$, but all other elements of $x(i \rightsquigarrow j)$ are equal to the elements of x [7]. The network-objective function $f_i^{[X]}$ consists of network effects (endogenous) and covariate effects (exogenous). A widely used definition of this objective function is a weighted sum of the various effects $s_{ik}^{[X]}(y)$,

$$f_i^{[X]}(\beta^{[X]}, y) = \sum_k \beta_k^{[X]} s_{ik}^{[X]}(y).$$

The weights of the effects $s_{ik}^{[X]}(y)$ are indicated by $\beta_k^{[X]}$. These are indicators of the strengths of the effects. If $\beta_k^{[X]} = 0$, then the corresponding effect does not play a role. If $\beta_k^{[X]} > 0$, then the probability to change to a certain state is larger when the corresponding effect is larger and the opposite holds for the case where $\beta_k^{[X]} < 0$. Now only the possible effects $s_{ik}^{[X]}(y)$ need to be specified in order to do the full model specification for the simple model. The effects can be divided into network effects and covariate effects and will be discussed in section 3.5.4.

3.5.2 Dynamics of the behaviour

The actor i can also get the opportunity to make a change in his behaviour in stead of in his network. When actor i gets an opportunity for change in the behaviour, this actor can move one level up, one level down or do nothing to his behaviour component h. In total there are H behaviour components. Which change in the behaviour is the most likely depends on the objective function for the behaviour $f_i^{[Z_h]}$

The actor can change his behaviour by changing one level in category h. Conditional on the fact that actor i is allowed to make a change in the behaviour, the choice probability is given by

$$\Pr(z(i \updownarrow_h \delta)|x(t), z(t)) = \frac{\exp(f_i^{[Z_h]}(\beta^{[Z_h]}, x(t), z(i \updownarrow_h \delta)(t)))}{\sum_{\tau \in \{-1,0,1\}} \exp(f_i^{[Z_h]}(\beta^{[Z_h]}, x(t), z(i \updownarrow_h \tau)(t)))}. \tag{3.2}$$

Here $z(i \updownarrow_h \delta)$ stands for the behavioural configuration that results from a micro step in which actor i changes the score on the behavioural variable Z_h by δ. So $z(i \updownarrow_h \delta)_{hi} = z_{hi} + \delta$, while all other elements of $z(i \updownarrow_h \delta)$ are equal to those of z [7]. Here the same interpretation as for the network change is possible; the direction of change is towards the maximum of the changing probabilities.

Now the objective function for the behaviour $f_i^{[Z]}(\beta^{[Z_h]}, y)$ can be specified in a similar way as for the network changes is done:

$$f_i^{[Z_h]}(\beta^{[Z_h]}, y) = \sum_k \beta_k^{[Z_h]} s_{ik}^{[Z_h]}(y).$$

Here $s_{ik}^{[Z_h]}(y)$ are again effects which will be specified later on. These effects for the behaviour will be discussed in section 3.5.4, together with the network effects.

3.5.3 Time heterogeneity

In the previous two sections the network- and behaviour-objective functions were discussed. Here it is assumed that the parameters ($\beta_k^{[X]}$ and $\beta_k^{[Z_h]}$) are equal for all $M-1$ time periods between t_1 and t_M. However, it is possible that the values of $\beta_k^{[X]}$ and $\beta_k^{[Z_h]}$ are not equal for the different periods. In order to take this into account a dummy variable $\delta \in \{-1, 0, 1\}$ can be added [5]. In that case the objective functions will look like this

$$f_i^{[X]}(\beta^{[X]}, y) = \sum_k (\beta_k^{[X]} + \delta_{k,m}^{[X]}) s_{ik}^{[X]}(y),$$

$$f_i^{[Z_h]}(\beta^{[Z_h]}, y) = \sum_k (\beta_k^{[Z_h]} + \delta_{k,m}^{[Z_h]}) s_{ik}^{[Z_h]}(y).$$

Here are $\delta_{k,m}^{[X]}$ and $\delta_{k,m}^{[Z_h]}$ the dummy variables for the effect k in time period m for the network respectively behaviour objective function.

3.5.4 Effects

As already mentioned in the above sections, the objective functions consist of a sum of effects, but what these effects are, is not specified so far. In this section these effects will be explained a little bit more. First the network effects will be discussed, followed by the covariate effects and in the end the behaviour effects will be explained.

Network effects

There are many possible network effects for actor i. These effects model some structural figures in the network. There are many effects for the network that can be used. Below some of them (the most commonly used ones) will be discussed.

- **Out-degree effect**
  This effect controls the density of the network, the average degree. It measures the overall tendency to form ties. This can be understood as the formation of a tie, as can be seen in figure 3.3a. Mathematically the out-degree effect is denoted by

  $$s_{i1}(x) = x_{i+} = \sum_j x_{ij}.$$

  If the value of $\beta_1^{[X]}$ would be 0 (and there will be no other effects present), the total degree of the network will be 50%, as follows from $\frac{e^\beta}{1+e^\beta} = \frac{1}{1+1} = 0.5$. For social networks mostly less than 50% of all possible ties will be present. Social networks are often more sparse, and therefore they have often a value of $\beta_1^{[X]}$ (that belongs to $s_{i1}(x)$) that is negative.

- **Reciprocity effect**
  This effect measures the number of reciprocated ties. That means that it measures the tendency to change a tie that is already there into a reciprocated one. Schematically this is shown in figure 3.3b. Mathematically this effect is indicated by

$$s_{i2}(x) = \sum_j x_{ij} x_{ji}.$$

  If the value of $\beta_2^{[X]}$ is positive, this means that there is a tendency to form reciprocated ties. In friendship networks it is often preferred to have a friendship tie that is reciprocated. Therefore in most friendship networks a positive β-value for the reciprocity effect will be found.

- **Transitive triplets effect**
  This measures the number of transitive triplets. A transitive triplet looks like the right part of figure 3.3c, so it consists of this pattern: $(i \to j, i \to h, h \to j)$. As schematically shown in figure 3.3c the effect includes the tendency to form transitive triplets. Mathematically the transitive triplet effect can be included as

$$s_{i3}(x) = \sum_{j,i} x_{ij} x_{ih} x_{hj}.$$

  It is described in literature, for example in [6] and [9], that triangular structures are important in friendship networks. A lot of friendship networks have a positive value for $\beta_3^{[X]}$, what indicates that the formation of transitive triplets is favourable in the networks. A negative value for $\beta_3^{[X]}$ would indicate that the formation of these transitive triplets is not favourable. The transitive triplets induce a kind of hierarchy to the network. This can be seen by the fact that actor A is friends with actor B and actor B is also friends with actor C, then actor A has the tendency to see actor C as friend, but actor C sees actor A not as his friend.

- **Transitive ties**
  This effect measures the triadic closure of the neighbourhood. That means that it measures if there is a drive to become friends with the actors that you are indirectly connected to. Indirectly connected means that there is at least one intermediary to whom i is connected to and this intermediary is connected to j, but the connection is not direct: $x_{ih} = x_{hj} = 1$ and $x_{ij} = 0$. The transitive ties effect measures if this situation is avoided by becoming friends with your indirect neighbourhood, as is shown schematically in figure 3.3d. This effect can be included by

$$s_{i4}(x) = \sum_j x_{ij} \max_h(x_{ih} x_{hj}).$$

  This effect also measures the closure of the network. If the value of $\beta_4^{[X]}$ is positive the network closure takes place and the actors become friends with their neighbourhood. If the value of $\beta_4^{[X]}$ is negative, the network tends not to close and the indirect neighbours stay indirect neighbours, there is no drive to become friends with them.

- **Three-cycle effect**
  This indicates the number of three-cycles in i's ties. A three-cycle looks like the right part of figure 3.3e, it has this connection structure: $(i \to j, j \to h, h \to i)$. This effect measures the tendency to form three-cycles. The schematically picture of this tendency can be found in figure 3.3e. The three-cycle effect can be calculated in this way:
  $$s_{i5}(x) = \sum_{j,h} x_{ij} x_{jh} x_{hi}.$$
  This represents the absence of hierarchy and it is a kind of generalized reciprocity. In literature (for example in [6]) it is described that the formation of three-cycles is not favourable in friendship networks. Therefore for friendship networks mostly a negative value of $\beta_5^{[X]}$ is found, what indicates that three-cycles are not likely.

For the model fitting the out-degree effect is always included (likewise a constant term is always included in a regression model). Almost always the reciprocity effect is included. Besides this list, there are many more effects who can be included. Of course also combinations of these effects can be included (for example reciprocity $\times$ transitivity). For each model you can decide which ones are useful to include and which ones are probably not necessary.

Covariate effects

As already mentioned, there may be covariates associated with the actors. Therefore the effects of these covariates can be included in the model. Some of these effects will be discussed below.

- **Covariate-related popularity**
  This covariate effect measures if one covariate value or background is more popular than others. This covariate effect looks at the main effect of the covariates of others and if that influences the choice to become friends with these actors or not. This covariate determines the popularity in the network. Schematically this is shown in figure 3.3f. In the picture it is assumed that the white balls are the actors with a low score on the covariate, the black ones have a high score on the covariate and the grey ones have an arbitrary score. The popularity of the covariates is measured as the sum of the covariates of all i's friends:
  $$s_{i6}(x) = \sum_{j} x_{ij} v_j.$$
  This measurement is related to the covariates of the other actors, so it is also called an 'alter'-effect. If a covariate rises the probability that everybody will be friends with you, this effect will give a positive value of $\beta_6^{[X]}$. A negative value for $\beta_6^{[X]}$ will be found as that covariate makes you less popular.

- **Covariate-related activity**
  Here the covariates of i are measured, so it is also called "ego"-effect. This effect measures if you have more ties with a certain covariate or not. It is schematically shown in figure 3.3g. In order to get a measurement for this effect, the out-degree of i is weighted by the covariate. Therefore this effect is included as
  $$s_{i7}(x) = v_i x_{i+}.$$

A positive value of $\beta_7^{[X]}$ states that this covariate makes that you form more ties, whereas a negative value of $\beta_7^{[X]}$ means that this covariate makes you form less ties.

- **Covariate-related similarity**
  This is a sum of the measurements of the similarity between i and his friends. It measures to what extend the covariate is similar between i and his friends, as is shown in figure 3.3h. This is calculated as

$$s_{i8}(x) = \sum_j x_{ij} \text{sim}(v_i, v_j).$$

  Here $\text{sim}(v_i, v_j)$ is the similarity between v_i and v_j:

$$\text{sim}(v_i, v_j) = 1 - \frac{|v_i - v_j|}{R_V},$$

  where R_V is the range of V. If this effect has a positive coefficient, this means that ties will be formed between actors with a similar covariate value. Relations will then be formed between actors with a similar background, age, etcetera. A negative value for $\beta_8^{[X]}$ will mean that ties are formed between actors with different values for their covariates. If $\beta_8^{[X]} = 0$ this means that this covariate has no effect.

Behaviour effects

For the objective function for the behaviour, also effects might be included. Some of the possible effects for the behaviour will be discussed below. It can be seen that these effects include behaviour values and sometimes also network values are included.

- **Shape: Linear and quadratic**
  The first effect for the behaviour that is always included is the shape effect. This effect will be added in two parts, in a linear and a quadratic effect-term. These effects are defined as

$$s_{i1}^{[Z_h]}(z) = z_{ih},$$
$$s_{i2}^{[Z_h]}(z) = z_{ih}^2.$$

  This effect is important for the fit of the model. When a negative quadratic tendency parameter is found, the model for behaviour is a unimodal preference model. The graph from this effect will then look like figure 3.2. Here it is clearly visible that there is a maximum for one of the values, in this case for the value 2.

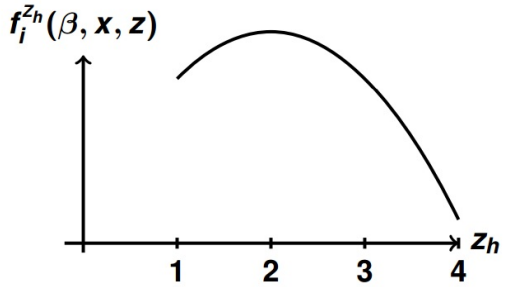


Figure 3.2: Graph of the quadratic tendency parameter with negative coefficient. This ends up in a unimodal preference model. [10]

When the coefficient is positive, the behaviour objective function can be bimodal because a parabola is obtained which has in the domain two maximum values (that can also be seen as positive feedback).

- **Behaviour-related average similarity**
  This effect is defined as the average of behaviour similarities between i and his friends. It measures the assimilation to the neighbours average behaviour. This is shown in figure 3.3i. Now it is assumed that the white balls are the actors with a low score on the behaviour component h, the black ones have a high score on the behaviour component and the grey ones have an arbitrary score. It is calculated as

$$s_{i3}^{[Z]}(x, z) = \frac{1}{x_{i+}} \sum_j x_{ij} \text{sim}(z_{ih}, z_{jh}),$$

  where $\text{sim}(z_{ih}, z_{jh})$ is the similarity in behaviour h between z_i and z_j:

$$\text{sim}(z_{ih}, z_{jh}) = 1 - \frac{|z_{ih} - z_{jh}|}{R_{Z_h}},$$

  where R_{Z_h} is the range of Z_h. If this effect has a positive parameter value, that means that the actors tend to change their behaviour in the direction of the average behaviour of their neighbours. If this parameter is negative, this means that they change their behaviour in away from the behaviour of their neighbours. Of course again when the parameter is zero, this means that this effect is not important in the friendship and behaviour evolution.

- **Popularity-related tendency**
  This is an in-degree effect. It measures if the popularity in the network has an effect on the behaviour. The effect measures if the amount of ingoing ties for actor i has an effect on the behaviour value. This is shown in figure 3.3j. The effect is calculated by

$$s_{i4}^{[Z_h]}(x, z) = z_{ih} x_{+i}.$$

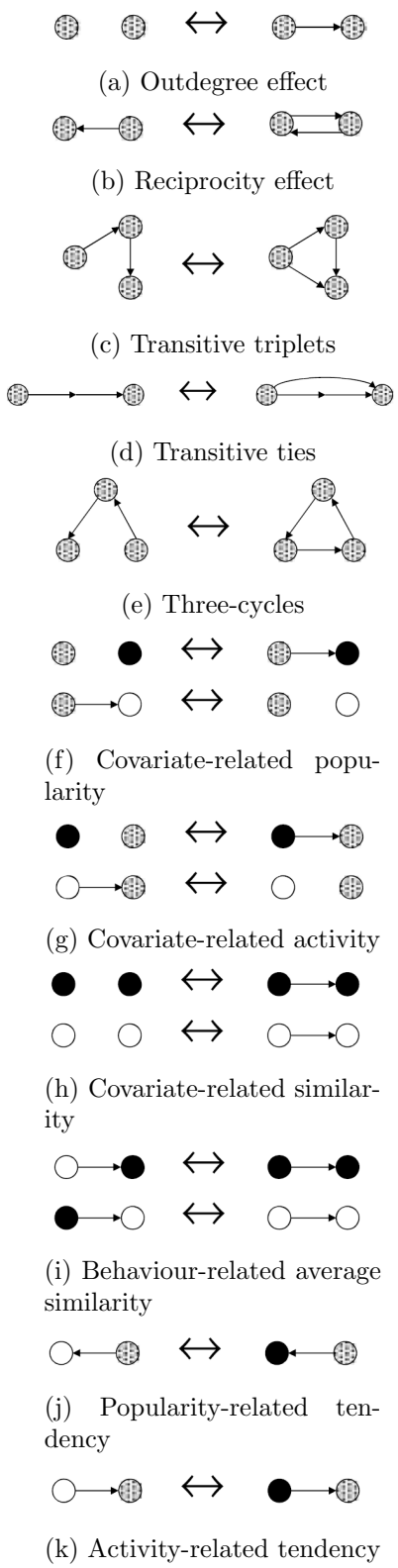
  A positive parameter for this effect means that if you have more ingoing ties, you will have a higher value in the behaviour h. When the parameter is negative, this means that more ties give rise to a lower value in the behaviour. Again if the parameter is zero, there is no relation between the ingoing ties and the behaviour value.

- **Activity-related tendency**
  This is an out-degree effect. It has therefore a similar form as $s_{i4}^{[Z_h]}$, however now the outgoing ties will be included instead of the ingoing ties. The schematic picture will look like figure 3.3k. The effect will be calculated by

  $$s_{i5}^{Z}(x, z) = z_{ih} x_{i+}.$$

  Here the parameter value can be interpreted in the same way as for the popularity-related tendency, only the ingoing ties are replaced by outgoing ties. So this effect also measures the influence of the network structure on the behaviour.

For the network position and the behaviour many more effects are known that can be included in the model. Furthermore, many combinations of effects can be included.



(a) Outdegree effect

(b) Reciprocity effect

(c) Transitive triplets

(d) Transitive ties

(e) Three-cycles

(f) Covariate-related popularity

(g) Covariate-related activity

(h) Covariate-related similarity

(i) Behaviour-related average similarity

(j) Popularity-related tendency

(k) Activity-related tendency

Figure 3.3: Schematic pictures of what tie formation is measured by the different effects

Chapter 4

Parameter estimation and inference

In the previous chapter the theoretical model specification is discussed. Now it is time to come up with parameter estimates. The equations cannot be solved analytically, so in this chapter it will be discussed what algorithms will be used to give the parameter estimates. Furthermore it will be discussed when these estimates are significant.

4.1 Markov process

The total model consists of the first wave of observations that is seen as the initial state of the stochastic process. The rate function defines the rate of changes in the network or behaviour and the objective function defines the choice probabilities for each possible micro step. This process $Y(t)$ can be computer simulated, since it is a continuous-time Markov process. That means that the process satisfies the Markov property, so conditional on the present state, the future and the past are independent. The process can be fully described by its starting value (in this case the first observation $y(t_1)$) and its matrix of transition intensities between the states at any moment t [7]. This matrix of transition intensities can be build in the following way, where $y = (x, z)$ is the current state and y' is the next outcome [7].

$$q(y; y') = \begin{cases} \lambda_i^{[X]}(y)\Pr(x(i \rightsquigarrow j)|x, z) & \text{if } y' = (x(i \rightsquigarrow j), z), \\ \lambda_i^{[Z_h]}(y)\Pr(z(i \updownarrow_h \delta)|x, z) & \text{if } y' = (x, z(i \updownarrow_h \delta)), \\ -\sum_i \Big\{ \sum_{j \neq i} q(y; (x(i \rightsquigarrow j), z)) + \sum_{\delta \in \{-1,1\}} q(y; (x, z(i \updownarrow_h \delta))) \Big\} & \text{if } y' = y, \\ 0 & \text{otherwise} \end{cases}$$

The model is a Markov chain, so for the algorithm we can make use of this fact. The simulation algorithm can be defined by giving the step of a single change in the process [6]. For the starting point a certain configuration of the network and behaviour $(x(t), z(t))$ is taken. Then a waiting time is drawn from the exponential distribution with parameter λ_{total} and the time parameter is incremented by this waiting time (and the process stops when the end of the time period is reached). If the process continues (using the probabilities $\frac{\lambda_i^{[X]}}{\lambda_{total}}$ and $\frac{\lambda_i^{[Z_h]}}{\lambda_{total}}$), it will be determined if the next change is a network change or a behaviour change and which actor makes the change. Therefore the probabilities from the objective functions will be used ((3.1) and (3.2)). This process repeats itself until the end of the period is reached [6]. At the end point, the configuration of the network and

behaviour will be evaluated. However, the model is too complex to have nice closed form solutions for the probabilities and expected values, so it is hard to obtain the parameter estimates via maximum likelihood methods. This approach is discussed in [11], but in this thesis the method of moments will be used to estimate the parameters. This approach will be discussed in the next section.

4.2 Method of moments

For a general statistical model with data Y and parameter θ, the method of moments-estimator is based on a statistic $u(Y) = (u_1, ..., u_K)(Y)$. This statistic is defined by the parameter value $\hat{\theta}$ for which the expected and observed values of $u(Y)$ are the same:

$$E_{\hat{\theta}}(u(Y)) = u(y),$$

where $u(y)$ is the observed value. This equation is called the moment equation [7].

Now we need to define θ for our model and we need to come up with some statistics that are useful in our model in order to apply the method of moments. For θ the parameters of the rate function and the objective function will be included. Therefore we define $\theta = (\lambda_m^{[X]}, \lambda_m^{[Z_h]}, \beta_k^{[X]}, \beta_k^{[Z_h]})$. There is no formal method to obtain the statistics u_k, but the statistics u_k should be chosen so that they are relevant for the components of the parameter θ in the sense that the expected values of u_k are sensitive for changes in the components of θ [12]. A manner to specify this is to require that

$$\frac{\partial \mathrm{E}_\theta u_k}{\partial \theta_k} > 0 \quad \text{for all } k.$$

As was proposed in [7], [12] and [6], the statistics can be build in the following way. For the rate function parameters $\lambda_m^{[X]}$ and $\lambda_m^{[Z_h]}$, the natural statistic for these parameters are

$$u_m(Y(t_{m-1}), Y(t_m)) = \sum_{i,j} |X_{ij}(t_m) - X_{ij}(t_{m-1})| \quad \text{for estimating } \lambda_m^{[X]}$$

$$u_m(Y(t_{m-1}), Y(t_m)) = \sum_{i} |Z_{hi}(t_m) - Z_{hi}(t_{m-1})| \quad \text{for estimating } \lambda_m^{[Z_h]}.$$

For these choices we can see that if $\beta = 0$, the model will be reduced to the trivial situation where $X_{ij}(t)$ and $Z_{hi}(t)$ are randomly changing 0-1 variables [12]. Therefore these are sufficient statistics for $\lambda_m^{[X]}$ and $\lambda_m^{[Z_h]}$.

For $\beta_k^{[X]}$ and $\beta_k^{[Z_h]}$ we would like to have a statistic that gives a high value when $\beta_k^{[X]}$ or $\beta_k^{[Z_h]}$ is high. Furthermore the values of $\beta_k^{[X]}$ or $\beta_k^{[Z_h]}$ are estimated for the whole time range. Therefore a summation over m is included. The expressions that therefore are proposed are (based on [7] and [12])

$$u_k(Y(t)) = \sum_m \sum_i s_{ik}^{[X]}(Y(t_m)) \quad \text{for estimating } \beta_k^{[X]}$$

$$u_k(Y(t)) = \sum_m \sum_i s_{ik}^{[Z_h]}(Y(t_m)) \quad \text{for estimating } \beta_k^{[Z_h]}.$$

For these expressions you can see intuitively why they are proposed. When β_k is larger, the actors have a stronger drive to go for a large value of s_{ik}. Therefore it might be expected that $u_k(t_m)$ is larger for all m.

However, the formulae that are proposed do not distinguish between influence and selection. That means that if we use the statistics that we proposed, we will not be able to say anything about the influence and selection. Therefore the formulae need to be modified. That can be done by including the time. The time order is the basis of causality and therefore by including the time order we will be able to distinguish between influence and selection. For selection an earlier configuration of attributes leads later on to a change in ties, whereas influence can be determined by an earlier configuration of ties that lead to a change of attributes later on [7]. Therefore the statistics that are used in the moment equations are given by

$$u_k(Y(t)) = \sum_m \sum_i s_{ik}^{[X]}(X(t_m), Z(t_{m-1})) \quad \text{for estimating } \beta_k^{[X]}$$

$$u_k(Y(t)) = \sum_m \sum_i s_{ik}^{[Z_h]}(X(t_{m-1}), Z(t_{m-1}), Z(t_m)) \quad \text{for estimating } \beta_k^{[Z_h]}.$$

Here $s_{ik}^{[Z_h]}(X(t_{m-1}), Z(t_{m-1}), Z(t_m))$ is defined by employing for the behavioural variables the value at t_m for Z_h and the value at t_{m-1} for $Z_{h'}$, for all other h' [7]. This can also be expressed as

$$s_{ik}^{[Z_h]}(X(t_{m-1}), Z(t_{m-1}), Z(t_m)) = s_{ik}^{[Z_h]}(X(t_{m-1}), Z^*)$$

with

$$Z_{h'}^* = \begin{cases} Z_{h'}(t_{m-1}) & \text{if } h' = h, \\ Z_{h'}(t_m) & \text{if } h' \neq h. \end{cases}$$

Also when the same components are present in the evaluation function for different behaviour variables, the proposed statistics can be used to separate these effects from each other [7].

Now we defined $\theta = (\lambda_m^{[X]}, \lambda_m^{[Z_h]}, \beta_k^{[X]}, \beta_k^{[Z_h]})$ and we found expressions for the statistics that belong to the model parameters. Therefore the moment equations can be constructed by

$$E_{\hat{\theta}}(u(Y)) = u(y),$$

with the statistics that we found above:

$$u_m(Y(t_{m-1}), Y(t_m)) = \sum_{i,j} |X_{ij}(t_m) - X_{ij}(t_{m-1})| \quad \text{for estimating } \lambda_m^{[X]},$$

$$u_m(Y(t_{m-1}), Y(t_m)) = \sum_i |Z_{hi}(t_m) - Z_{hi}(t_{m-1})| \quad \text{for estimating } \lambda_m^{[Z_h]},$$

$$u_k(Y(t)) = \sum_m \sum_i s_{ik}^{[X]}(X(t_m), Z(t_{m-1})) \quad \text{for estimating } \beta_k^{[X]},$$

$$u_k(Y(t)) = \sum_m \sum_i s_{ik}^{[Z_h]}(X(t_{m-1}), Z(t_{m-1}), Z(t_m)) \quad \text{for estimating } \beta_k^{[Z_h]}.$$

The idea now is to solve these equations. However, since there is no closed form solution, a computer simulation will be needed. This iterative procedure will be discussed in the next section.

4.3 Iterative procedure

In order to obtain the parameter estimates, we need to estimate θ. This will be done in an iterative manner, that starts with $\hat{\theta}_0$ and gives $\hat{\theta}_k$ at step k. At step k the value of U_k^{sim} is determined. This is the simulated statistic at time k and it is determined in the following way. For each $m = 1, ..., M$, the process $Y(t)$ is simulated, starting at time t_m. Letting time run from t_m to t_{m+1} for parameter value $\hat{\theta}_k$, $Y(t)$ is simulated with $Y(t_m) = y(t_m)$, where $y(t_m)$ is the observed value at t_m. The simulated value $\hat{\theta}_k$ that is obtained for time t_m to t_{m+1} is denoted by $Y^{sim}(t_{m+1})$. The corresponding value from the components of U_k^{sim} are denoted by

$$u_k(y(t_m), Y^{sim}(t_{m+1})),$$

where u_k is the statistic that belongs to the corresponding component of $\hat{\theta}_k$ [7]. In a similar way, the observed values of $u(Y)$ (denoted by $u(y)$) are defined as

$$u_k(y(t_m), y(t_{m+1})).$$

Now $\hat{\theta}_{k+1}$ can be obtained by an iteration step. The stochastic approximation methods of Robbins and Monro (1951) can be used to solve the moment equations. The iteration step is then given by

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_{k+1} D_0^{-1} (U_k^{sim} - u(y)).$$

Here is D_0 an approximation of the matrix of partial derivatives of the statistic $u(Y)$ at $\hat{\theta} = \hat{\theta}_0$. a_k is a converging sequence of numbers that approaches zero at rate k^{-c}, where c is chosen from the interval $0.5 < c < 1$, in order to have good convergence properties. This iteration rule defines also a Markov chain, so the algorithm is a Markov chain Monte Carlo algorithm [7].

The final estimate $\hat{\theta}$ is then obtained by the tail average of the trial values

$$\hat{\theta} = \frac{1}{R} \sum_{r=1}^{R} \hat{\theta}_{r_0 + r}.$$

Here is r_0 the number of iterations after which there are only small changes in $\hat{\theta}$ and R is large enough to obtain a stable estimator.

4.4 Test of significance

As also described in [12], [6] and [7], the approximate covariance of $\hat{\theta}$ can be obtained using the delta method and the implicit function theorem and is given by

$$\mathrm{Cov}(\hat{\theta}) \approx D_\theta^{-1} \Sigma_\theta D_\theta'.$$

Here D_θ is the matrix of partial derivatives,

$$D_\theta = \left(\frac{\partial E_\theta(u(Y))}{\partial \theta} \right),$$

and Σ_θ is the covariance matrix of the statistic $u(Y)$

$$\Sigma_\theta = \mathrm{Cov}_\theta(u(Y)).$$

This shows that the efficiency of the method of moments-estimation depends on the statistic $u(Y)$ that was chosen. The covariance matrix Σ_θ and the matrix of partial derivatives D_θ can be obtained by Monte Carlo methods [7].

The covariance matrix of $\hat{\theta}$ can be used to obtain the standard errors of the estimators. At the diagonals of the covariance matrix, the variance of $\hat{\theta}$ can be found. If we take the square root of these diagonal elements, the standard error is obtained.

In [12] it is noted that these estimators have approximately normal distributions (with $\mu = 0$ and $\sigma =$ standard error) and therefore the parameter estimates can be divided by the standard errors in order to do a t-test. A t-test with infinite degrees of freedom will be done. This means that we have the hypotheses

$$H_0 : \quad \theta = 0$$

$$H_1 : \quad \theta \neq 0.$$

For the t_∞-test, with a significance level of 0.05, the null hypothesis can be rejected when

$$\left| \frac{\hat{\theta}}{\text{standard error}} \right| > 1.96.$$

When we reject the null hypothesis, this means that we take the alternative hypothesis, so we know that $\hat{\theta} \neq 0$. Therefore we can then say that $\hat{\theta}$ is significant (with significance level 0.05) when $\left| \frac{\hat{\theta}}{\text{standard error}} \right| > 1.96$.

Chapter 5

Data for the study

5.1 Requirements of the data

In this section the requirements of the data, in order to use the stochastic actor based model, will be discussed. These requirements consist of requirements for the number of observations and the number of actors, as well as requirements on the changes between the measurements (expressed as the Jaccard index) and the number of missing data points. These requirements will be explained in the sections below. The data that is obtained for this thesis will also be tested for these requirements.

5.1.1 Number of measurements and number of actors

In order to fit the stochastic actor-based model, there need to be at least two measurements. So the number of measurements M needs to be larger than or equal to two. Furthermore the number of measurements is mostly smaller than 10. Models with more than 10 measurements can be fitted, but that needs to be done very carefully, see [9] for more information about this. Our model has three measurements, with 3 months time between each measurement. So $M = 3$ for our study, thus the stochastic actor-based model can be applied.

As [9] also states, the number of actors, N, needs to be larger than 20. On the other hand, the number of actors cannot be larger than a few hundred. If there would be more than a few hundred actors, the assumption that in principle all actors are network partners for all other actors does not hold any more. All actors can in principle meet each other and become friends, but if there are too many actors, this assumption is not true any more. In our study the actors are the first year mathematics students. In the beginning this are 66 actors. This is a nice number of actors for the application of the stochastic actor-based model.

5.1.2 Jaccard index

The total number of changes in the network between the different measurements should be large enough. Normally under 40 changes is not enough, but of course this also depends on the sample size [9]. On the other hand, the number of changes should also be not too high. In that case it is not known any more if the changes that are measured indicate the right changes in the measurement period. If the number of changes is too large, the time between the measurements is probably too large.

The difference between two networks can be measured by the Jaccard index. This index is given by

$$J = \frac{N_{11}}{N_{11} + N_{01} + N_{10}}.$$

Here N_{11} indicates the number of ties that is present in both measurements, N_{01} is the number of ties that is newly formed and N_{10} is the number of ties that is not present any more in the second measurement. As [9] states, from earlier studies it is clear that a $J > 0.3$ indicates a useful data set.

For our data the Jaccard index for the network between measurement 1 and 2 is $J_{12} = 0.559$. Between measurement 2 and 3 the Jaccard index is $J_{23} = 0.434$. Therefore our data set has enough changes between the measurements in order to use the stochastic actor-based model.

5.1.3 Missing data

In practice often some of the data is missing. For example some of the actors are absent during the measurement, some of the actors don't want to participate in the research, some actors forget to complete a question and so on. In [13] it is stated that when more than 20% of the data is missing, the model fitting is very difficult. Furthermore it is stated that for the missing data "NA" is added to the data that is given to R. Then in RSiena it is programmed how to handle the missing data, because the simulations will be run as if all data would be available. In order to do so, the missing spots need to get a number. For the network-part, if the missing spots were in the first measurement, a 0 will be given to these spots. This is done, because it is more likely that a tie is not present than that a tie is present. This is a result of the fact that friendship networks are often sparse. When the data is missing in later measurements, the same value as at the measurement before is used. If this data is not available, a 0 is assigned. For the behaviour-part, when the dependent behaviour value is missing, it will get the value of the measurement before. For the first measurement the modus of the values for the behaviour variable at measurement 1 will be used. When data is missing for the covariates, this will be replaced by the average of the covariate. For the calculation of the estimators, the missing data will not be included.

For our data set the number of missing data is quite large. For the first measurement the fraction of missing data is 0.288, for the second measurement the missing fraction is 0.364 and for the third measurement the missing fraction is 0.379. These fractions are way larger than the 20% that is maximum for the model to give good estimates. However, we think that for this data the model can give reliable estimations, the missing data estimate is higher than the actual missing data. This is because during the academic year some of the students stop with the study, because they don't like it or because they don't have the right level. These students will not show up any more at the measurements, but they still can be nominated by other students. It is for us not clear which students quit, but in principle these students leave our research group. However, it is not possible to eliminate these students, because it is not known which students did stop. These values are added as missing values, but in principle these are not missing values. Furthermore, there are 16 students who only completed one of the measurements. All measurements were executed during compulsory courses/meetings, so it doesn't seem logical that the students missed two of these meetings. Therefore they probably have stopped their studies or they are not

following all the first year courses. So the fraction of missing data obtained from R is a maximum estimate of the total amount of missing data.

5.1.4 Network autocorrelation

In order to measure if relation partners have indeed a more similar behaviour than actors who have no connection, the network autocorrelation can be calculated. This measures if the behaviour similarity is higher for relation partners than for randomly chosen partners. There are two widely used statistics that measure the network autocorrelation, Moran's I and Geary's c. These two coefficients measure slightly different aspects of the association between behavioural homogeneity and presence versus absence of a relational tie [6], therefore it is good to calculate them both. The I-coefficient is based on cross-products of behavioural scores of relational partners. For Moran's I values between -1 and 1 are possible. Values close to zero indicate that relational partners are not more similar than one would expect under random pairing, while values close to one indicate a very strong network autocorrelation and values close to -1 indicate a negative autocorrelation [6]. Moran's I is defined as

$$I = \frac{n \sum_{ij} x_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\left(\sum_{ij} x_{ij}\right)\left(\sum_i (z_i - \bar{z})^2\right)}.$$

Geary's c is based on squared differences of the behavioural variable between two partners. For Geary's c values between 0 and 2 are possible. Values close to one indicate that random pairing gives almost similar results as the network gives, whereas values close to zero indicate a strong network autocorrelation and values close to 2 indicate a strong negative autocorrelation. The coefficient is calculated by

$$c = \frac{(n-1) \sum_{ij} x_{ij}(z_i - z_j)^2}{2\left(\sum_{ij} x_{ij}\right)\left(\sum_i (z_i - \bar{z})^2\right)}.$$

For our network the correlations between the network and the behaviour (in our case the average grades) are calculated. For the first measurement $I = -0.003$ and $c = 0.86$, for the second measurement $I = -0.03$ and $c = 1.05$, and for the third measurement $I = -0.02$ and $c = 0.90$. From these values only the $c = 0.85$ and $c = 0.90$ indicate somewhat network autocorrelation, but all other measurements show that there is barely no network autocorrelation. In order to measure the network autocorrelation all people that did not complete the questionnaire or that did not include their grades in the questionnaire were removed. Therefore only a sub-network was left to calculate the autocorrelation from. People who nominated people who did not complete the questionnaire therefore have less ties left than they indicated in the questionnaire. Therefore these values for the network autocorrelation might not be totally correct. However, it seems that for university students there is not so much drive to become friends with people with almost equal grades as there is at high school.

5.2 Observed data and particularities

The measurements were performed at the end of November, at the end of February and at the beginning of June of the academic year 2014-2015. The first measurement took place during the mentor meetings. The second questionnaire was distributed after a compulsory computer practical and the third measurement was handed out during a compulsory course (propaedeutic project). On the list that was obtained in September from the study advisor were 77 names. This were the names of all people that were registered in September. Probably some of them didn't start the study at all. That also turned out to be the case. There are 11 people who did not complete the questionnaire and who were also never nominated by someone else. Therefore in the preparation of the data these twelve people were removed. This was done by first adding all information to R, so that matrices and vectors were formed. Afterwards the rows and columns that belong to these twelve people were removed. The numbers assigned to the actors that remained therefore changed. This is no problem, as long as this approach is executed in a consistent manner. There are also six students who don't want to participate themselves in the research, but they are nominated. These students are kept inside the list, because these students are really part of the group under research.

In total there are 16 students who participated only once in the research. These students might have stopped their studies after their participation in the research. It is normal that some of the students drop off before February because the study is not what they hoped. There are some questions in the questionnaire that try to get a feeling for the students who quit. A small analysis with this data is performed in section 5.4.7. It is also possible that some of these students participated once and the questions were too sensitive or it took too long so that they don't want to participate in the next data wave. In order to avoid a lot of stoppers in the research I removed the question about the people that they don't like and about the popular persons. After the first measurement I received a lot of comments that these two questions were not appropriate. There were also (almost) no students who nominated a person for these questions. Furthermore, there are some students who did not complete all questions. For the non-completed questions NA is added to R. During the three measurements there were in total three different students who were not on the list. They completed the questionnaire and their names were added to the list. However, they were not present at other meetings. It seemed that these students were only present at the courses where the second or third measurement was held, but that they are not really first year students. Therefore I did not include these students in the study.

Furthermore, after the first measurement I removed some of the questions that do not change over time (like the performance at high school or the nationality), in order not to annoy the students and to make the questionnaire shorter. However, it turned out that some of the students did not show up at the first measurement, but they did complete the second questionnaire. Therefore for these students there is no background information available. This is the reason that I added all questions again for the last measurement. There were some students who only completed the second measurement. From these students the background is not available. For the other students the background is known (except for the six students who don't want to participate). What is also quite remarkable is that for some students a difference is observed for the answers to background questions between the first and the third measurement. The difference is observed in the question about the reasons to study in Groningen and in the question about the level at high

school. Most of the students who show this difference (6 students) rate themselves in the third measurement lower than in the first measurement with respect to the grades in high school. The reason to choose for Groningen also yielded different answers. The fact that it is the closest university is not named that often any more and the nice city and the international character is named more often. I used the answers that the students gave at the first measurement. For another kind of study it might be interesting to have a look how the perception of the students on questions like this changes over time.

5.3 Data preparation

In order to fit the model, the data that is obtained from the questionnaires needs to be processed. How this is done for the different variables is explained in this section. First the variables for the analysis will be discussed and later on also the control variables will be studied.

5.3.1 Variables for the analysis

In this section the variables that were mentioned in section 2.2 will be extracted from the questionnaires in order to do the analysis.

Grades
The students were asked about their grades in the last exam week. Mostly they completed three exams. Some of the students are only taking two courses or some of them do a resit. In this research the average grade is calculated from these values. This is done because in this way we have one number that indicates the academic performance of the students. Moreover the problem that not all students took the same amount of courses is solved. Of course it is debatable if this is the right way to go, because a student who passes four courses has a better performance than a student who passes only two courses. However, over the whole line it will be a good measurement for the academic performance to take the average grade. In the questionnaire some of the students only indicated pass or not pass. I set the pass equal to 6 and not pass equal to 5. This might not be perfect, but it are only 5 students who did this and mostly the people who think in pass-non pass indeed go for a 6 in order to pass the course. The grades are rounded to halves, because the grades will be the behaviour variable. The behaviour variable needs to be discrete. The distribution of the grades is given in figure 5.1.

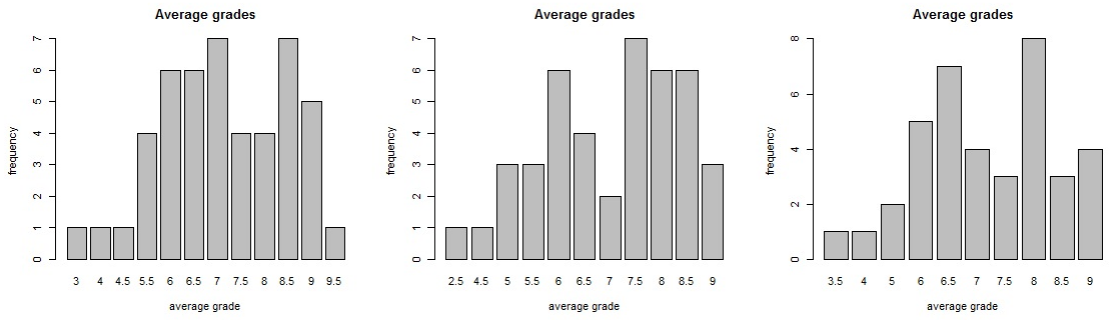


Figure 5.1: The distribution of the average grades

What is interesting to see in these diagrams, is that there kind of are two classes. There are many people who have a grade around a 6 and there are many people who have a grade around an 8.

Nationality
The students were asked what their nationality is. In total there are 12 different nationalities present. There are four British students, seven students with all a different nationality and the rest of the students are Dutch. It might be argued that the British people form a cluster, but we think that 4 persons are not enough to reserve a whole category for them. Furthermore they speak British, so that language will be understood by everybody and therefore no others are excluded from the possible friendship cluster of the British students. In the whole population there are no other people who will communicate in a different language than English with each other, apart from the Dutch students. That is why we decided to make two categories: 0=Dutch, 1=international. With this classification we can find out if the international students mix with the Dutch students. There are 42 Dutch students, 14 international students and from 10 students we have no information about the nationality.

Attitude towards the international bachelor
In order to know a little bit more about the opinion about the international bachelor, the students were asked to agree or disagree to some statements. We can combine the answers to the statements "I think the international character is an advantage to the study", "I find it hard that everything is in English" and "If it was possible to do a Dutch bachelor mathematics in Groningen, I would prefer that over the international bachelor". Here the students have to agree or disagree in a five fold scale (from strongly disagree to strongly agree). For the scoring of this question for the first statement 5 points were given for strongly agree, 4 points for agree and so on. For the other two statements 5 points were given for strongly disagree, 4 points for disagree and so on. The scores for these three statements were summed in order to get one variable that indicates the opinion about the international bachelor. The frequencies are given in figure 5.2.

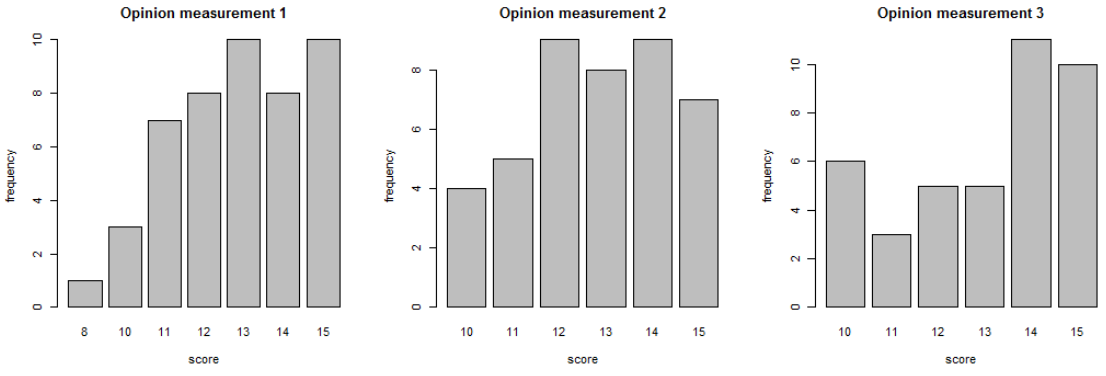


Figure 5.2: Frequency of the scores at statements about the different opinions towards the international bachelor

From this figure it can be seen that there is only one student with a score of 8, and all others have a score of 10 or higher. That means that all statements were rated at or above 3. Therefore all students are positive towards the international bachelor. This makes it

useless to include this variable in the model, because all students are positive.

However we can also take a look at the behaviour in the international bachelor. For this variable a similar approach as for the opinion about the international bachelor was used. Now the statements "I speak English during tutorials" and "I speak English during breaks" are combined in order to get an idea about how the language use is during a study day. Here 5 points were given for all of the time, 4 points for most of the time and so on. These statements are summed. The frequencies of the different scores are shown in figure 5.3.

Figure 5.3: Frequency of the scores at statements about the behaviour towards English speaking with respect to the international bachelor

Here is a lot more variation than in the opinion about the international bachelor. Therefore it seems more useful to include this statistic in the model fitting in stead of the opinion about the international bachelor. The bar with the score 10 is quite large, but this is partly because most of the international students will always talk English, because they have no other option. We can try to investigate if people who always speak English have more international friends.

Presence

The students were asked about the hours of lectures and tutorials that they attend. This factor can influence the friendship formation because the more serious students might pick more serious students to become friends. Also the students who are most present at lectures and tutorials, have a good chance to meet others and to become friends. The hours that the students attend the lectures and the hours that the students attend the tutorials are summed. This then gives the variable that we call "presence" to indicate how many lectures and tutorials the students attend. The distributions of this variable at the different measurements are shown in figure 5.4.

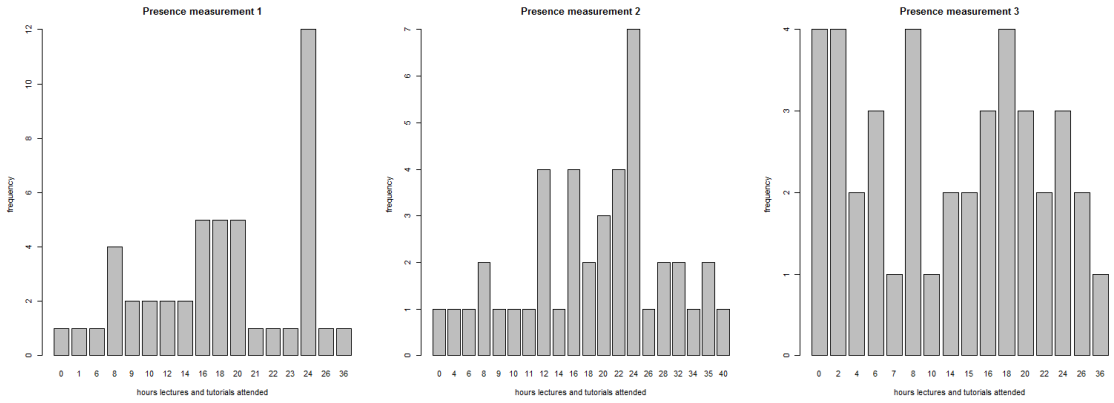


Figure 5.4: The distributions of hours that students attend the lectures and tutorials

In the first measurement there are many students who attend 24 hours of lectures and tutorials. This might be because they don't exactly know and then for both lectures and tutorials estimate to be there 12 hours. The other possibility is that the students in the beginning are mostly present and later in the academic year they know better which lectures they want to attend and which things they will study at home. For the last measurement the distribution is quite diverse.

5.3.2 Control variables

For the control variables also the data needs to be obtained from the questionnaires. How this is done, is discussed in this section.

Gender
The students were asked what their gender is. This can be either male or female, so a coding variable with two categories is logical. There are 44 males and 12 females (and 10 students from whom the gender is unknown) in our population.

Living at home/living away
In the questionnaire the students were asked what their living situation is. In order to answer this question they can pick one of the following answers: 1="I live in a student house with one or more other students", 2="I live with my parents", 3="I live alone in an studio/apartment", 4="I live together with my partner", 5="Other". The distribution of the answers is shown in figure 5.5.

Figure 5.5: The distribution of the living situation of the students

From these figures it is clear that there are only a few people who answered that they live together with a partner or else. Therefore it is decided to leave these answers out and then make a binary variable, consisting of living at home and living away. This gives the distribution as is given in table 5.1 for the three measurements.

Measurement	Living at home	Living away	No information available
Measurement 1	18	26	22
Measurement 2	15	26	25
Measurement 2	9	30	27

Table 5.1: The living situations of the students

From this table we see that there are people who move out during the academic year. This variable is included to control for the living situation that probably has an influence on the friendship formation of the students.

5.4 Exploratory data analysis

In this section we will have a closer look at the data, for example the friendship nominations will be studied. Furthermore we will search for outliers and clusters.

5.4.1 Network part: friendship nominations

In this section the changes in the nominations are discussed. How many friendship nominations there are at the different measurement moments can be seen from table 5.2. This table also shows the average number of friendship nominations that the actors do. This average is only calculated on the available data, so it gives a good estimate for the real average number of ties in the network.

Nominations	Moment 1	Moment 2	Moment 3
Total number of nominations	293	242	199
Average number of nominations	6.234	5.762	4.854

Table 5.2: The number of nominations in the friendship network

The average number of nominations in our study is lower than the average number of nominations that is found in the study at high school [5]. This can be explained by the difference between university and high school. At high school students normally have mostly friends from high school and sometimes some people from sports or some neighbours. At university people tend to have more friends outside the study environment (for example people from high school, from student associations, from sports, house mates and so on). Moreover the attendance at the university mostly is not compulsory, so not everybody is present during all study activities. Therefore not everybody has many contacts with fellow students. Furthermore we see that the average degree of nominations is going down. This might be because some students stop studying and the students who stay at the study have less friends left. It can also be that the students get to know each other better and become real friends with some people and contacts with others are broken.

The changes of the ties can be seen in table 5.3. Here $1 \Rightarrow 2$ indicates the change between measurement 1 and 2, whereas $2 \Rightarrow 3$ indicates the change between measurement 2 and 3. The number of ties that stay absent are indicated by $0 \Rightarrow 0$, $0 \Rightarrow 1$ indicates the number of newly formed ties, $1 \Rightarrow 0$ indicates the ties that were broken, and $1 \Rightarrow 1$ indicates the ties that stay present.

Changes	**0⇒0**	**0⇒1**	**1⇒0**	**1⇒1**
$1 \Rightarrow 2$	1866	66	57	156
$2 \Rightarrow 3$	1789	39	65	122

Table 5.3: The changes in the friendship network

The distributions of the nominations are shown in figure 5.6.

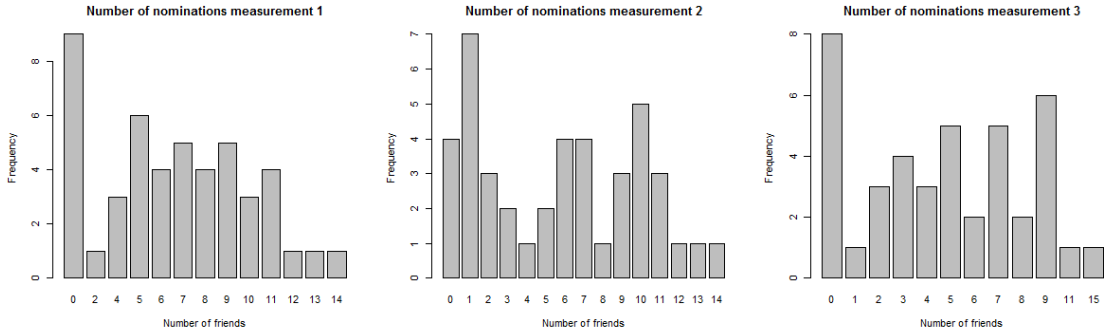


Figure 5.6: The distribution of the friendship nominations

What is notable from these distributions is the fact that there are quite a lot of people who nominate nobody. Some of these people added to the questionnaire that they prefer to sit and work alone for studies. For others it might be the case that they are only present at lectures when they think it is necessary to pass the exam. These students have almost no interaction with fellow students, they have friends outside the study environment. This is a big difference compared to the study at high school [5]. Of course there is missing data and some of the students probably stopped their study, so these plots might not give the whole picture. However it seems that the network consists of quite some people who have no friends at the study. It also might be the case that some of these students were too lazy to mention their friends at the questionnaire. For the rest there are no strange numbers, for example where people nominate around 30 people to be their friends or so. This indicates that people understood the question.

Similar analyses can be done for the network of the neighbours during lectures or the best friends network. These networks can be studied later on, in this thesis we focus on the friendship network.

5.4.2 Behaviour part: academic performance

The academic performance is seen as the behaviour variable. We have already seen what the distribution of the average grades is. Now we can have a look at how this variable is distributed over time for all actors. This results in the graphs of figure 5.7. Here the not

rounded numbers are used, so that not all actors have a spot at the exact same location. All actors are spread over 6 plots, otherwise the individual actors cannot be distinguished because the graph is too crowded.

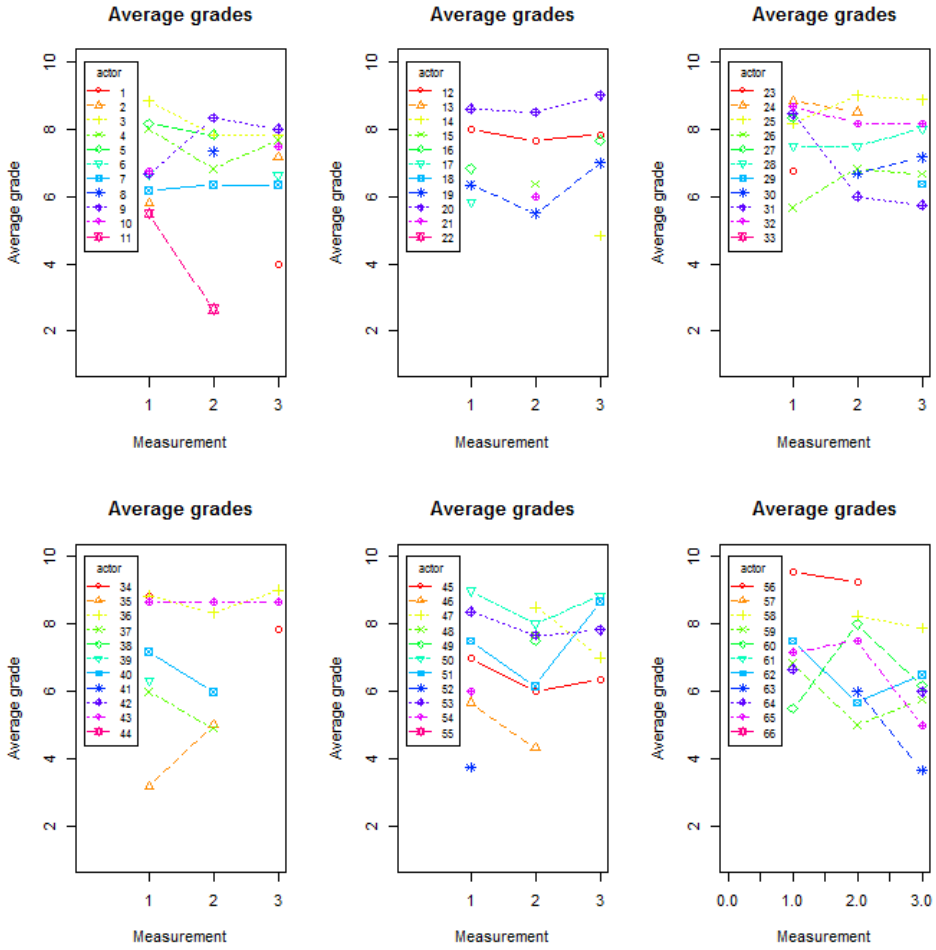


Figure 5.7: The evolution of the average grades per actor

From these figures we see that some of the actors have grades that form a parallel line. These actors might be friends that stimulate each other to go to the lectures. Furthermore the people with really low grades at measurement 1 and/or 2 that have no completed questionnaire for measurement 3 might have stopped their studies. We don't see any outliers or values that cannot be true.

5.4.3 Covariates: Presence

For the covariates we can also look at the development over time, in order to spot outliers and to get a better feeling about the data. The development of the presence over time is shown in figure 5.8.

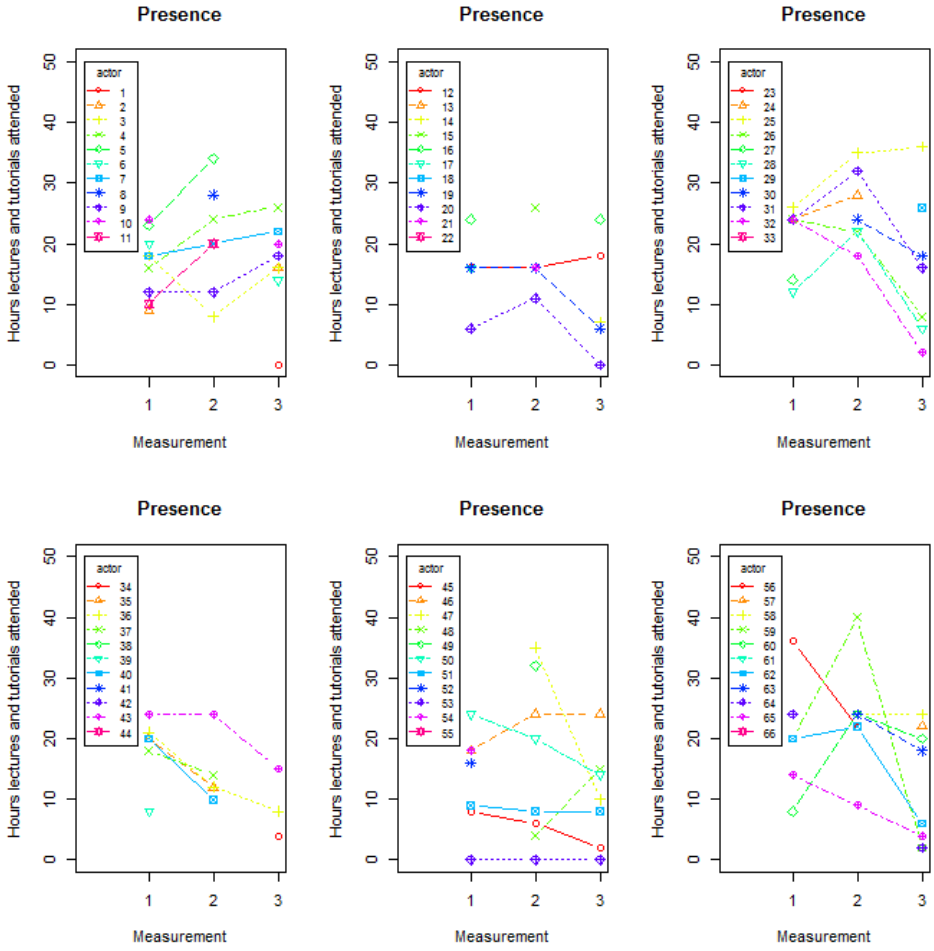


Figure 5.8: The evolution of the presence at lectures and tutorials per actor

It is clear that there is a broad range in hours that the different students attend the lectures and tutorials. We see that there are two students who attend around 40 hours. This is quite a lot. However, after checking the schedule we concluded that it is possible that the students indeed attend around 36-40 hours in the second measurement period. Therefore there are no outliers that need to be removed. It can be seen that for example the lines of attending of the actors 19, 20, 32, 26 and 28 look similar. It might be the case that these people stimulate each other to go or not to go to the lectures and tutorials.

5.4.4 Covariates influencing network: clusters

In this section we will see if the students with similar values for the covariates form clusters in the network. The networks will be coloured for the different covariates.

Nationality

In order to make the possible clusters visible, the Dutch students will be marked with orange, the international students will be marked with blue and the students for whom no nationality is available will be marked with grey. The different networks for the three measurements will then look like figure 5.9, 5.10 and 5.11.

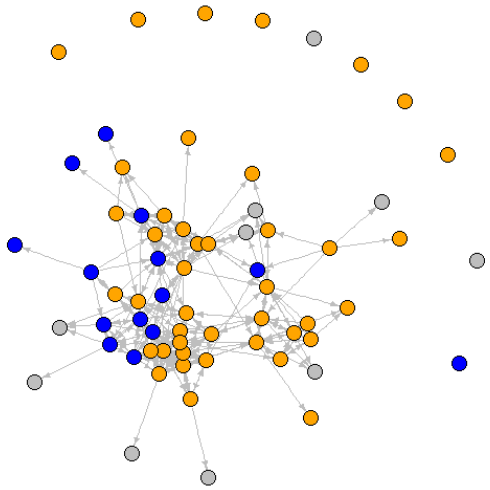


Figure 5.9: The network for the first measurement with nationality highlighted (orange=Dutch, blue=international, grey=no information available)

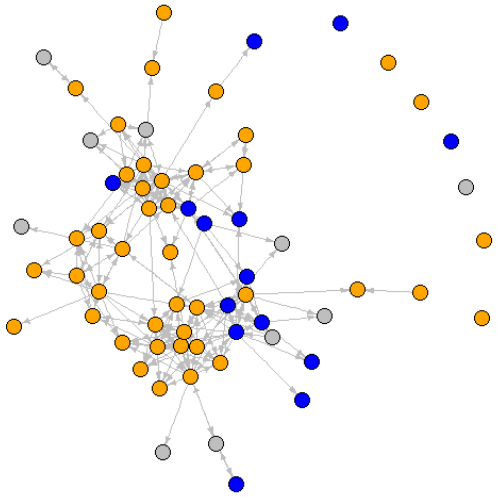


Figure 5.10: The network for the second measurement with nationality highlighted (orange=Dutch, blue=international, grey=no information available)

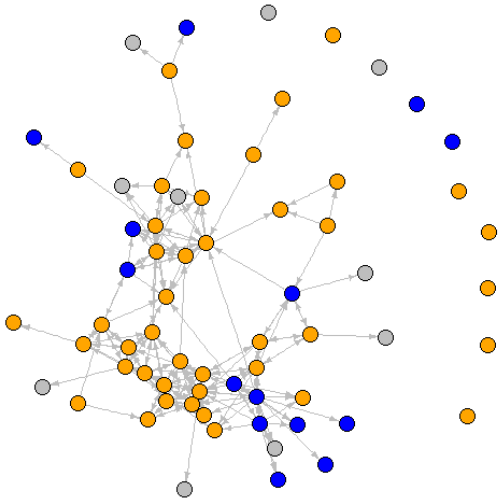


Figure 5.11: The network for the third measurement with nationality highlighted (orange=Dutch, blue=international, grey=no information available)

It looks like there is some cluster formation in the first measurement between some of the international students. However, in the second and third measurement these clusters are not visible any more. It seems that the international students mix quite well with the Dutch students.

Gender

In order to get an idea of the cluster formation of the male and female students, similar plots as for the nationality are generated. Now the males are blue and the females are pink. The students from whom no information available is, are indicated by grey dots. The networks for the three measurements look like figure 5.12, 5.13 and 5.14.

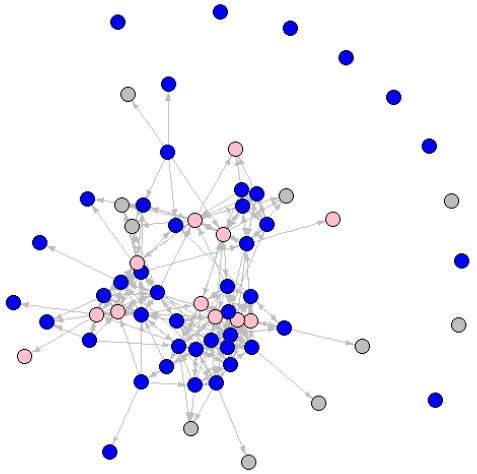


Figure 5.12: The network for the first measurement with gender highlighted (blue=male, pink=female, grey=no information available)

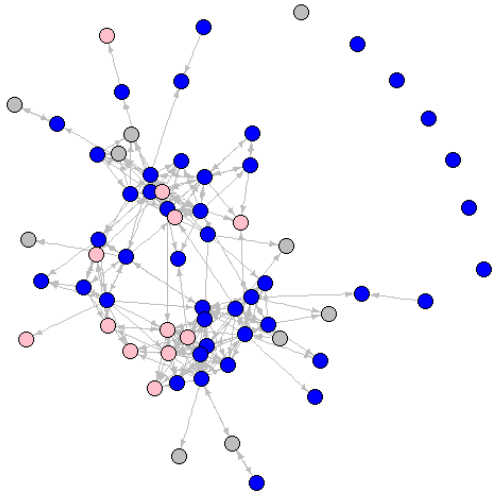


Figure 5.13: The network for the second measurement with gender highlighted (blue=male, pink=female, grey=no information available)

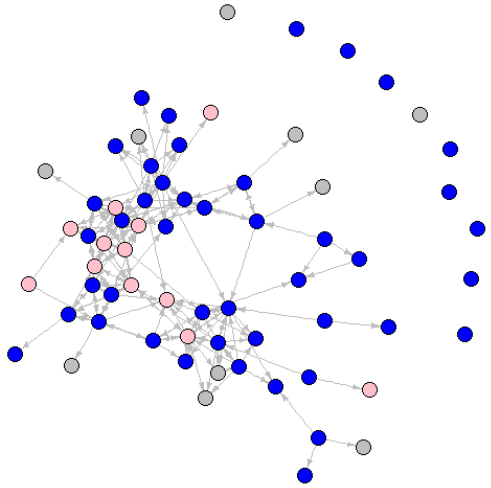


Figure 5.14: The network for the third measurement with gender highlighted (blue=male, pink=female, grey=no information available)

There seems to be somewhat cluster formation between the boys and girls. Therefore it is good to use "gender" indeed as a control variable. It seems that the gender explains a part of the possible cluster formation in the network.

Presence

In order to get insight in the possible cluster formation between the people who are frequently present at lectures and tutorials and people who are not frequently present, plots with different colours will be generated. In figure 5.4 it is seen that there is a group of students who is less than 14 hours present and there is a group of students who is 14 hours or more present. Therefore we can make a plot where we indicate the people who are less than 14 hours present in red, the people who are 14 hours or more present are blue and the people from whom is no information available are grey. This results in the networks

as shown in figure 5.15, 5.16 and 5.17.

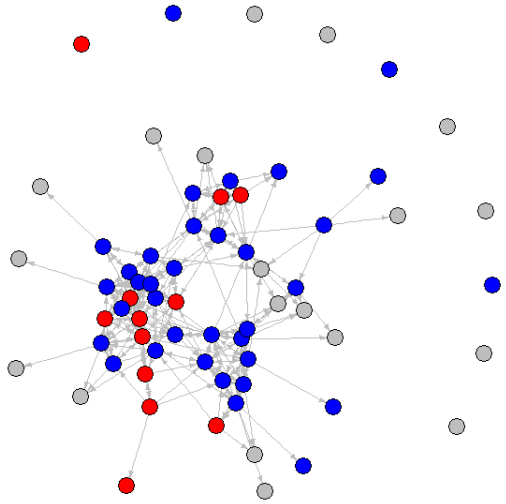


Figure 5.15: The network for the first measurement with presence highlighted (red=less than 14 hours present, blue=14 hours or more present, grey=no information available)

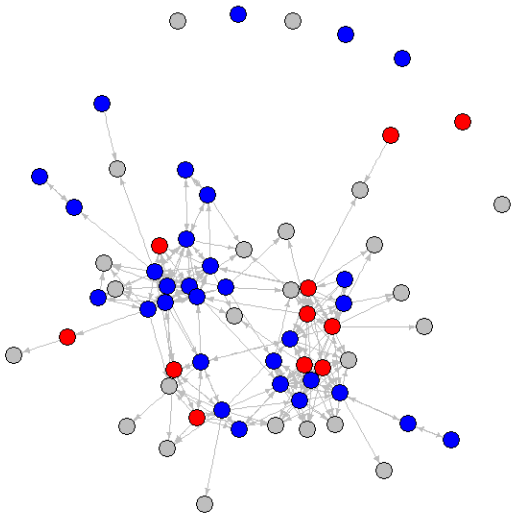


Figure 5.16: The network for the second measurement with presence highlighted (red=less than 14 hours present, blue=14 hours or more present, grey=no information available)

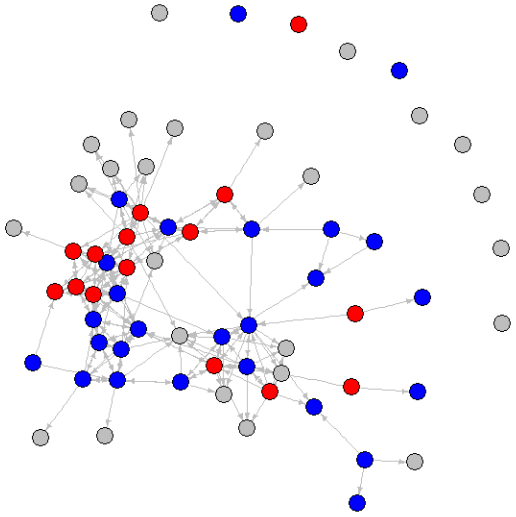


Figure 5.17: The network for the third measurement with presence highlighted (red=less than 14 hours present, blue=14 hours or more present, grey=no information available)

From the plots of the networks there are no clear clusters visible for the differences in presence. It seems that the students do not form clusters based on their presence at lectures and tutorials.

5.4.5 Covariates influencing network: average out-degree

In order to see if there are differences in the number of nominations for certain covariate values, the average out-degree for each covariate group is calculated. This is the average number that each actor nominates to be his friend. For the gender, nationality, presence and living situation the calculations are performed and the results are shown in figure 5.18.

From the graphs it is clear that girls tend to nominate more friends than boys. Furthermore Dutch students tend to nominate less people than the international students do. The people who are not so many hours present at the lectures and tutorials, nominate in the last measurement significantly fewer students than the students who are many hours present. This is probably because these students have more friends from outside the study and they don't spend so much time together with their fellow students, so the contacts will fade. From the living situation we see that the average number of nominations of people who live at their parents stays almost equal over all measurements, whereas the average number of nominations of the people that are living away decreases. We see from the graphs that there are differences between the values of the covariates with respect to the number of nominations. Therefore the covariates indeed are likely to have an influence on the network.

Note that for the second measurement the average of the whole population is not a weighted average of the values for the separate groups based on the covariates. This difference is caused by the missing data at the second measurement. At this measurement there are students who completed the questionnaire, but from whom the nationality or gender is unknown.

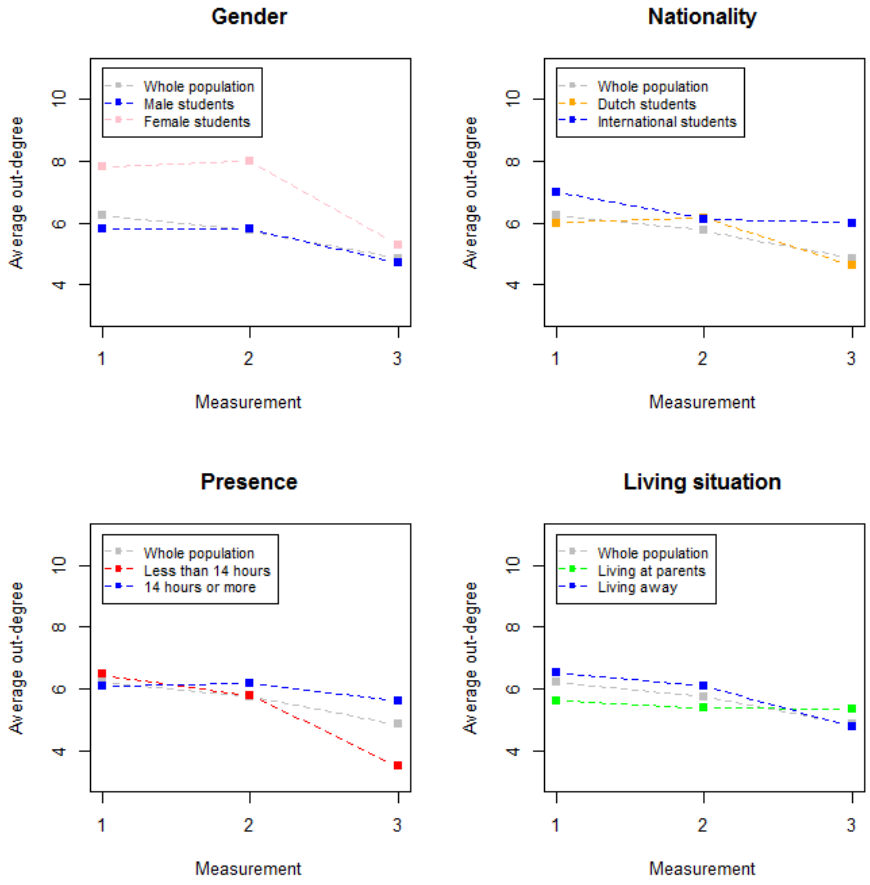


Figure 5.18: The development of the average out-degree for different covariates (in all pictures: grey=average of the whole population, gender: blue=male, pink=female, nationality: orange=Dutch, blue=international, presence: red=less than 14 hours, blue=14 hours or more, living situation: green=living at parents, blue=living away)

5.4.6 Covariates influencing the average grade

In order to see if the covariates have an influence on the behaviour variable (average grade), we can calculate the average grades of the different covariate groups. We calculated the average grade of the girls and the average grade of the boys for all measurement moments. This was also done for the Dutch and international students, for the people who are less than 14 hours present and the ones who are 14 hours or more present and for the students who live at their parents and the students who live away. The results are shown in figure 5.19. Here the same colours are used as for the plots of the clusters in the previous section. In all plots in grey the average grade of the whole population is shown.

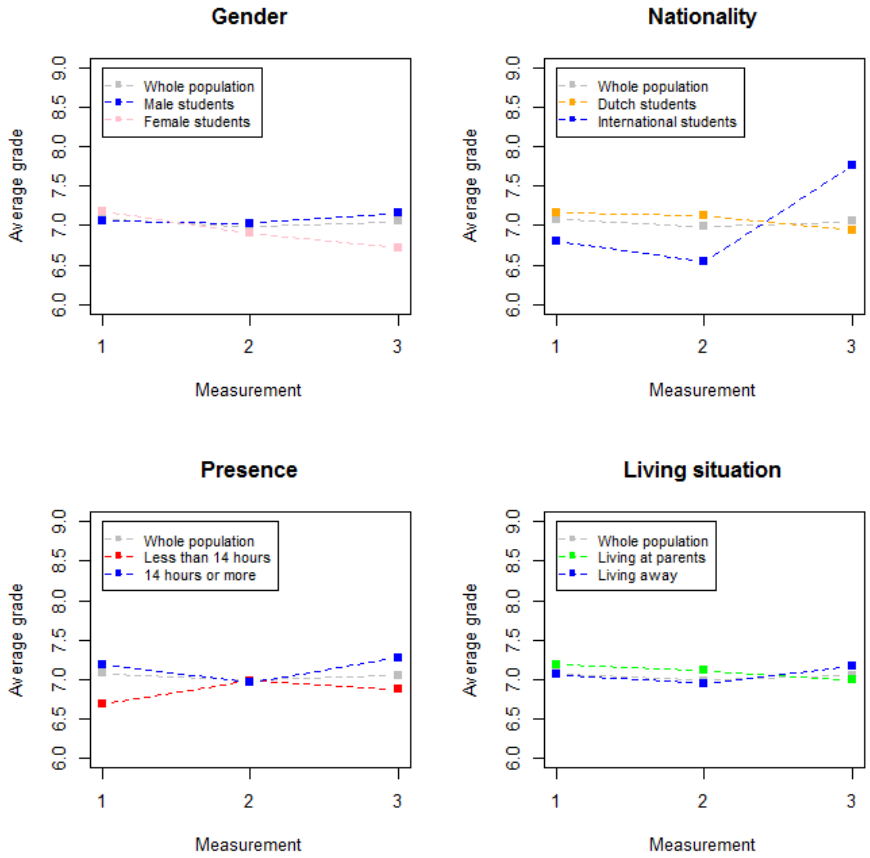


Figure 5.19: The development of the average grade for different covariates (in all pictures: grey=average of the whole population, gender: blue=male, pink=female, nationality: orange=Dutch, blue=international, presence: red=less than 14 hours, blue=14 hours or more, living situation: green=living at parents, blue=living away)

For the gender it can be seen that the girls average is going down over time, whereas the average of the boys is slightly increasing. This is remarkable, because in the beginning the average grade of the girls is higher than the boys', but in the end the girls average grade is significantly lower. Apparently the girls are distracted by the student life, maybe they moved out or the courses are more difficult for the girls. Another important aspect to keep in mind is that the population of girls is quite small, so the changes in the average grade can be large if one or two students have a bad exam week.

From the nationality-plot it can be seen that for the first two measurements the performance of the international students is significantly worse than the performance of the Dutch students. However, in the last measurement the performance of the international students is significantly better than the academic performance of the Dutch students. This seems somewhat logical, because the international students are mostly motivated to finish the study. They come to a good university and they pay a lot, so they are motivated to study hard. However, in the beginning they have to get used to the Dutch system. The preliminary knowledge might be different for the international students than for the Dutch students. Moreover they are far away from home, so they have to get used to the situation. Once they know how it works in the Netherlands and they are at the same level as the Dutch students with respect to preliminary knowledge, the motivation will pay off.

However, here we again have to remember that the population is small, so the influence of one or two good or very bad students will be large in the average degree.

For the presence we see that for the first and third measurement the people who attend less than 14 hours of lectures and tutorials have a lower average grade. For the second measurement both groups have the same average grade. The lower average grade is expected, because the lectures and tutorials try to be useful to understand the topics. For the second measurement maybe topics were covered that were easier to learn by self study. From the plots in figure 5.8 it is also clear that there are some students who have a peak in the presence hours at the second measurement. This shows that some of the students increase their attendance after the first exam week (and they cannot persevere that for the rest of the year).

The average degree of the students who live at their parents and the ones who live away does not vary that much. Apparently the influence of the living situation on the average degree is not so large. The students who live away loose time that they in principle could use for study because they have to do the household, whereas students who live at their parents loose this time when they are travelling.

5.4.7 Indications for stoppers

As already mentioned, the fraction of missing data is quite large. This is also because we don't know which students stopped their study. In order to get an idea about this the students were asked to agree or disagree on a five-fold scale to these three statements:

1. I like the mathematics bachelor.
2. I think I will finish my mathematics bachelor.
3. The study is harder than I expected.

For these three statements the students get a score. For the first two statements a 5 is given for strongly agree, a 4 for agree and so on. For the third statement a 5 is given for strongly disagree, a 4 indicates disagree and so on. The scores on these three statements are summed. The scores for all actors at the first and second measurement moments are investigated. This score is for the third moment not interesting, because there all students that are present did not stop until the last measurement.

The actors with low scores are candidates for stopping. For the first measurement the lowest score is 8 (for actor 24), then 9 (for actor 20 and 27) and 10 (for actor 5, 6, 7, 10, 12, 13, 23, 32, 41, 53, 56). In order to see if some of these actors stopped, it is checked if they completed a questionnaire later on. From all actors that we found as candidates, actor 24, 12, 53 and 56 completed no questionnaire any more. Therefore it is likely that these actors stopped their study. However, they still are nominated in the later measurements. Most of them are just one time nominated in the last measurement. By the nomination it is shown that they are still friends with (former) fellow students, therefore we don't remove these students. If a similar research project is performed in the future, it is wise to remove the students who stopped the study from the list with names and numbers, so that these students are removed from the population. Mostly it is hard to get to know who stopped the study.

For the second measurement the lowest scores are 8 (for actor 22), 9 (for actor 11, 27, 32 and 45) and 10 (for actor 5, 10, 59 and 64). Again for these candidates it is checked if they completed another questionnaire. Actor 22, 11 and 5 did not complete the third measurement, so it is likely that these students also stopped. However, again most of these students were nominated at least once in the last measurement. So we also don't remove these students from the data set. Therefore we keep the whole data set in to do the analysis in the next chapter.

Chapter 6

Analysis of the data

By applying the stochastic actor-based model we would like to answer the research questions. Therefore already in section 2.2 the variables were specified that we use in the model. For the network we study the friendship relationships between the actors. In order to fit the network, the out-degree, reciprocity, transitive triplets and transitive ties effects are included. This is done because in friendship networks reciprocity plays an important role. The behaviour part that we are interested in is the academic performance. This is measured by the average grade of the students. Then there are some covariates that are useful in order to answer the research questions. These are the nationality, the attitude towards the international bachelor and the presence at lectures and tutorials. Besides these variables we include two control variables that we think have influence in the friendship formation. These two are the gender and the living situation of the students. All covariates get each three effects: similarity or same, alter and ego. In this way it is tested if the people with similar or the same values for the covariate become friends. It is also tested if people with a high value for the covariate are more popular, and thus get nominated more often. Moreover the ego-effect tests whether the higher value for the covariate influences the number of friends that the actor nominates. Furthermore the effect of gender, nationality, presence and living situation on the academic performance is included. Also the average academic performance of the neighbours is included, in order to observe influence. The last effect that we include is an interaction effect between the nationality and the behaviour towards English speaking. In this way we can see if Dutch people who are speaking more English have more international friends. With all these variables included, the model is fitted with the use of the *RSiena*-package.

6.1 Results

The effects that are described above, are included in the model. The results are shown in table 6.1. Here the first column exists of the effects that are included. The upper part shows the effects of the network dynamics, whereas the lower part consists of the results of the behaviour part (so the average grades). The second column shows $\hat{\theta}$, this is the parameter estimate. Besides that the standard error ($s.e.$) is shown and $\left|\frac{\hat{\theta}}{s.e.}\right|$ is calculated to see which effects are significant. The significant values are indicated with a star. The full interpretation of the results will be discussed in section 6.2.

Network Dynamics			
	$\hat{\theta}$	$s.e.$	$\lvert\frac{\hat{\theta}}{s.e.}\rvert$
rate (period 1)	7.50	1.03	
rate (period 2)	6.61	0.90	
out-degree	-3.27	0.33	9.78*
reciprocity	1.96	0.36	5.47*
transitive triplets	0.14	0.03	4.14*
transitive ties	1.19	0.44	2.73*
gender alter	0.74	0.22	3.30*
gender ego	-0.42	0.24	1.74
same gender	0.41	0.16	2.56*
nationality alter	-1.07	0.32	3.34*
nationality ego	0.34	0.39	0.86
same nationality	-0.63	0.23	2.77*
academic performance alter	0.07	0.09	0.92
academic performance ego	0.10	0.09	1.12
similar academic performance	-0.22	0.88	0.26
presence alter	-0.01	0.01	0.71
presence ego	0.01	0.01	0.74
similar presence	0.29	0.43	0.67
living situation alter	-0.42	0.18	2.28*
living situation ego	-0.27	0.18	1.49
same living situation	0.22	0.14	1.58
English speaking alter	0.10	0.05	1.90
English speaking ego	-0.06	0.06	0.97
similar English speaking	0.29	0.45	0.64
nationality ego × English speaking alter	0.25	0.14	1.80
Behaviour Dynamics (Academic Performance)			
rate (period 1)	4.43	2.04	
rate (period 2)	2.52	0.88	
linear shape	0.04	0.12	0.32
quadratic shape	-0.05	0.04	1.12
average alter	0.07	0.31	0.22
gender	0.02	0.25	0.09
nationality	0.02	0.27	0.06
presence	0.01	0.01	0.27
living situation	0.02	0.21	0.10

Table 6.1: Results of the first model for the friendships network

The fitting was done three times, where the second and third time the previous answer was used as starting point and for the last fitting n_3 was set to 5000 in order to obtain accurate estimates for the standard error. In this way all convergence t-ratios are in absolute value smaller than 0.1. Moreover the overall maximum convergence is 0.21, this is smaller than 0.25. Therefore the model is nicely converged [13]. In order to see how the different

structures are modelled, we can have a look at the goodness of fit diagrams. These diagrams give an idea about the various structures in the model. The red line indicates the data, the box plots show the simulated values and below the diagram a p-value is visible. This is the probability that the distribution is indeed as is shown in the diagram. At a significance level of 0.05 this means that a p-value higher than 0.05 shows a nice fitted model. The diagrams for the indegree, outdegree, triangular structures and distances are shown in figure 6.1, 6.2, 6.3 and 6.5. The codes on the x-axis of the triad census goodness of fit are explained in figure 6.4.

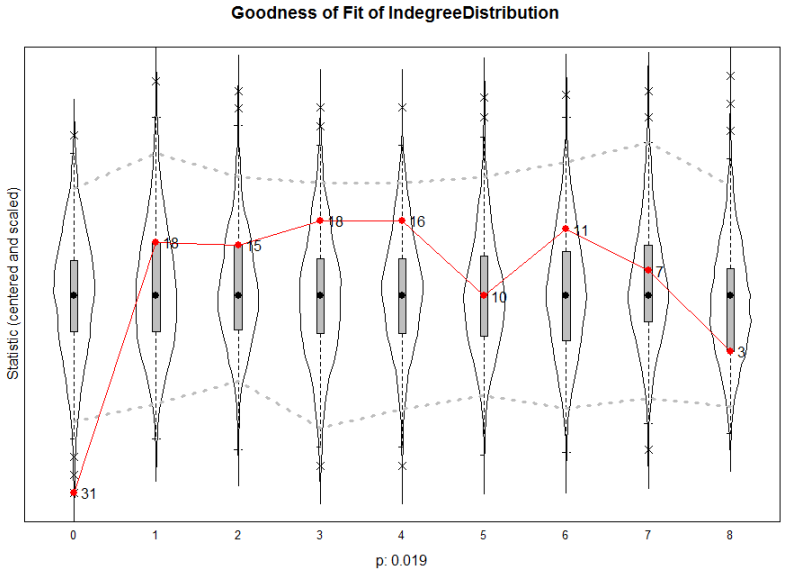


Figure 6.1: The distribution of the incoming ties

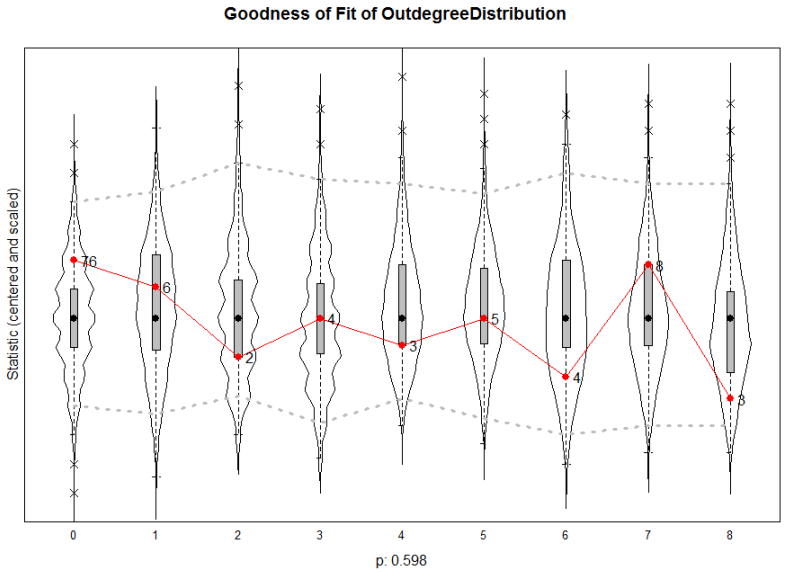


Figure 6.2: The distribution of the outgoing ties

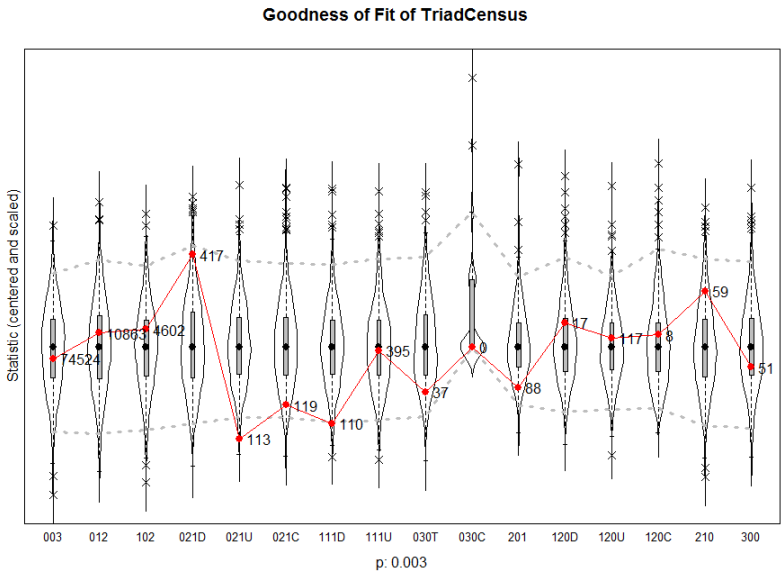


Figure 6.3: The distribution of the triangular structures (the codes on the x-axis correspond to different triangular structures as shown in figure 6.4)

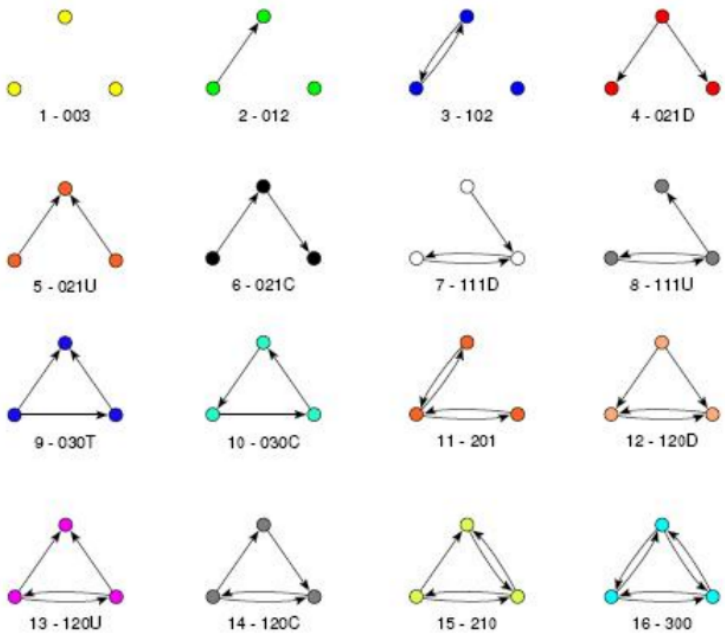


Figure 6.4: The triangular structures that are shown in the goodness of fit of the triad census (figure 6.3)

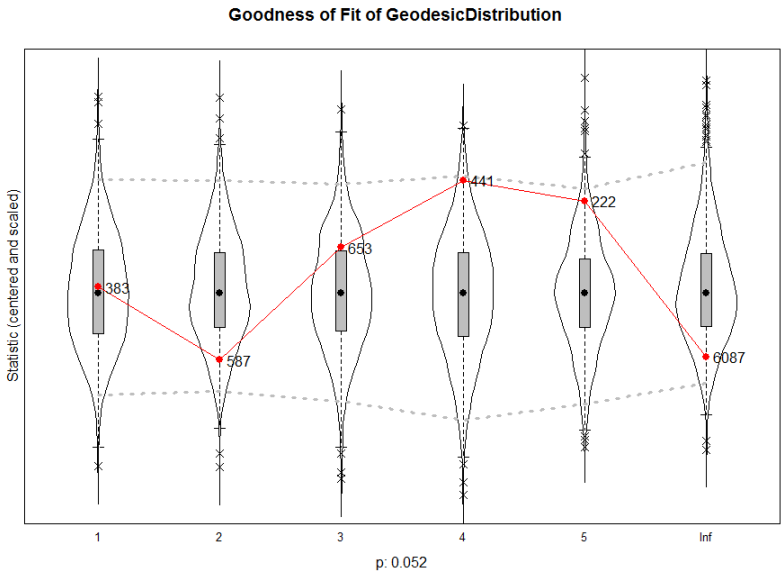


Figure 6.5: The distribution of the distances

From these plots it can be seen that the in-degree and the triangular structures are not well modelled. The isolated actors (with zero incoming ties) and the triangular structures 021U and 111D are worst modelled, here 95% of the simulated data does not include the observed value. The outdegree is nicely modelled and also the distances have a p-value that is slightly higher than 0.05.

It can be tested if there is time heterogeneity, so that a dummy term must be included. In the graphs of figure 6.6 and 6.7 the horizontal black line shows the estimate for the first period, between measurement one and two. The right dot is the estimate for the second period. The vertical red stripe shows 95% confidence interval for this estimate. So if the horizontal line ends up in the red 95% confidence interval, the same estimate also holds for the second period and therefore no time dummy is needed. Otherwise we need to include a time dummy.

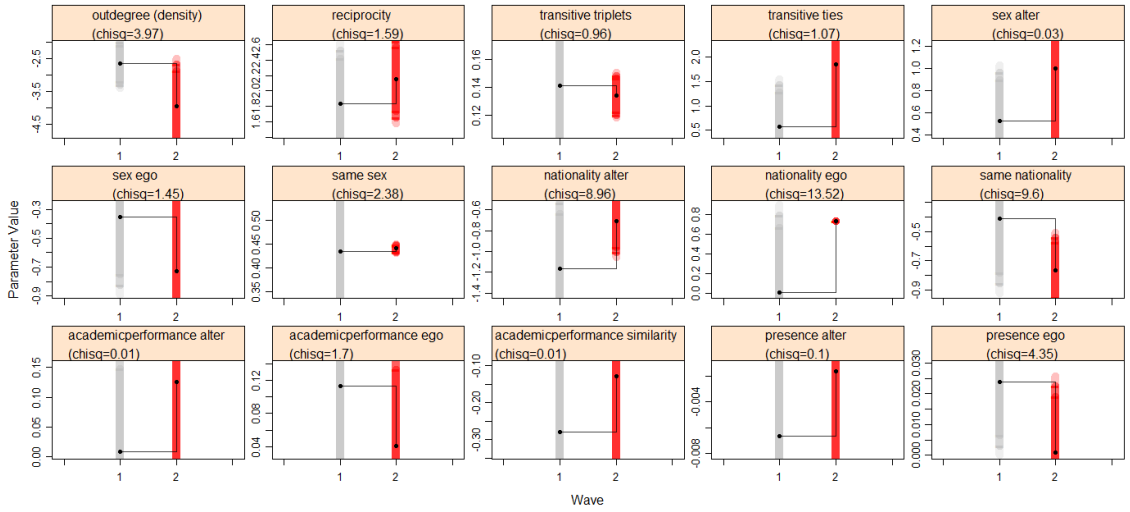


Figure 6.6: The tests for time heterogeneity for the first 15 effects

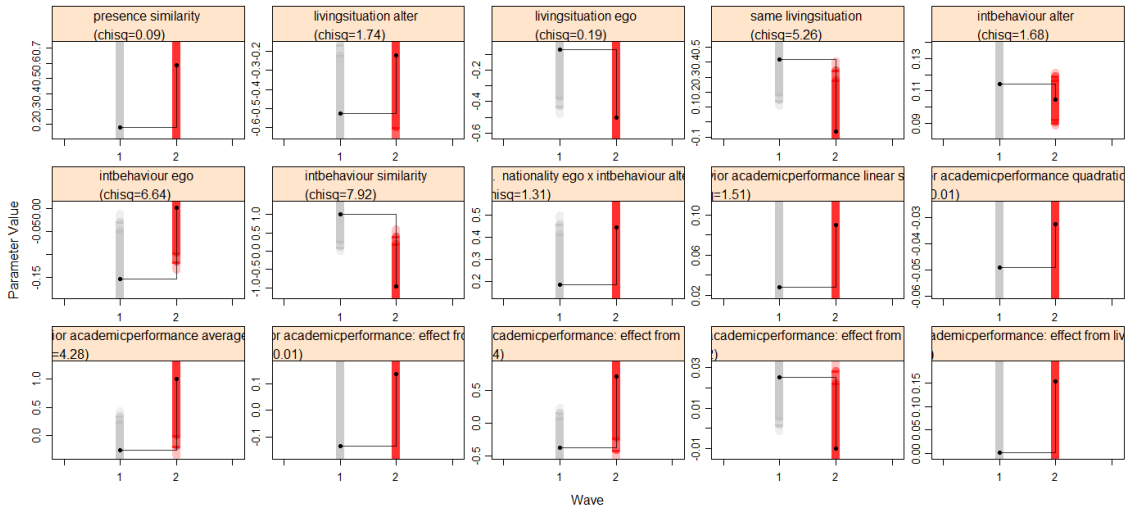


Figure 6.7: The tests for time heterogeneity for the second 15 effects

From these plots it is clear that the nationality ego needs a dummy term. Furthermore the effects nationality alter, same nationality, the attitude towards the international bachelor (in the plot called intbehaviour) ego and the attitude towards the international bachelor alter are outside the confidence interval and therefore we have to include a dummy term in our new fit.

6.2 Interpretation

As already explained in section 4.4, an effect is significant when $\left|\frac{\hat{\theta}}{\text{standard error}}\right| > 1.96$. Therefore we see that the structural effects out-degree, reciprocity, transitive triplets and transitive ties are significant. The sign of the out-degree is negative, what means that the network is more sparse than half of the ties present. The sign for the other effects is positive, what means that people tend to form reciprocated ties, transitive triplets and

transitive ties. For the gender we see that the alter and same effect on the network formation are significant. They both have a positive sign. This means that girls are more often nominated than boys (girls are coded as 1, boys as 0) and that the girls tend to become friends with girls and boys with boys. For the nationality we see that the alter and same effect on the network are significant. The alter effect has a negative sign. The nationality is coded as 0=Dutch, 1=international. The negative sign now shows that a higher value of the nationality has a lower popularity. Therefore the international students are less popular than the Dutch students, the Dutch students are nominated more often. The same nationality effect has a negative sign. This means that people tend to become friends with people of different nationalities. So the international students mix well with the Dutch students. There are no separate clusters. The effect of the academic performance and the presence on the network formation are both not significant in the model. Therefore these effects don't play an important role. So we don't see that the students with high grades have friends with high grades. This clustering was observed for high school, but our model shows that it is not the case for university. It also doesn't matter if a student is many hours present. He/she does not have more or less friends than students who are not so many hours present. The alter-effect of living situation on the network formation is significant. This has a negative sign and living situation was modelled with 0=living at home, 1=living away. Therefore the people who live at their parents are more popular. This might be because the students who live at home usually are no member of student associations or sport associations and therefore these students actively try to make friends at their study. The effects of the behaviour towards English speaking on the network are not significant. The effects of the gender, nationality, presence and living situation on the academic performance are also not significant. Because of the relatively small data set for the academic performance it was also tried to include only one of the effects. In this way it would be easier to find an effect that is significant. However, also the models with only one effect included don't show a significant effect on the academic performance.

6.3 Include time heterogeneity

Because of the observed time heterogeneity for the effects nationality alter, same nationality, the attitude towards the international bachelor ego and the attitude towards the international bachelor alter, a dummy term is added for these effects. The model that is fitted with these dummy terms included gives the results that are shown in table 6.2.

Network Dynamics			
	$\hat{\theta}$	$s.e.$	$\lvert\frac{\hat{\theta}}{s.e.}\rvert$
rate (period 1)	7.05	0.93	
rate (period 2)	6.20	0.81	
out-degree	-3.40	0.35	9.74*
reciprocity	1.99	0.34	5.80*
transitive triplets	0.13	0.03	3.74*
transitive ties	1.37	0.46	2.92*
gender alter	0.78	0.22	3.48*
gender ego	-0.48	0.25	1.90
same gender	0.44	0.17	2.60*
nationality alter period 1	-1.21	0.34	3.60*
nationality alter period 2	-0.35	0.61	0.57
nationality ego period 1	0.69	0.42	1.63
nationality ego period 2	2.22	0.88	2.54*
same nationality period 1	-0.73	0.23	3.13*
same nationality period 2	-1.10	0.31	3.52*
academic performance alter	0.07	0.09	0.81
academic performance ego	0.06	0.09	0.67
similar academic performance	-0.34	0.96	0.35
presence alter	-0.01	0.01	0.81
presence ego	0.02	0.01	1.73
similar presence	0.16	0.43	0.37
living situation alter	-0.38	0.18	2.06*
living situation ego	-0.19	0.19	1.00
same living situation	0.20	0.14	1.40
English speaking alter period 1	0.15	0.05	2.57*
English speaking alter period 2	0.03	0.13	0.22
English speaking ego period 1	-0.16	0.07	2.37*
English speaking ego period 2	-0.25	0.15	1.65
similar English speaking	0.11	0.47	0.24
nationality ego × English speaking alter	0.35	0.14	2.44*
Behaviour Dynamics (Academic Performance)			
rate (period 1)	4.41	1.49	
rate (period 2)	2.61	0.88	
linear shape	0.04	0.11	0.34
quadratic shape	-0.04	0.04	1.12
average alter	0.07	0.30	0.25
gender	0.02	0.26	0.10
nationality	0.03	0.27	0.11
presence	0.01	0.01	0.50
living situation	0.02	0.21	0.14

Table 6.2: Results of the model for the friendship network with time dummies included

From this table we see that the model is similar to the model without the dummy terms. The effects have the same sign and order of magnitude. The effects where the time dummies are included show the largest changes. For these effects there are now two estimates, for each time period one. For the estimate for the second time period, the time dummy term that was obtained from the model was added to the estimate for period 1. The variance for this estimate is calculated by variance(estimate 1) + variance(estimate time dummy) + 2×covariance (estimate 1, estimate time dummy). The standard error can be obtained by taking the square root of the variance.

Note that for example for the effect "nationality ego" the parameter estimate for the model without time dummies is lower than for both separate periods in the model with time dummies. This is remarkable, but this discrepancy might be caused by compensation (for example the rate parameters are in the model with time dummies lower than in the model without time dummies). Furthermore, the effect is in the model without time dummies not significant.
The nationality alter effect only has a significant influence on the network during the first time period. This means that the Dutch students are more often nominated than the international students in the first time period. The effect of nationality ego on the network evolution is only significant in the second time period, with a positive parameter. This means that the international students tend to nominate more friends than the Dutch students do in the second time period. The same nationality effect has for both periods a significant influence on the network. It has a negative parameter. This was also obtained in the model from the previous section. In the second time period the mixing of the different nationalities is even larger than in the first time period, as can be seen from the larger parameter estimate. The behaviour towards the international bachelor (the amount of English usage during breaks and tutorials) has for the first time period a significant effect on the network evolution. In this period the alter effect has a positive sign and the ego effect has a negative sign. This means that during the first time period people who talk English more often are more popular and less active. The last difference that is observed between the model with and the model without time dummies is that the interaction effect nationality ego × English speaking alter is significant only for the last model. This means that the international students tend to have more friendship nominations towards students who most often speak English.

The overall convergence of the model is 0.2402. This is smaller than 0.25. Furthermore all convergence ratios are smaller than 0.1, so the model is nicely converged. The goodness of fit plots can be found in the appendix (figure A.1, A.2, A.3 and A.4). From these plots it is clear that the model does not fit the in-degree, triangular structures and distances so well. Again the isolates and 012U and 111D are badly fitted. The out-degree is nicely fitted. The new test for time heterogeneity is also shown in the appendix (figure A.5 and A.6). It is clear that now all variables end up in the red 95% confidence interval and thus there is now no more time heterogeneity that needs to be solved.

The addition of the time dummy terms indeed solves the time heterogeneity. The goodness of fit statistics still are, except for the out-degree, not so good. The addition of the dummy terms makes the model more difficult and extensive. However, the data set is not so large. The addition of the dummy terms does solve the time heterogeneity, but it does not give so much new information, whereas it takes a lot more calculations to fit the model. Therefore it is likely that we are over-fitting with the dummy terms. Therefore it

might be a better idea to use the model without the dummy terms. The goodness of fit diagrams of this model are not optimal. Especially the isolated actors, and the triangular structures 012U and 111D are not nicely modelled. In principle it is possible to include these effects in the model. In that case the goodness of fit will be improved, but this is a trouble shooting approach without much physical meaning. Therefore we don't do this and we decide to take the model without the time dummies as the preferred model, where we keep in mind that there is time heterogeneity and the goodness of fit is not everywhere perfect.

6.4 Impact of the covariate effects

In this section we will get a better feeling for the effects. What are the largest effects and what people have a preference to form ties to people with certain covariate values? In order to get some insight in the effects, so called ego-alter selection tables are formed for the covariates that had a significant effect in model 6.1. The covariates gender, nationality and living situation all had at least one significant effect and therefore for these covariates the ego-alter selection tables are constructed in the same way as in [14] is done. In the tables the hypothetical case is shown that all actors have the same network position and the same values for the variables included in the model, except for one covariate v. For this covariate, the ego-, alter- and similarity effect have an influence on the formation of a tie between actor i (in the table shown as ego) and actor j (in the table shown as alter). This influence is given by

$$\beta_{\text{ego}} v_i x_{i+} + \beta_{\text{alter}} \sum_j x_{ij} v_j + \beta_{\text{sim}} \sum_j x_{ij} (\text{sim}_{ij}^v - \widehat{\text{sim}^v}), \tag{6.1}$$

where sim_{ij}^v is the similarity (as was also shown in 3.5.4), given by $\text{sim}_{ij}^v = 1 - \frac{|v_i - v_j|}{R_V}$ and $\widehat{\text{sim}^v}$ is the mean of all similarity scores [13]. Now the contribution of the formation of a tie from i to j to (6.1) can be represented by the single tie variable x_{ij}. This means that we have a look at the difference between the values for (6.1) when $x_{ij} = 1$ and when $x_{ij} = 0$. Because *RSiena* centers all values, the mean values used for the centering are subtracted. Therefore the contribution of covariate v to the evaluation function of actor i is given by

$$\beta_{\text{ego}}(v_i - \bar{v}) + \beta_{\text{alter}}(v_j - \bar{v}) + \beta_{\text{sim}}(\text{sim}_{ij}^v - \widehat{\text{sim}^v}). \tag{6.2}$$

For different values of the covariates, equation 6.2 can be evaluated. The results of these calculations are shown in the tables for the different covariates (gender, nationality and living situation), as can be seen in figure 6.3, 6.4 and 6.5.

Gender

		Alter	
		Male	Female
Ego	Male	0.079	0.388
	Female	-0.771	0.401

Table 6.3: Ego-alter selection table for the covariate gender

Nationality

		Alter	
		Dutch	International
Ego	Dutch	-0.058	-0.499
	International	0.913	-0.787

Table 6.4: Ego-alter selection table for the covariate nationality

Living Situation

		Alter	
		Living at home	Living away
Ego	Living at home	0.320	0.117
	Living away	0.262	-0.375

Table 6.5: Ego-alter selection table for the covariate living situation

From the ego-alter selection table for the gender, we can see that male students have a tendency to prefer relations with female students (0.388) and in much smaller extend also male students (0.079). The female students tend to prefer relations with other female students (0.401) and not with male students (-0.771). In the model the alter effect showed that female students are more popular than male students. This matches with the values found in table 6.3. The model also showed a same gender effect, but the ego-alter selection table shows that this effect is only present for the female students.

For the covariate nationality, it can be seen that the Dutch students show a tendency not make a connection to international students (-0.499) nor to Dutch students (-0.058), whereas the last tendency is very weak. The international students tend to strongly prefer friendships with Dutch students (0.913) and not with international students (-0.787). The model showed that there is a same nationality effect that is negative (so a preference for the other nationality). From the ego-alter selection table it is seen that this effect is very strong for the international students and it is not present for the Dutch students. Moreover, the model showed that the Dutch students are more popular. That is also observed in the ego-alter selection tables.

For the living situation, students who live at home exhibited a tendency to prefer relations with other students who live at home (0.320) and to a smaller extend also with students who live away (0.117). The students who are living away tend to prefer a connection to students who live at home (0.267) and not to students who live away (-0.375). The model showed a significant effect of the living situation alter. The effect was negative, what means that the students who live at home are more popular. That is indeed what is observed in the ego-alter selection tables.

As a summary for this section we can have a look at the biggest effects. The biggest effects in the model are the tendency of the international students to prefer the formation of a connection to Dutch students and not to international students. Another large effect is the tendency from male students to prefer the friendship formation with female students.

6.5 Coupling to research questions and discussion

The structural effects that are significant in this model, are more often observed in friendships networks. People tend to have a reciprocated friendship.

Considering the variables in the model, we see that the effect of nationality on the network is significant. We saw that the Dutch students are more popular. We saw in section 5.4 that the international students tend to nominate more students, but this effect was too small to be significant in the model. When time dummies are included this effect is significant for only one time period. However, the time dummies make the model more complex and rise the probability of over-fitting. What we further see in the model is that the nationalities mix, they don't form separate clusters. This also answers one of the research questions: there are no separate clusters of Dutch and international students, but they tend to mix. This is a nice result, because this is what the board of the university likes to see. This gives the international bachelor indeed a real international character. The ego-alter selection tables (shown in section 6.4) show that this preference for the other nationality is only observed for international students. The tendency of the international students to prefer the formation of a connection to Dutch students and not to international students is large.

The effect of the attitude towards the international bachelor on the network position is not significant in the simple model. In the model with the time dummies included the ego and alter effect of the English speaking behaviour are significant for the first time period. Therefore it seems that for the first period the English speaking behaviour is important, but later on this does not play a role any more. That can be because the English speaking behaviour is especially important for the first acquaintance. Another observation is that the interaction of nationality ego $\times$ English speaking alter is significant in the model with the time dummies. This might state that in the beginning international students nominate more friends who speak mostly English. Maybe later on this effect is not present any more and therefore it is not significant in the simple model. This is also an answer to one of the research questions: the attitude towards the international bachelor (in particular towards the use of English) has an influence on the friendship formation between Dutch students and international students only in the beginning of the academic year.

The presence at lectures and tutorials is has no significant effect on the network formation nor on the academic performance. That means that the presence of the students has no influence on the friendship formation and not on the academic performance. This also answers one of the research questions: there is no influence of the presence at lectures and tutorials on the friendship formation nor on the academic performance. This is somewhat counter-intuitive because the lectures and tutorials are there in order to help the students pass their exams. That there is no effect of the presence on the academic performance can be a little bit disappointing for the teachers. Now it looks like the lectures are not really needed or the exams are too easy so that you can also pass it by not attending the lectures. However, it can also be the case that the students who are a little bit smarter decide not

to go to the lectures and tutorials, because they don't need them. The smarter students can teach themselves and therefore there is no effect visible. There is also no effect of the presence on the friendship networks. That means that there is no clique of students who are always present that don't allow others to become friends with them. The people who are more often present have the same amount of friends as people who are less present. There are also no clusters of frequently present people and almost never present people. This shows that the group is quite open.

For the control variables we see that the effect of gender and living situation are indeed significant. Therefore it was good to include these control variables. We see that the female students tend to be more popular and there is a drive to become friends with people of the same gender. The ego-alter selection tables also show that female students are more popular. The same gender effect is only present for the female students (see section 6.4). In section 5.4 it was shown that girls tend to nominate more people than boys. This effect was not found in the model, so it is probably too small to be detected. From the living situation we see that the people who live at their parents are more popular. The reason for this might be that the students who live at home are usually not a member of a student association or have other friends from student houses or so. Therefore these students are more actively making friends at their study environment. The effect of the same living situation is not significant, so there are no clusters of people who live at their parents and people who live away. This is in accordance to what we already observed in section 5.4.

We saw in the model that the effect of the academic performance on the network evolution is nowhere significant. Therefore we see that people with high grades do not form friendships only with people with high grades. This answers one of the research questions: there is no cluster formation between the people with high grades. Furthermore the grades do not have an influence on the number of friends that the students nominate. This is different from what is found at high schools [2]. At high schools the differences between the people are larger. In the first year mathematics at the university mostly people with high grades at exact courses at high school enter. Therefore the differences between the level of the students is smaller than at high school. Furthermore at the university the emphasis is more on passing a course. In that way you obtain your EC's and therefore you come closer to your degree. This is more important than the grades that you obtain. This mentality is more often seen at the university and this can be the reason why these kinds of clusters are not visible. As the behaviour variable the academic performance is not significant either. The effects of the gender, nationality, presence and living situation on the academic performance are not significant. However, in section 5.4 there were some differences in academic performance observed for the different classes. These differences are apparently quite small and therefore not observed in the model. For example for the performance of Dutch versus international students the effect is opposite for the first and last measurement and therefore this is difficult to see with the model. There is no effect of the nationality on the academic performance. That gives the answer for the last research question: no differences in academic performance are observed between Dutch and international students. The fact that the academic performance has no influence on the network also tells that there are no differences in number of friends observed for people with high grades and for people with low grades. This also matches with the network autocorrelation that was found in section 5.1.4. Here we found that there was almost no correlation between the network and the academic performance.

Chapter 7

Conclusion and discussion

7.1 Conclusion

In this thesis the co-evolution of the friendship networks and the academic performance of the first year mathematics students (cohort 2014-2015) of the University of Groningen is studied. Since two years the mathematics bachelor is completely taught in English. The influence of this change is studied as well. Data collection was done by distributing questionnaires to the first year students three times during the academic year. The obtained data set is analysed by the use of the stochastic actor-based model. The theory behind this model is discussed in chapter 3 and 4 of this thesis. In chapter 5 and 6 the data is analysed and the stochastic actor-based model is applied.

We were interested in answering five research questions:

- Are there clusters of people with high grades and clusters of people with low grades? And if this is the case, are these clusters formed by selection or by influence?
- Do the international students get higher grades than the Dutch students and does this influence the number of friends that they have?
- Is there nationality homophily observed? In other words do the international students mix with the Dutch students or do they form a separate cluster?
- Does the attitude towards the international character of the bachelor have an influence in the friendship formation between Dutch and international students?
- Does the presence of the students at lectures and tutorials have an influence on the academic performance and/or on the friendship networks?

In order to answer these questions the friendship network was modelled. For the behaviour part (academic performance) the grades of the students were used. Furthermore some variables were included as covariates: nationality, attitude towards the international bachelor and presence at lectures and tutorials. Furthermore we included two control variables: gender and living situation.

In the corresponding model an effect from the gender on the network formation was observed. The female students are more popular, they are nominated more often. The same gender effect was significant for the influence on network formation. The ego-alter selection tables showed that this effect is largest for female students.

For the living situation it turned out that students who live with their parents are more popular and get more nominations. The reason for this might be that students who live at home have less friends outside the study and therefore they more actively try to make friends at the study. So the two control variables gender and living situation indeed have an influence on the friendship networks.

For the variables that were included in order to answer the research questions we saw that the negative nationality alter effect on the network showed that the Dutch students are more popular. They get more nominations. The effect of same nationality is significant with a negative sign, what means that the students of different nationalities tend to mix. From the ego-alter selection tables it is clear that the international students have a strong tendency to prefer a friendship formation with Dutch students and not with international students. There are no separate clusters based on nationality. There is time heterogeneity in the nationality effect. The nationality ego-effect is only significant for the second time period. This effect shows that the international students tend to nominate more people than the Dutch students do in the second time period. The effect of nationality on the academic performance is not significant. This means that there is no difference observed between the average grades of the international students versus the Dutch students.

The attitude towards the international bachelor is measured by the differences in the usage of English during tutorials and breaks. This is done because there was no difference observed in the opinion about the international bachelor and therefore this opinion variable was not included. For the simple model there was no effect of the attitude towards the international bachelor on the network. When the time dummies were included it turned out that this attitude has effect only in the first time period. This effect was that the people who speak more English during breaks and tutorials, nominate less people to be their friend, but these people are nominated more often. The interaction term nationality ego $\times$ English speaking alter was only significant in the model with time dummies. This effect indicates that international students tend to nominate more people who speak English all the time. The addition of time dummies to the model makes it more complicated, so we have to keep in mind that this model might suffer a bit from over-fitting.

The presence at lectures and tutorials had no influence on the network or on the behaviour. That means that it was also not observed that a higher presence gives better exam results. This looks counter-intuitive, because the lectures and tutorials should increase the chance to pass the exam. It might be the case that the more talented students don't follow the lectures because they don't need them and they study the material themselves. Another option is that there is not enough data available to find this effect.

An interesting observation is that the effect of the academic performance is nowhere significant. This means that the academic performance has no influence on the network, so there are no clusters of people with high grades and clusters of people with low grades. This is different from what was observed by [2] for high school students. This difference might be explained by the fact that at the mathematics bachelor mostly students enter with high grades at high school so the students already have a more similar level. Furthermore the focus at the university lies more on passing the courses than on the grades. The level of the grades has no influence on the number of nominations. This not observed influence of the academic performance on the network matches with the fact that we found

almost no network autocorrelation between the friendships and the academic performance.

As a summary the answers that were found to the research questions are given below.

- There are no clusters of people with high grades and clusters of people with low grades. This is different from what is found in high schools. There is no influence or selection based on the academic performance.
- There is no difference observed between the academic performance of Dutch students and international students. The international students don't get higher grades than the Dutch students and there is also no effect on the number of friendship nominations caused by the level of academic performance.
- The international students and the Dutch students mix. The international students tend to prefer the connection to Dutch students and not to other international students.
- The opinion towards the international bachelor is positive for all the students. The usage of English is not similar for all students. This makes a difference in the friendship formation only in the first time period. The people who talk more often English are more popular and they have more nominations from international students than the people who talk more often Dutch. These effects were only found in the model with time dummies, so the effects are not present in the simple model. For the model with time dummies we have to be careful, because here is a higher probability of over-fitting.
- The presence at lectures and tutorials has no effect on the friendship formation nor on the academic performance.

7.2 Discussion

The fraction of missing data is quite large for this study. That makes the reliability of the model smaller. The large fraction of missing data is partly explained by the students that stop with the study, but that are still on the list with names and numbers. Therefore they still can be nominated and in this way it is really hard to determine which students stopped the study. If a similar research is performed in the future it is wise to take more effort in order to know which students stopped the study and to remove them from the population. However, it is quite hard to know this, because most students don't let the university know that they quit. What can be a nice suggestion for a follow-up study is the modelling of the people who stop. Do people who have almost no friends stop earlier than people with more friends? Are only the grades indicative for what students stop? Questions like this can be part of a new research. Another part of the missing data is because in the second questionnaire some of the constant questions were removed in order to make the questionnaire shorter. However, there were students who were not present at the first measurement that completed the second questionnaire. Therefore the background of these students is not known. For a similar research in the future it is wise to never remove questions in order to avoid unnecessary missing data. Moreover the first year mathematics students are quite hard to reach, so some of the students are missed in some of the measurements. This is also because the mathematics bachelor has only a limited amount of compulsory courses. Maybe for another research in the future it is easier to get a more complete data set if another study (with more compulsory activities) is chosen.

The whole population is quite small. Therefore some groups (for example international students or female students) are in particular quite small. This might have a big influence on the obtained results.

An assumption of the stochastic actor-based model is that there are no more than one changes at the same moment. In practice this is not the case, it is possible that there are two changes at the same time. Therefore the model is a simplification of reality.

In section 5.4.6 it was shown that the average grades of the international students differ from the average grades of the Dutch students. In the beginning the international students obtained lower grades, but at the last measurement they obtained higher grades. The fact that the grades are first lower and later higher, makes it hard to see this effect in the model. Therefore it seems that there is no effect of nationality on the average grades. However, if a similar research would be performed at the second academic year, a difference in grades might be observed. When there are more data points it might be a good idea to use time dummies for this part of the model as well. In our study the data set is not large enough to do this, but in follow-up studies this might be a good idea.

It might be the case that there is a difference between the formation of a tie and the removal of a tie. Losing a friend might cost more energy than getting a new friend gains. This inequality in ties can be included in the model by adding an extra term to the objective function: a so-called endowment function. In later researches this effect might be included.

Another point to be careful about is the definition of friendship. This definition might be different in various countries. In the questionnaire it is partly explained what we see as friendship and how the students should answer the question, but for researches with different international backgrounds it is good to keep in mind that different cultures might use different definitions of friendship.

In this research the influence of the nationality, presence, attitude towards the international bachelor, gender and living situation on the friendship network and the academic performance are studied. Of course there are many more effects that might have an influence on the friendship formation. Furthermore there are many effects that can be studied as behaviour. For example the usage of English can be seen as behaviour variable. Some first implementations of this suggestion show that this new behaviour variable is (like the academic performance) also uncorrelated with the network changes. Other examples for behaviour variables to study in the future are sports, hobbies, contact with parents, ways of financing the study time or memberships of study associations. From the questionnaires information like this can be obtained, so it is possible to do a follow-up study with this data set. The data set already consists of csv-files of various constant background variables (like age, nationality, religion, siblings etc.) and time varying variables (like the presence at lectures, the way of financing the study time and the time spent at sports and hobbies). Furthermore, the network development of the best friends network, the network of collaboration for homework, the neighbours during lectures and the participation at non-study related activities can be studied in the future by using the data set.

Chapter 8

Acknowledgements

Almost one year back I had my first conversation about the master research project with Ernst Wit. There were quite some possibilities for subjects to study, but the subject of this thesis was the one that I immediately liked. With this subject the goal and the idea can be understood by for example your mother or your friends who study Arts. The formation of friendships is something that happens everywhere and everybody can come up with a reasoning why one becomes friends with another person or not. However, to make these theories quantitative can be hard. That is also shown in this project. The stochastic actor based model is used, but until now this model is based on the method of moments, because the likelihood is hard to calculate. Another difficult point in studies like this is to obtain all the data. Especially at universities there are not so many compulsory meetings, so it is hard to reach everybody. Furthermore it was good to see what the questions in your questionnaire can tell you and determine if this is ethically approved. With the help of Ernst Wit, Nynke Niezink and the ethical committee for social sciences the questionnaire was designed in a proper way. Thank you all for helping and giving suggestions on how to improve the questionnaire. After the design of the questionnaire, it was time to distribute the questionnaires. For this I had some help of the mentors (2014-2015). They helped me to reach the students and to ask them to complete the questionnaires. Mentors, many thanks for your contributions to my data collection. Moreover, thanks to Roel Luppes who made it possible to distribute the questionnaires during his practical courses. After the data collection, the analysis started. Because of the large amount of data that was present, it was sometimes hard to determine what you are looking for and what data is not immediately used. Fortunately Nynke Niezink was really helpful at this moment. She asked the right questions to move me into the right direction. Furthermore when I had questions or problems with the model or the interpretation, Ernst and Nynke were both really willing to help. Thanks a lot for the nice supervision. I really liked the project and with your help it was even nicer.

Bibliography

[1] Notes from a programming assignment at the university of washington. `https://courses.cs.washington.edu/courses/cse143/14wi/notes/notes10.html`. Accessed: 02-04-2015.

[2] Jennifer Flashman. Academic achievement and its impact on friend dynamics. *ASA*, 85(1):61–80, 2011.

[3] Mc Pherson and Cook Smith-Lovin. Birds of a feather: Homophily in social networks. *Ann. Rev. Soc.*, 27:415–444, 2001.

[4] N. Friedkin. *A Structural Theory of Social Influence*. C.U.P, 1998.

[5] Ferdi Doddema. *Introductie tot het stochastisch actor-georienteerd model*. Masteronderzoek, 2014.

[6] Christian E.G. Stechlich, Tom A.B. Snijders, and Michael Pearson. Dynamic networks and behaviour: separating selection from influence. *Sociol Methodol*, pages 329–393, 2010.

[7] Christian E.G. Stechlich, Tom A.B. Snijders, and Michael Schweinberger. *Longitudinal models in the behavioral and related sciences*. Mahwah, NJ: Lawrence Erlbaum, 2007. Edited by K. van Montfort, H. Oud and A. Satorra, page 41-71.

[8] Christian E.G. Stechlich, Tom A.B. Snijders, and Patrick West. Applying siena: an illustrative analysis of the co-evolution of adolescents' friendship networks, taste in music, and alcohol consumption. *Meth Eur J Res Meth Behav Soc Sci*, 2(1):48–56, 2006.

[9] Tom A.B. Snijders, Gerhard G. van de Bunt, and Christian E.G. Stechlich. Introduction to stochastic actor-based models for network dynamics. *Soc. Netw.*, 2009.

[10] Ramiro Berardo. Stochastic actor oriented models for longitudinal networks using rsiena. Presentation at the University of Wisconsin-Milwaukee, May 2014. Based upon copyrighted material of Tom Snijders.

[11] Tom A.B. Snijders, Johan Koskinen, and Michael Schweinberger. Maximum likelihood estimation for social network dynamics. *Ann Stat*, 4(2):567–588, 2010.

[12] Tom A.B. Snijders. The statistical evaluation of social network dynamics. *Sociol Methodol*, 31, 2001.

[13] Christian E.G. Steglich Ruth M. Ripley, Gerhard G. van den Bunt. *Manual for RSiena*. 2014.

[14] Slides from christian steglich, university of groningen. `http://www.gmw.rug.nl/~steglich/workshops/MixedTopics.pdf`. Accessed: 29-08-2015.

Appendix A

Figures from chapter 6

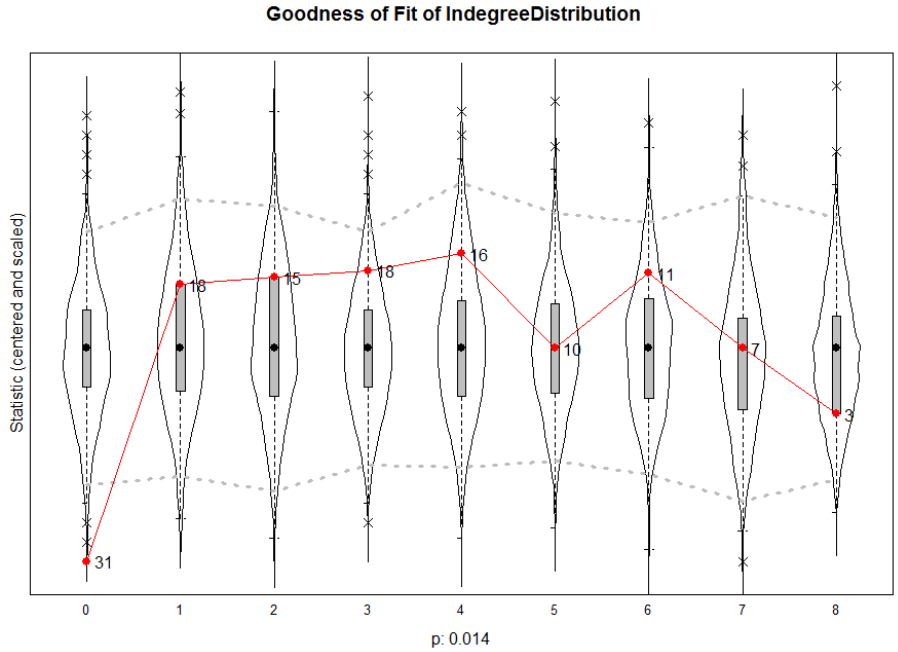


Figure A.1: Goodness of fit for the indegree distribution

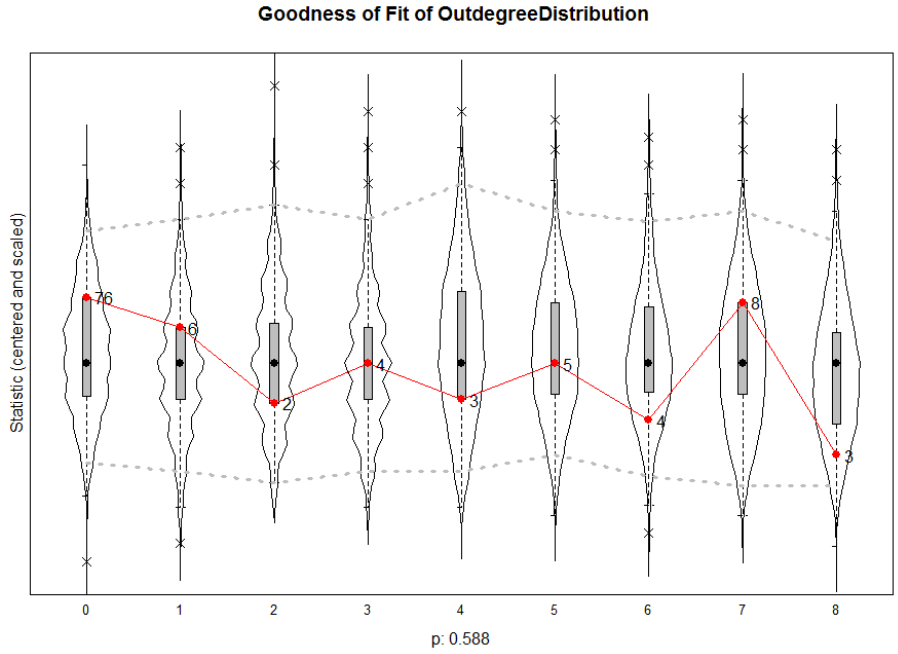


Figure A.2: Goodness of fit for the outdegree distribution

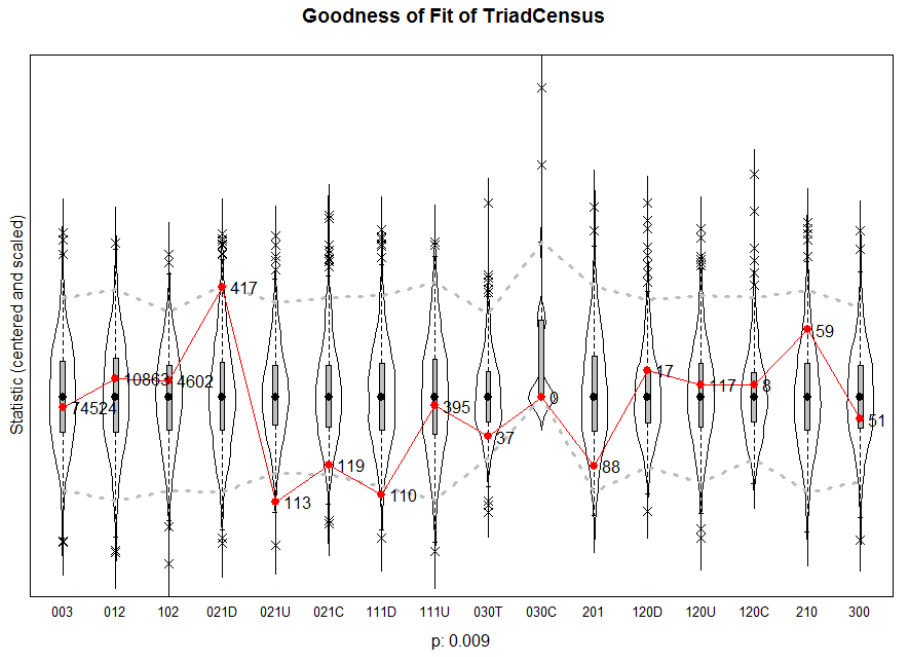


Figure A.3: Goodness of fit for the distribution of triangular structures

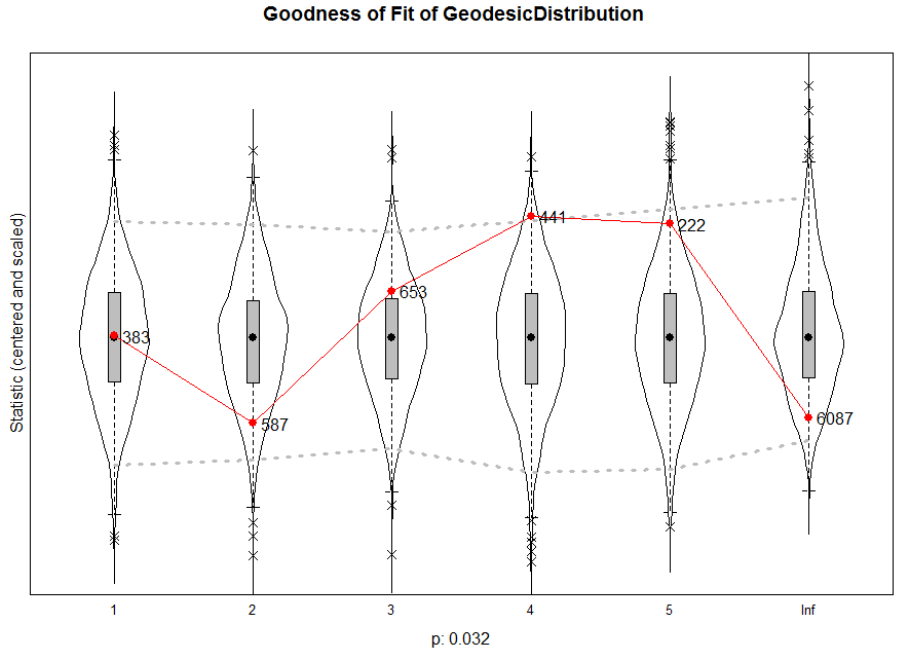


Figure A.4: Goodness of fit for the distribution of distances

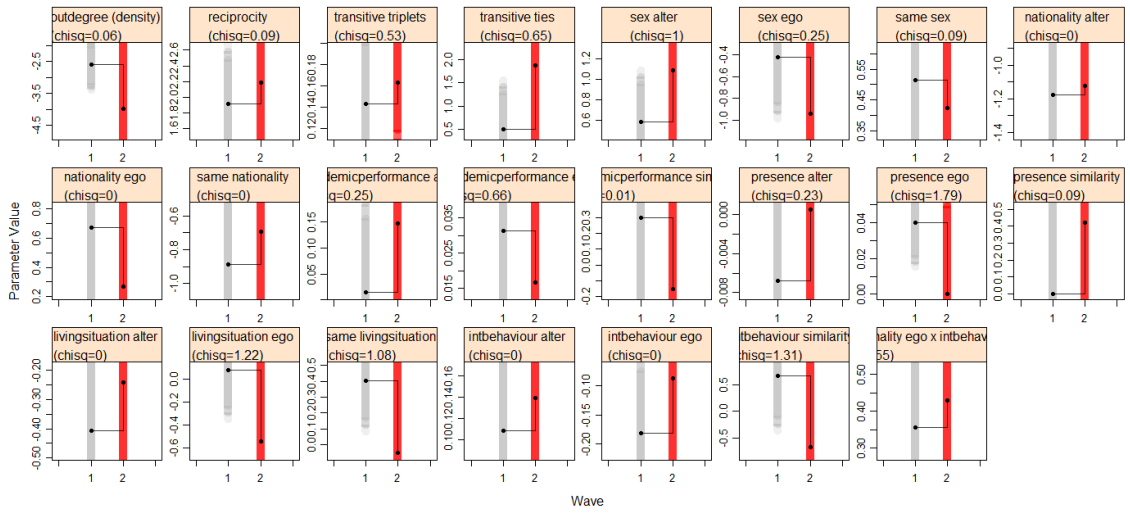


Figure A.5: The tests for time heterogeneity for the network effects

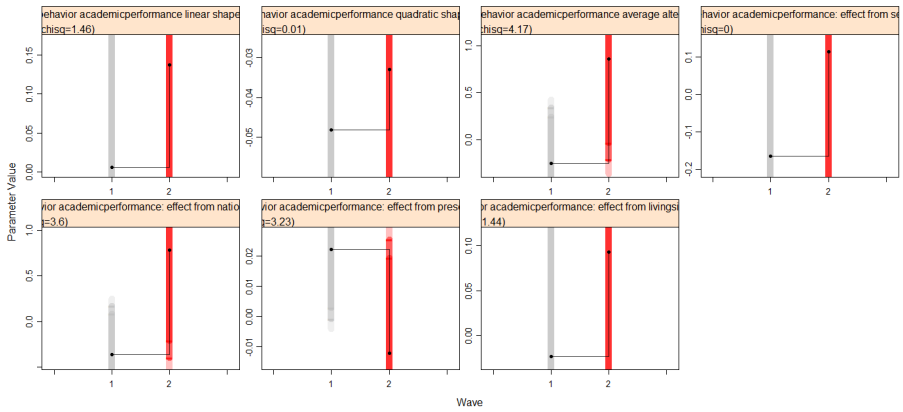


Figure A.6: The tests for time heterogeneity for the behaviour effects

Appendix B

Questionnaire (with consent letter)

Dear first year mathematics student,

My name is Lianne Jansen and I am a Masters student in mathematics. I am currently conducting my 30 EC Master's research project under the supervision of Professor Ernst Wit (Professor of statistics and probability). When considering a research project, we realised that the emergence of the new international bachelor has thrown up some interesting questions. Now more than ever there are students with different backgrounds working together and we wondered whether this influences the (social) behaviour of the students. Do students from different nationalities mix? Are there significant differences in study results? Does the financial situation of the students play a role in how well they perform academically? And how will the relationships between students change over time? Answers to these questions may inspire improvement for the international bachelor.

In order to answer these questions, I have to collect some data, this is the reason why I am asking your assistance by filling out a questionnaire. I will analyse the collected data by using a statistical model: the stochastic actor-based model (Snijders, T.A.B., et al., "Introduction to stochastic actor-based models for network dynamics," *Social Networks*, 2009).

For this study I will ask you to complete a questionnaire three times this academic year. The questionnaire consists of general questions, in order to have some insight in your living situation and background, combined with some questions about your study mates and your leisure time. The questionnaire will take about 15 minutes and we will only ask you to fill in a designated number (not your student number). The list with names with corresponding numbers will be stored in a safe place and separate from the questionnaires. After the third wave of data collection this list will be removed. This will mean that no individual student can be identified and confidentiality will be maintained, none of your answers will be shared with others.

Your participation in this study is voluntary; you are under no obligation to participate and may withdraw at any time. The completed study will be reported in my Master's thesis, but only anonymized data will be used. The data will not be used in a manner which would allow identification of your individual responses. The information provided by you in this questionnaire will be used for research purposes only. If the results are published, only the anonymised data will be available for third parties in order to verify the results.

If you are willing to participate, please sign this form below giving consent for the study. If you can hand the signed form to your mentor, he/she will give you the questionnaire. You are free to ask any questions about the research or about being a participant by e-mail: l.r.jansen@student.rug.nl. If you are interested in receiving the results of the study, you can leave your e-mail address. I will send you the conclusion of the research as soon as it is finished.

Many thanks in advance.

Lianne Jansen, Masterstudent Mathematics, RuG
Ernst Wit, Professor of Statistics and Probability, JBI, RuG (supervising professor)

……

I am willing to participate in the Masters research of Lianne Jansen

Name ……, Date ……., Place ……

Signature ……..

I am interested in receiving the results of the study: yes, my e-mail is …… / no

Questionnaire student research 1

Lianne Jansen
University of Groningen

university of groningen

This questionnaire is part of a study concerning the social behaviour of first year students in the context of the new international bachelor. The study is performed by a master student, associated with the statistics department of the Johann Bernoulli Institute, University of Groningen.

This questionnaire will be treated confidentially. Please do not discuss answers with your fellow students and do not write your name down on the questionnaire. Only write down your number at the start of the questionnaire (this is not your student number). In this way no individual student can be identified.

This questionnaire consists of general questions in order to get insight in your living situation and background. Other questions concern your study mates and your leisure time. Please fill in these questions giving answers that you feel are right for you. Note that there are no wrong answers and that your answers will not be shared with other students or teachers.

Above each question it is explained how to answer the question. Feel free to add comments or explanations. If you pick the wrong answer and you would like to change it, cross out the wrong answer and indicate the right one with an arrow. It will take about 15 minutes to complete the questionnaire, but you can take all the time you need.

General Information

For this question you need the list with names that is also provided to you. Please look up which number you are on the list.

1	What is your number on the list with names? Please write down your own number in the box.	I am number ☐ (please fill in the number)

For the next questions, please check the right answer and/or fill in the answer on the dotted line

2	What is your gender?	☐	Male
		☐	Female

3	What is your age?	I am …….. years old

4	What is your nationality?	……..

5	Do you have any siblings?		Yes
			No

6	In which area have you mostly lived during your life?		In the north of the Netherlands
			In the middle of the Netherlands
			In the south of the Netherlands
			Not in the Netherlands, namely ……

7	What did you do last year?		Last year of high school
			A gap year
			Studied a different subject in Groningen
			Studied a different subject in the Netherlands (Excluding Groningen)
			Studied a different subject not in the Netherlands
			I did something else, namely ……

High school, choice of study and study results

Please check the right answer and/or fill in the answer on the dotted line

8	What description most matches your level at high school?		My grades were mostly excellent and I belonged to the 25% of the students with the best grades
			My grades were just better than the average grade
			My grades were mostly equal to the average grade
			My grades were mostly worse than the average grade
			My grades were quite bad and I belonged to the 25% of the students with the worst grades

9	Why did you choose mathematics in Groningen? (multiple answers possible)		I like the university when I saw it during introduction days and information days
			I think the quality of mathematics in Groningen is high
			It is the nearest university to study mathematics at
			I like the international bachelor course
			My friends are also studying in Groningen
			Else, namely ……

10	Please indicate how many hours you spend in a typical study week on lectures, tutorials and preparation/exercises besides the lectures and tutorials. Note that the answers to this question have absolutely no consequence for you or your study and the answers will not be shared with other students or teachers.	I spend in a typical study week approximately …… hours for lectures …… hours for tutorials …… hours for study, preparation, exercises, assignments etc. besides the lectures and tutorials

11	Did you pass the first basic skills test so that you didn't have to follow remedial teaching?		Yes
			No

12	Please fill in the grades you got in the last examination period. Note that also for this question confidentiality is maintained, so nobody will get access to your grades via this questionnaire.	My grades in the last examination period were: Course: …… Grade: …….. Course: …… Grade: …….. Course: …… Grade: …….. Course: …… Grade: ……..

13	***Please indicate to what extent the following propositions apply to you.***	**None of the time**	**A little of the time**	**Some of the time**	**Most of the time**	**All of the time**
	I speak English during tutorials					
	I speak English during breaks					
	I feel confident at the university					
	During tutorials, I cooperate with my neighbours in order to solve the questions					
	When I cannot solve a question, I ask the teacher for help					

14	***Please indicate to what extent you agree with the following propositions.***	**Strongly disagree**	**Disagree**	**Neither agree nor disagree**	**Agree**	**Strongly agree**
	I think the international character of the bachelor is an advantage to the study					
	I find it hard that everything is in English					
	If it was possible to do a Dutch bachelor mathematics in Groningen, I would prefer that over the international bachelor					
	I like the mathematics bachelor					
	I think I will finish my mathematics bachelor					
	The study is harder than I expected					
	I know where to go when I have study-related problems (for example when I have problems with enrolment for courses or the BSA)					

Leisure time

Please indicate in the table below how many hours per week you spend doing the different activities (approximately). Fill in 0 if you don't spend any time on the activity that is mentioned.

15	**Activity**	**How many hours do you spend per week (approximately)?**
	Sports	
	Other activities in a sport association (for example parties)	
	Activities from the *study* association (note: not student association) like outings and parties	
	Study-related activities from the *study* association like lectures, practise sessions etc.	
	Activities from a *student* association	
	Committee work (include preparation, your own tasks and meetings)	
	Hobbies (other than sports)	
	Socialising/going out with friends	
	(Part time) job	

Personal circumstances

Please check the right answer and/or fill in the answer on the dotted line

16	What is your living situation?		I live in a student house with one or more other students
			I live with my parents
			I live alone in an studio/apartment
			I live together with my partner
			Other, namely ……..

17	Is this living situation the one you wish?		Yes
			No

18	How often do you meet your parents?		Almost every day
			A few times a week
			Approximately once a week
			A few times a month
			Approximately once a month
			A few times a year
			Once a year
			Not applicable

19	How often do you have other ways of contact with your parents? (for example by phone, skype, e-mail etc.)		Almost every day
			A few times a week
			Approximately once a week
			A few times a month
			Approximately once a month
			A few times a year
			Once a year
			Not applicable

20	What is your religion?		Protestant
			Catholic
			Islamic
			Hindu
			Buddhist
			Jewish
			No religion
			Other, namely ……

21	Do you smoke?		Yes, approximately …… cigarettes a ***day / week*** (please encircle the right option)
			No

22	Do you drink alcoholic drinks?		Yes, approximately …… drinks a ***day / week*** (please encircle the right option)
			No

The next question is about the different manners to finance your study time. Various money sources are given. Please use percentages to indicate how you finance your study time. Note that nobody gets to know something about your answers.

For example: I need around 800 euro a month to finance my study time. I get around 300 euro from "studiefinanciering", 200 euro from my parents, 100 euro from my parttime job and I borrow 200 euro each month from the IBG, so I fill in: studiefinanciering 40%, parents 25%, job 10%, borrowing 25%.

You are allowed to approximate, but please check if the total adds up to 100%.

23	***Please indicate to what extend you use from the following manners to finance your study time.***	**Percentage of the total amount that you need each month**
	Money from "studiefinanciering" or a scholarship	
	Borrowing money from the bank or IB-group	
	Support from your parents	
	Money gained from a part time job that you still have	
	Money that you saved or gained from an earlier job	
	Other, namely ……	

The next question is about your feeling <u>of the last month</u>. Please indicate how many times you felt the way that is mentioned by checking the right box.

24	***In the past 30 days how often …***	**None of the time**	**A little of the time**	**Some of the time**	**Most of the time**	**All of the time**
	Did you feel tired out for no good reason?					
	Did you feel nervous?					
	Did you feel so nervous that nothing could calm you down?					
	Did you feel hopeless?					
	Did you feel restless or fidgety?					
	Did you feel so restless that you could not sit still?					
	Did you feel depressed?					
	Did you feel that everything was an effort?					
	Did you feel so sad that nothing could cheer you up?					
	Did you feel worthless?					

Study mates

For this question you need the list with names that is also provided to you. For each question think of which person matches the question and write down the number that corresponds to the person in the box next to the question. You are allowed to fill up all boxes, just a few, one or even none if you don't think that there is a person that matches the question.

The first question is about friends. With "friends" I mean the persons that you like to talk to and share your experiences with. Mostly you do more things together, like going to parties or prepare your homework. So I am looking for the ones that you have a little bit more contact with than with the rest of your study mates.

25	From your fellow students, who are your friends?	☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

26	**Please encircle in the last question who are your closest friends.** ***By closest friends, I mean those persons who know you better than most of the other fellow students do, the persons that you share the most information with, the ones that mean the most for you.***

27	Which of your fellow students do you collaborate with on homework and assignments?	☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

28	With which of your fellow students do you participate in non-study related activities, e.g. going out or having dinner?	☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

29	During lectures, who are the ones that you usually sit next to?	☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

30	Which of your fellow students do you think are very popular?	☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

31	Who of your fellow students do you dislike?	☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

32	Do you have a relationship at this moment?	☐	Yes, with a fellow student, namely ☐ (please fill in the number)
		☐	Yes, but not with a fellow student
		☐	No

33	Besides your fellow students, do you have other friends? If yes, can you mention how many friends you have and how many times you meet these friends?	☐	Yes, I have other friends, approximately …….. (please fill in a number) and I meet them **……** times a ***week / month / year*** (please encircle the right option)
		☐	No, I don't have other friends besides my fellow students

This is the end of the questionnaire. Please check once again if you completed every question.
Thanks a lot for your cooperation!

Appendix C

Data and R-code

All data used for this master research project is stored in csv-files. These data sets contain anonymized data points and these files can be obtained from prof. dr. E.C. Wit (e.c.wit@rug.nl) or L.R. Jansen (liannejansen@live.nl). The questionnaires are securely stored by prof. dr. E.C. Wit until May 2018. The R-code that is used for the analysis of the data is, because of its length, digitally available (again from E.C. Wit or L.R. Jansen).