

# Friendship and Identity in School

## **Data Manual**

Sebastian Pink, Aitana Gräbs Santiago, David Kretschmer,  
Lars Leszczensky, Frank Kalter

Mannheim Centre for European Social Research (MZES),  
University of Mannheim, Germany

May 25, 2023

Version 1.1.0

Citation of the Data Manual: Pink, S., Gräbs Santiago, A., Kretschmer, D., Leszczensky, L. & Kalter, F. (2023): Friendship and Identity in School. Data Manual (Version 1.1.0). German Center for Integration and Migration Research (DeZIM). Berlin.

Data provider: DeZIM.fdz

DOI: 10.34882/dezim.fis.c.1.1.0 (SUF C)

# Table of Contents

Table of Contents .....	2
1. Introduction.....	3
2. Overview .....	4
3. Study description.....	5
3.1. Design.....	5
3.2. Sampling.....	5
3.3. Sample description .....	6
4. Questionnaire.....	9
4.1. Question program .....	9
4.2. Supplements to the questionnaire.....	10
5. Variable names .....	10
5.1. Overall naming scheme .....	10
5.2. Naming conventions for changes between waves.....	11
5.3. Edited variables .....	12
6. Longitudinal data files.....	12
6.1. Identifiers .....	12
6.2. Composition file .....	13
6.3. Survey data file: Questionnaire variables.....	13
6.4. Coding of open questions.....	25
6.5. Edited variables .....	26
6.6. Data anonymization .....	29
6.7. Response rates .....	29
7. Examples of data use.....	32
7.1. Plotting networks.....	33
7.2. Longitudinal network analysis (RSiena): Same-sex homophily.....	35
7.3. Friendship similarity in Stata .....	39
8. References .....	43
8.1. Publications based on the data .....	43
8.2. Field reports .....	44

# 1. Introduction

The study “Friendship and Identity in School,” which we collected this dataset for, aimed to investigate the *interplay between adolescents’ friendship networks and their ethnic identification* (Leszczensky et al. 2014).<sup>1</sup> To this end, the data provide *up to six waves* of longitudinal information for *ten schools* located in the German federal state of North Rhine-Westphalia, which features a high share of students with a migration background. We randomly sampled schools (lower secondary, intermediate secondary, and comprehensive schools) based on three strata of shares of non-native students. Within each school, we surveyed three grades. In the first wave in 2013, we started with the fifth, sixth, and seventh grade, respectively, and followed the students in these grades up to six times with a gap of about nine months between each wave. Written parental consent was mandatory to participate in the survey. The overall sampling population amounts to 2,928 students nested within 29 grades and 95 classrooms, which could have been subject to participation at any wave. Of those students, 2,701 students participated at least once in the survey. Response rates were high, with the wave-specific response rates amounting to almost 85%, on average, and the panel response being close to 90%.

To study the interplay between adolescents’ friendship networks and their ethnic identification, we captured students’ friendship networks via sociometric questions and developed a new measurement to capture ethnic identifications (see Leszczensky & Gräbs Santiago 2014a, 2014b, 2015). Concerning sociometrics, the data provide 29 grade-level networks on directed friendship relations (“Who are your ten best friends?”). Furthermore, we surveyed other kinds of relations such as negative ties, romantic relations, or visiting at home. The measurement of ethnic identifications takes into account different dimensions of national and ethnic identities as well as dual identities. Intense pre-testing showed that the measure is applicable to native and non-native students of different age groups. In addition to sociometrics and ethnic identifications, the data contain information on students’ background, such as demographics, leisure time activities, academic achievement, musical and sports preferences, perceived discrimination, attitudes and opinions on a broad range of topics, religion, the familial situation, opportunity structures for contact, and parents’ opinions. This makes the data suited for a wider range of research questions than the research question that guided the data collection initially.

This data manual seeks to ease researchers’ usage of the data to carry out their own empirical analyses. To this end, we will first explain the study in more detail, including design (3.1), sampling (3.2) and the resulting sample (3.3). Thereafter, we describe the questionnaire program (4.1) and supplements to the questionnaire (4.2). In the fourth section, we explain the variable naming scheme. We use the question program from the first wave as a basis for naming the variables (5.1.) and explain how variables introduced in later waves are included (5.2). Furthermore, we explain the edited variables we created to further ease working with the data (5.3). After that, we discuss the structure of the longitudinal data files provided (6.). First, we show the identifiers for schools, students, grades, and so on (6.1). Next, the composition file is discussed, which allows to specify network boundaries over time (6.2). Thereafter, we provide an overview of the questions and the link between questions in the questionnaire and variable names (6.3). Moreover, we explain the coding of open items (6.4) and describe edited variables (6.5). We also show response rates over time at both the grade and the classroom level (6.6), which may support in selecting networks for

---

<sup>1</sup> For more information, also see the project’s webpage: <http://www.mzes.uni-mannheim.de/d7/en/projects/friendship-and-identity-in-school>

empirical analyses. In the last section (7.), we illustrate typical use case scenarios for the data with examples focusing on social network analysis. We demonstrate how networks can be plotted (7.1) and how to analyze networks longitudinally using stochastic-actor oriented models (7.2) in R. Finally, we show how the network data may be used in Stata (7.3).

## 2. Overview

For a quick orientation, Table 1 provides a brief overview of the most important data and working features.

Table 1: Project overview

Title	Friendship and Identity in School
Institution responsible	Mannheim Centre for European Social Research (MZES)
Survey institutes	None
Funding	German Research Foundation (DFG)
Project team	Frank Kalter, David Kretschmer, Lars Leszczensky, Sebastian Pink
Survey population	5 <sup>th</sup> to 7 <sup>th</sup> grade students in upper secondary comprehensive, intermediate secondary, and lower secondary schools located in the German federal state North Rhine-Westphalia with an elevated share students with foreign, especially Turkish, citizenship
Survey method	PAPI
Survey period	10.04.2013 to 14.12.2017
Survey sample	Random sample of 10 schools
Selection process	All 5 <sup>th</sup> , 6 <sup>th</sup> , and 7 <sup>th</sup> grade students within one school
Survey documents	Standardized questionnaire
Interview duration	Up to two school lessons
Waves	6
Total sample	$n = 11,221$ ( $N = 2,928$ students)
Interviews carried out	$n = 9,292$ ( $N = 2,701$ students)
Response rate	82.8 percent interviews, 92.2 percent students
Citation of the data	Leszczensky, L., Pink, S., Kretschmer, D., & Kalter, F. (2023). Friendship and Identity in School. Dataset version: 1.1.0. German Center for Integration and Migration Research (DeZIM). Berlin. DOI.
DOI – Data access	10.34882/dezim.fis.c.1.1.0 (SUF C)

### **3. Study description**

#### **3.1. Design**

The main goal of the study “Friendship and Identity in School” was to investigate the interplay of adolescents’ social networks and their ethnic identification. Because a ready-to-use German-language measure of ethnic identity was not available at the beginning of the project in 2012, we constructed an accurate measurement of ethnic identity before the actual data collection. Drawing on previous research, we proposed a new measure and conducted cognitive pretests to assess the comprehensibility of the items (see Leszczensky & Gräbs Santiago 2014a). We selected the approved items for the measurement, which captured key dimensions of both national and ethnic identities as well as the presence of dual identities. A primary study supported the applicability of our measurement to children and adolescents. A final test based on data of the first wave demonstrated the validity of the measurement (see Leszczensky & Gräbs Santiago 2014b, 2015). We also ascertained invariance of the measure across native and non-native students, different migration generations and age groups.

The second central aspect of the study are the social networks of adolescents. Adolescents spend a substantial amount of their time at school and form significant social ties to peers in this context. Previous research also indicates the influence of social networks in schools on attitudes and behavior (see Veenstra et al. 2013 for a recent review). We therefore decided to collect network data in this specific context and capture a variety of adolescents’ social ties, such as their friends, but also negative ties such as bullying. Networks were assessed at the grade level, meaning that students could indicate social ties to all members of their own grade. In comparison to networks assessed at the classroom level, this has the advantage of providing more information about the students and, therefore, allowing the specification of more complex multivariate network models, which need large amounts of data to be estimated (see Leszczensky & Pink 2015; Valente et al. 2013).

To disentangle the mechanisms underlying the formation and change of social networks it is crucial to observe both networks and individual attributes repeatedly at several points in time. This longitudinal design raises the question of the optimal amount of time between subsequent surveys. On the one hand, the distance between subsequent assessments has to be short enough to obtain precise and comparable information over time; on the other hand, it has to be long enough to allow for meaningful change in both networks and individual characteristics (Snijders et al. 2010: 49 f.). For this reason, we planned the survey as a panel with a gap of about nine months in between subsequent wave (see Leszczensky et al. 2014 for details).

#### **3.2. Sampling**

As one of this study’s core goals was to study ethnic identification, the sample had to contain a sufficient number of students with a migration background. Our focus is on students of Turkish origin, who constitute the largest and the most disadvantaged immigrant group in Germany. To take different opportunity structures for interaction among students into account, we needed to include schools with different shares of Turkish students in the sample. For this purpose, we constructed three strata: schools with more than 15% students with Turkish citizenship, schools with 10 to 15% students with Turkish citizenship, and schools with at least 15% students with foreign citizenship, but less than 5% students with Turkish citizenship (see Leszczensky et al.

2014). Due to a high share of students with a migration background and organizational advantages, all of the data collection took place in the German federal state of North Rhine-Westphalia.

To capture the influence of students' social background, we included different types of schools in the sample. In the relevant age range, the most important school types in the German context are upper secondary (*Gymnasium*), comprehensive (*Gesamtschule*), intermediate secondary (*Realschule*), and lower secondary schools (*Hauptschule*). However, due to an overall low share of students with a migration background in upper secondary schools, we restricted the sampling procedure to lower secondary, intermediate secondary and comprehensive schools. School type and ethnic composition define a two-dimensional space to sample schools from (see Table 2). As a final restriction of the pool of schools to sample from, we had to account for the fact that we intended to sample grade-level networks. These networks have to be of moderate size in order to be sure that students within a grade are aware of each other. We therefore restricted sampling to schools with reported grade sizes of 45 to 120 students. Conditional on this restriction, we randomly sampled one school from each of the cells spanned by the two-dimensional strata (see Table 2). The school response rate was 10%. The most important reason for schools' refusal was that many schools were already participating in other school-based studies for which there has been a huge increase in Germany in recent years. More information on sampling can be found in the study's field report (Leszczensky et al. 2015).

Table 2: Schools sampled

Strata	Lower secondary	Intermediate secondary	Comprehensive schools	Total
>15% Turkish	1 ( <i>+1</i> )	1	1	3
10-15% Turkish	1	1	1	3
>15% foreign; <5% Turkish	1	1	1	3
Total	3 (4)	3	3	9 (10)

After the first three waves of data collection had been completed and a follow-up study had been approved by the German Research Foundation (DFG), we attempted to sample additional schools from the strata defined above. From the 55 schools contacted for this refreshment sample, one school agreed to participate in the study. This school is another lower secondary school with the share of students with Turkish citizenship exceeding 15%, which is marked in italics in Table 2. Data collection from that school is comparable to the procedures applied in the other schools but started in parallel with the fifth wave in the remaining schools (see below for more information).

### 3.3. Sample description

Information on the sample over time is provided in Table 3. For the nine schools that participated in the study from the beginning, the first wave of data was collected in April and May 2013. In the first wave, 2,185 students were registered in the fifth, sixth and seventh grade of the nine sampled schools. On average, a grade contained 84, and a class contained 26 students. To participate in the study, students had to submit written consent from their parents. As an incentive, every participating student received five Euro after completing the questionnaire. The students filled

out paper-and-pencil questionnaires during two school lessons in about 90 minutes. In the first wave, 1,668 students participated in the survey, yielding an overall participation rate of 76.3%. The students were between 10 and 17 years old, and the average age was 12.6 years. 63.3% of the respondents had a migration background, i.e., they themselves or at least one of their parents or grandparents were born abroad.

Table 3: Panel overview

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6
<i>Initial School Sample (9 schools)</i>						
Period of data collection	04-05 2013	01-02 2014	10-12 2014	09 2015	04-06 2016	02-03 2017
Schools	9	9	9	6	6	5
Grades	26	26	26	18	18	10
Classes	86	83	85	57	57	35
Students	1,668	1,860	1,920	1,334	1,249	811
Response	76.3%	82.7%	86.5%	88.4%	82.1%	87.8%
Panel response*		90.9%	91.0%	90.3%	84.3%	90.2%
<i>Refreshment School Sample (1 school)</i>						
Period of data collection	04 2016	03 2017	12 2017			
Schools	1	1	1			
Grades	3	3	3			
Classes	8	9	9			
Students	132	166	152			
Response	66.7%	74.8%	78.8%			
Panel response		88.5%	86.5%			
<i>Entire School Sample (10 Schools)</i>						
Students surveyed across all waves and all schools	At least once 2,701	At least twice 2,295	At least three times 1,871	At least four times 1,185	At least five times 872	Six times 368

\* Panel response is corrected for schools no longer participating in the study.

The second wave of data was collected in January and February 2014. In total, 2,250 students were registered in the relevant grades of the nine participating schools at this time. 1,860 students answered the questionnaire. Thus, the overall participation rate was 82.7%. The panel mortality was very low, with 90.9% of the students from the first wave also participating in the second wave.

The third wave of data collection took place in October, November, and December 2014. In wave 3, 1,920 students participated, while 2,219 were enrolled in the relevant grades, yielding a response rate of 86.5%. Again, the panel mortality from the second to the third wave was very low, with 91.0% of the students from the second wave participating in the third wave as well.

Table 4: School participation across waves

School	W1	W2	W3	W4	W5	W6	Note
1	X	X	X	X	X	X	original grade 7 has left school in W6
2	X	X	X	X	X		
3	X	X	X				
4	X	X	X				
5	X	X	X	X	X	X	original grade 7 has left school in W6
6	X	X	X	X	X	X	original grade 7 has left school in W6
7	X	X	X	X	X	X	original grade 7 has left school in W6
8	X	X	X	X	X	X	original grade 7 has left school in W6
9	X	X	X				
10	X	X	X				W1 surveyed in parallel to W5 in other schools

After wave three, a number of schools did not continue to participate in the study.<sup>2</sup> School-level participation across the six waves is illustrated in Table 4: after the third wave, data collection ended in four of the ten schools. In the remaining five schools, the fourth wave of data was collected in September 2015, with 1,334 of the 1,509 students enrolled in the relevant grades participating. This corresponds to a response rate of 88.4% and a panel response rate of 90.3%. The fifth wave of data was collected from April to June of 2016, with 1,249 of the 1,521 students enrolled in the sampled grades participating, corresponding to a response rate of 82.1% and a panel response of 84.3%. Of the six schools remaining from the initial sample, one school left the study after wave 5. Furthermore, as illustrated in Table 3, the initial grade 7 had left school by the time data for the sixth wave of data was collected. Therefore, only 924 students were eligible for participation in the sixth wave, which was conducted in February and March 2017. 811 students participated, yielding a response rate of 87.8% and a panel response rate of 90.2%.

As illustrated in Table 4, a 10<sup>th</sup> school started to be surveyed in parallel to the fifth wave of data collection in the remaining schools. In this school, students from the sixth, seventh, and eighth grade were sampled in the first wave (rather than students from the fifth, sixth, and seventh grade, as in the remaining nine schools). Results for this school are displayed in the middle panel of Table 3. The school was first surveyed in April 2014, with 132 of the 198 eligible students participating. This corresponds to a response rate of 66.7%. In the second wave of data collection, conducted in March 2017, 166 of the 222 eligible students participated, which amounts to a response rate of 74.8% and a panel response of 88.5%. Finally, the last wave of data was collected in December 2017. The response rate was 78.8%, with 152 of 193 eligible students participating. This amounts to a panel response of 86.5% between the second and the third wave.

The lower panel of Table 3 displays how many students participated in the survey across waves when aggregating across all of the ten schools. Overall, there is information on 2,701 students, with 2,295 students having participated at least twice and 1,871 students having participated at least three times. Given that more than three waves of data were only collected in six of the schools, there are fewer respondents who provide data across four or more waves: 1,185 students participated at least four times, 872 participated at least five times, and 368 students participated in all six waves.

<sup>2</sup> The main reason for the three schools dropping out of the survey was that they had no interest in providing further information. Initially, all nine schools had been asked to participate three times.



## 4. Questionnaire

### 4.1. Question program

Given the longitudinal design of the study, in which respondents were repeatedly surveyed, they were asked largely the same set of questions in all waves. The questions from the first wave's questionnaire (as well as their numbering) served as the basis for all subsequent questionnaires (and, as outlined later, the naming scheme within the dataset). The question program is organized in seven blocks A to G, which tie together thematically similar questions. The following provides a concise résumé of the topics surveyed in the first wave.

#### A

- sociometry (e.g., best friends, meetings in leisure time, bullying)
- friendship-related
- demographics
- academic achievement
- leisure
- music and sports
- school subjects

#### B

- national identity
- conditions for belonging to Germany
- country of birth of student, parents and grandparents (i.e., third generation migration background)

#### C Only for those *with* a migration background

- country of origin
- ethnic identity
- external categorization
- perceived discrimination and permeability
- acculturation
- ethnic identity of parents
- country of birth of grandparents (specific)

#### D Only for those *without* a migration background

- acculturation
- contact to foreigners
- proud of Germany
- attitudes towards foreigners
- perceived discrimination and permeability
- ethnic origin of friends

#### E

- religion and intensity of religiosity
- religious identity
- perceived discrimination
- religious identity of parents

#### F

- language
- language skills and use

## G

- neighborhood
- attitudes toward different groups
- national identity of parents
- influence of parents on friendships
- parental contact
- parents' occupation
- living situation
- family characteristics
- pocket money
- cultural capital
- elementary school
- city district
- grade-level course structure
- social trust
- personality traits (reduced version of Big Five)

The vast majority of the questions are repeated each wave. However, minor variations occur between waves as some questions are dropped and others are added. Section 5.3 provides a comprehensive overview of the items of the respective questions, tracing their occurrence across waves and linking them to their variable names. On the basis of this information, the variables can be found in the accompanying dataset, as elaborated in Section 4.

### 4.2. Supplements to the questionnaire

To capture social ties (such as friendship ties) between students at the grade-level, a list of all of the students' schoolmates in their respective grade accompanied each questionnaire. For confidentiality, students did not write down the names of their peers when nominating them, but used generic numbers, which we assigned to each student's name on the list. These grade-level lists were separated visually by the grade's different classrooms, showing each classroom's students in a separate column.

Furthermore, the survey asked students for the district of the city they live in. Therefore, each questionnaire entailed a list of the districts in the respective city, each with its own number. Students then identified the number of the district they live in.

## 5. Variable names

### 5.1. Overall naming scheme

Variable names in the data generally mirror the questionnaire structure from the first wave. As discussed above, the questionnaire consists of seven topic-specific blocks A to G. Within every block, items are numbered consecutively. In the case of questions with more than one item, the items are numbered sequentially within the question.

A variable name consists of up to four elements: the topic, an item number, an optional suffix for an additional item sub-number and further information, and an optional prefix to denote edited variables (see 4.2). Table 5 shows some examples.

The first element of the variable name refers to the **topic**, with letters a-g in the variable name referring to questionnaire blocks A-G, respectively. The topic indicator is followed by the **item** number within the respective block, which generally contains up to two digits. The exception are the sociometric questions, which consist of four digits. For the sociometric questions, the first two digits refer to the item number in the questionnaire and the second two digits refer to the different nominations given for each item number. For example, students could nominate up to ten friends, such that, for this sociometric question, the third and fourth (i.e., the last two) digits identify the first nomination, the second nomination, and so on.

An optional suffix in the variable name is separated from the previous elements by an underscore. The suffix indicates an additional **item** sub-number in the case of questions with several items. Furthermore, it displays information on the coding of the variable. Variables originally measured in string format (i.e., open questions asking respondents to write down short texts) are coded specifically, which is indicated by the **\_coded** suffix. Questions with multiple choices in the response categories are coded as individual dummy variables for each response category, and are indicated by the suffix **\_mc**. The suffixes **\_w1** and **\_w2** show that in wave 1 and wave 2 different scales are used in the respective variable, reflecting a change in the question design across waves (analogous suffixes are used for change between other waves). An optional prefix denotes edited (**ed\_**) variables, which we use to provide compound information based on already available information from the questionnaires. The suffix **\_grouped** indicates the aggregation of values as part of the overall anonymization procedures for the download version of this dataset. Ungrouped variables featuring higher resolution are only available for remote and on-site use at the DeZIM in Berlin.

Table 5: Examples for the variable naming scheme

b2_1	block <b>B</b> , question <b>2</b> in the block, item <b>1</b> within the question
a1_0104	block <b>A</b> , question <b>1</b> in the block, nomination <b>4</b> in a sociometric question
a8_1_w1	block <b>A</b> , question <b>8</b> in the block, item <b>1</b> within the question, version of question used in wave <b>1</b>
a12_mc_1	block <b>A</b> , question <b>12</b> in the block, answer <b>1</b> in a multiple choice question
b4_coded	<b>coded</b> version of the string variable b4, denoting the 4 <sup>th</sup> question in block B
ed_a6	<b>edited version</b> of variable a6
a7_2_group	<b>grouped version</b> of variable a7_2 due to anonymization

## 5.2. Naming conventions for changes between waves

Though much of the questionnaire remained unchanged across waves, there are some changes due to questions being dropped or added over time. Consequently, the question number of specific questions within the thematically structured block on the physical questionnaire changed over time. Variable *names*, however, have to be constant over time. Thus, even though a question's position in the questionnaire may change over time, its variable name does not, with the name most likely pertaining to the question's position in the questionnaire in the *first wave*. Thus, the eighth question in block A in the first wave (A8) carries the variable name a8 in all waves, although this question becomes question A7 in the physical questionnaire in later waves because of changes in other questions.

New questions, in the sense of questions that had not been asked in the first wave, are given variable names that put them at the end of the block. For example, in the third wave, we introduced

an open question about the favorite school subject as question A8. However, in the longitudinal dataset this question’s variable name is `a14_coded`. This is because prior to the third wave, there were already 13 questions with associated variable names in section A, the last being `a13_coded_5`.

Because a variable’s position in the questionnaire can vary across waves while its variable name cannot, dissociations between position in the questionnaire and name occur. Table 6 in section 5.1 provides a comprehensive overview of the variables used in the survey and provides information on (1) the variables’ (time-constant) names, (2) whether the variable was surveyed in any given wave, and (3) the variable’s number in the respective wave’s questionnaire. Note that dissociations between variable name and position are only relevant for identifying variables from specific physical questionnaires but not for data analysis per se.

### 5.3. Edited variables

The data contain *edited variables* that seek to make it easier for researchers to work with the data. As indicated already earlier, these variables start with the prefix **ed\_** and provide information about a wide range of topics. For example, we created a variable that indicates our best guess of the students’ country of birth, which was measured in all waves even though the information itself is time-constant. Furthermore, we constructed several variables referring to ethnic origin, including information from different variables, such as countries of birth of the students as well as their parents and grandparents (for a detailed description of the variables see section 5.4). In addition, we provide variables capturing the size of the class and the grade as well as the response rate at the grade-and classroom-level to ease sample selection for network analyses.

## 6. Longitudinal data files

The full data from the study consists of two data sets, the contents of which we describe in the sections below. On the one hand, we provide the longitudinal survey data, which contains all survey data from the questionnaires of all waves. Second, we provide *composition data*, which provides complete information on the student composition of grades and classrooms at any wave. This file is particularly relevant for social network analysis. In the Stata data files, users may switch between German and English variable and value labels typing “label language de” or “label language en”.

### 6.1. Identifiers

Due to the longitudinal and multi-level structure of the data, both data sets contain multiple identifiers of persons and entities (e.g., grades and classrooms) as shown in Table 6.

Table 6: Identifiers

Variable	Identifies ...	Across waves	Values
<b>id_s</b>	a school uniquely	constant	1,...,10
<b>id_g</b>	a grade uniquely	constant	1,...,29
<b>grade_qs</b>	a grade within a school in a wave	varying	5,...,10
<b>id_c</b>	a class within a grade in a wave	varying	1,...,5
<b>id_p</b>	a student (i.e., pupil) uniquely	constant	1001,...,10250

Only two of these identifiers are not constant across waves. These identifiers are provided to allow for analyses that are grade- (i.e., academic year) specific, as in an analysis that focuses on sixth-graders only. For this scenario, `grade_qs` can be used.

## 6.2. Composition file

A core goal of the data was to enable social network analysis of grade-level student networks. For network analysis (based on graph theory), it is necessary to have information on the entire pool of individuals who are part of the network (i.e., the classroom or grade context), not only on those participating in the network study. The longitudinal survey data file, however, only provides information on the students who participated in the survey (and the students they nominated in the sociometric questionnaire). The composition file complements this with information from the grade lists (the lists students used to nominate their peers), which provide full information on all students that have been part of any grade-level network in any given wave. This information can be used to both specify network boundaries and to differentiate non-participation of students in a given wave into subcategories, such as unit non-response and absence from networks (as observed for students switching between classrooms, grades, or schools over time), which then may be accounted for in the statistical models (Huisman & Snijders 2003).

Why is the composition file so important for longitudinal social network analysis? In very general terms, the set of actors of one grade subject to statistical analysis are all students that at least once have been part of that grade in any of the waves considered. The ties between students are captured by so-called adjacency matrices, one for each wave considered. In other words, if a longitudinal network analysis for a grade-level friendship network of a specific grade is supposed to be carried out for the first three waves, then the dimensionality (i.e., the number of rows and columns) of each adjacency matrix has to equal the number of students who at least once have been part of the specific grade in either of these three waves. In case students left or entered that grade-level network at any point, their rows and columns are to be marked as missing after or before leaving or entering. This granularity of the data is referred to as *complete information* in the context of longitudinal social network analysis, which is a precondition for a proper analysis. Without the composition file, meeting this requirement would not be possible. In the example data analysis discussed below (6.), we demonstrate how the information from the composition data can be used in cross-sectional and longitudinal social network analyses.

## 6.3. Survey data file: Questionnaire variables

The second data file contains the survey data from all students and all waves. In this section, we provide an overview in Table 7 of all variables included in the data set, displaying their *names* and their positions in the wave-specific questionnaires. If no information on the wave-specific position in the questionnaires is provided (i.e., if any of the columns 2-7 of Table 7 are empty), the corresponding item was not surveyed in that wave. The last column provides a brief description of the variable.

Table 7: Questionnaire variables

Variable name	W1	W2	W3	W4	W5	W6	Label
a1_0101	A1	A1	A1	A1	A1	A1	network: best friends
a1_0102							
a1_0103							
a1_0104							
a1_0105							
a1_0106							
a1_0107							
a1_0108							
a1_0109							network: friends from elementary school
a1_0110							
a1_0201	A1	A1				A1	
a1_0202							
a1_0203							
a1_0204							
a1_0205							
a1_0301	A1	A1	A1	A1	A1	A1	network: don't like at all
a1_0302							
a1_0303							
a1_0304							
a1_0305							
a1_0306							
a1_0307							
a1_0308							
a1_0309							network: best friend
a1_0310							
a1_0401	A1	A1	A1	A1	A1	A1	
a1_0501	A1	A1	A1	A1	A1	A1	network: most popular
a1_0502							
a1_0503							
a1_0504							
a1_0505							
a1_0601	A1	A1	A1	A1	A1	A1	network: meet in leisure time
a1_0602							
a1_0603							
a1_0604							
a1_0605							
a1_0606							
a1_0607							
a1_0608							
a1_0609							network: bully (sender)
a1_0610							
a1_0701	A1	A1	A1	A1	A1	A1	
a1_0702							
a1_0703							

Variable name	W1	W2	W3	W4	W5	W6	Label
a1_0704							
a1_0705							
a1_0801	A1	A1	A1	A1	A1	A1	network: romantic relationship
a1_0901	A1	A1	A1	A1	A1	A1	network: visit at home
a1_0902							
a1_0903							
a1_0904							
a1_0905							
a1_1001	A1	A1	A1	A1	A1	A1	network: bully (receiver)
a1_1002							
a1_1003							
a1_1004							
a1_1005							
a1_1101	A1	A1	A1	A1	A1	A1	network: talk about problems
a1_1102							
a1_1103							
a1_1201	A1	A1	A1	A1	A1	A1	network: write messages
a1_1202							
a1_1203							
a1_1204							
a1_1205							
a1_1206							
a1_1207							
a1_1208							
a1_1209							
a1_1210							
a2	A2	A2	A2	A2	A2	A2	localization: majority of friends
a3_1	A3	A3	A3	A3	A3	A3	preference for friends: same hobbies
a3_2							preference for friends: same musical taste
a3_3							preference for friends: same country of origin
a3_4							preference for friends: same sex
a3_5							preference for friends: same friends
a3_6							preference for friends: same religion
a4	A4	A4	A4	A4	A4	A4	preference for friends: parents like friends
a5	A5	A5	A5	A5	A5	A5	out of school friends: from Germany
a6	A6	A6	A6	A6	A6	A6	sex
a7_1	A7	A7	A7		A7		birthday: day
a7_2							birthday: month
a7_3							birthday: year
a8_1_w1	A8						school grade: math
a8_2_w1							school grade: German
a8_3_w1							school grade: English
a8_1_w2		A8	A10	A7	A8	A7	school grade: math
a8_2_w2							school grade: German
a8_3_w2							school grade: English

Variable name	W1	W2	W3	W4	W5	W6	Label
a9_1	A9	A9	A12	A9	A10	A9	leisure time: reading books
a9_2							leisure time: going to youth center
a9_3							leisure time: partying
a9_4							leisure time: family activities
a9_5							leisure time: going to the movies
a9_6							leisure time: smoking cigarettes
a9_7							leisure time: hanging out with friends
a9_8							leisure time: spending time in a club
a9_9							leisure time: drinking alcohol
a10_w1	A10	A10					ownership of smartphone
a10_w3			A13	A10	A11	A10	
a11_1	A11	A11	A14	A11	A12	A11	spending time: watching TV
a11_2							spending time: internet, chatting, social networks
a11_3							spending time: doing homework
a11_4							spending time: helping in household
a11_5							spending time: video and computer games
a11_6							spending time: taking care of relatives
a12_mc_1	A12	A12	A15	A12	A13	A12	music: electronic music (e.g., House, Techno, or Dubstep)
a12_mc_2							music: HipHop, Rap (also R'n'b)
a12_mc_3							music: Jazz, Blues (also Classic)
a12_mc_4							music: Pop (e.g., charts)
a12_mc_5							music: Rock (also Metal, Punk or Indie)
a12_mc_6							music: other
a12_coded_1							music: other (coded) nomination 1
a12_coded_2							music: other (coded) nomination 2
a13_mc_1	A13	A13	A16	A13	A14	A13	Sports: not on a regular basis
a13_mc_2							sports: soccer
a13_mc_3							sports: swimming
a13_mc_4							sports: basketball
a13_mc_5							sports: gymnastics
a13_mc_6							sports: cycling
a13_mc_7							sports: inline-skating
a13_mc_8							sports: martial arts
a13_mc_9							sports: dancing
a13_mc_10							sports: other
a13_coded_1							sports: other (coded) nomination1
a13_coded_2							sports: other (coded) nomination 2
a13_coded_3							sports: other (coded) nomination 3
a13_coded_4							sports: other (coded) nomination 4
a13_coded_5							sports: other (coded) nomination 5
a14_coded			A8				school subject: like most
a15_coded			A9				school subject: like least
a16			A11	A8	A9	A8	friends among themselves: importance of liking each other



Variable name	W1	W2	W3	W4	W5	W6	Label
a17_coded_1			A13	A10	A11	A10	smartphone (used) (coded)
a17_coded_2							smartphone (new) (coded)
b1_1	B1	B1	B1	B1	B1	B1	national identity: importance
b1_2							national identity: private regard: be satisfied
b1_3							national identity: private regard: be glad
b2_1	B2	B2	B2	B2	B2	B2	national identity: attachment: troubles me, if sb. speaks ill
b2_2							national identity: attachment: close to my heart
b2_3							national identity: attachment: feel strongly attached
b2_4							national identity: attachment: feel as part
b3_1	B3	B3	B3	B3	B3	B3	condition for belonging to Germany: born in Germany
b3_2							condition for belonging to Germany: speaking German
b3_3							condition for belonging to Germany: feeling German
b3_4							condition for belonging to Germany: German parents
b3_5							condition for belonging to Germany: respecting rules
b4	B4	B4	B4	B4	B4	B4	country of birth: student
b4_coded							country of birth: student: other (coded)
b5	B5	B5	B5	B5	B5	B5	age of migration
b6_mc_1	B6	B6	B6	B6	B6	B6	citizenship: German
b6_mc_2							citizenship: Turkish
b6_mc_3							citizenship: Italian
b6_mc_4							citizenship: Polish
b6_mc_5							citizenship: other
b6_mc_6							citizenship: don't know
b6_coded							citizenship (coded)
b7	B7	B7	B7	B7	B7	B7	country of birth: mother
b7_str							country of birth: mother (other)
b7_coded							country of birth: mother (coded)
b8	B8	B8	B8	B8	B8	B8	country of birth: father
b8_str							country of birth: father (other)
b8_coded							country of birth: father (coded)
b9_1	B9	B9	B9	B9	B9	B9	country of birth: grandmother (maternal)
b9_2							country of birth: grandfather (maternal)
b9_3							country of birth: grandmother (paternal)
b9_4							country of birth: grandfather (paternal)
c1_coded	C1	C1	C1	C1	C1	C1	family's country of origin (coded)
c2_1	C2	C2	C2	C2	C2	C2	ethnic identity: importance
c2_2							ethnic identity: private regard: be satisfied
c2_3							ethnic identity: private regard: be glad

Variable name	W1	W2	W3	W4	W5	W6	Label
c3_1	C3	C3	C3	C3	C3	C3	ethnic identity: attachment: troubles me, if sb. speaks ill
c3_2							ethnic identity: attachment: close to my heart
c3_3							ethnic identity: attachment: feel strongly attached
c3_4							ethnic identity: connection: feel part of
c4	C4	C4	C4	C4	C4	C4	self-categorization: closed
c5_1	C5	C5	C5	C5	C5	C5	categorization by others: Germans
c5_2							categorization by others: people from country of origin in Germany
c5_3							categorization by others: people from country of origin in country of origin
c6_1	C6	C6	C6	C6	C6	C6	ethnic identity: public regard: respect
c6_2							ethnic identity: public regard: like
c6_3							ethnic identity: public regard: positive view
c7	C7	C7	C7	C7	C7	C7	self-categorization: half-open
c7_coded							self-categorization (coded)
c8_1	C8	C8	C8	C8	C8	C8	dual identity: being both, German and from country of origin
c8_2							dual identity: sometimes German, sometimes from country of origin
c9_1	C9	C9	C9	C9	C9	C9	discrimination: country of origin: spoken badly
c9_2							discrimination: country of origin: insulted or offended
c9_3							discrimination: country of origin: treated badly or unfair
c10_1	C10	C10	C10	C10	C10	C10	permeability: difficult to be seen as a German
c10_2							permeability: impossible to be seen as a German
c11	C11	C11	C11		C11	C11	acculturation: importance of preserving German customs and traditions
c12	C12	C12	C12		C12	C12	cooking dishes from country of origin
c13	C13	C13	C13		C13	C13	visiting country of origin
c14	C14	C14	C14	C11	C14	C14	out of school friends: country of origin
c15_1	C15	C15	C15	C12	C15	C15	parents: talking about history of country of origin
c15_2							parents: contact with a person from country of origin
c15_3							parents: watch tv-shows of films from country of origin
c16_1	C16	C16	C16			C16	parents: ethnic identity: troubles them, if sb. speaks ill
c16_2							parents: ethnic identity: close to their heart
c16_3							parents: ethnic identity: feel as part
c16_4							parents: have lots of German friends
c16_5							parents: speak German well

Variable name	W1	W2	W3	W4	W5	W6	Label
c16_6							parents: advocate having friends from the country of origin
c17		C17	C17		C16	C17	country of birth: grandparents: grandmother (paternal)
c17_coded							country of birth: grandparents: grandmother (paternal, coded)
c18		C18	C18		C17	C18	country of birth: grandparents: grandfather (paternal)
c18_coded							country of birth: grandparents: grandfather (paternal, coded)
c19		C19	C19		C18	C19	country of birth: grandparents: grandmother (maternal)
c19_coded							country of birth: grandparents: grandmother (maternal, coded)
c20		C20	C20		C19	C20	country of birth: grandparents: grandfather (maternal)
c20_coded							country of birth: grandparents: grandfather (maternal, coded)
d1_1	D1	D1	D1	D1	D1	D1	acculturation: Germans should save customs/traditions
d1_2							acculturation: foreigners should adapt to Germany
d1_3							acculturation: Germans should be open for customs/traditions of foreigners
d1_4							acculturation: foreigners should do everything to preserve their customs/traditions
d2	D2	D2	D2	D2	D2	D2	romantic relationship: potential of partner with migration background
d3_1	D3	D3	D3	D3	D3	D3	contact to foreigners: like talking
d3_2							contact to foreigners: like spending time
d3_3							contact to foreigners: glad getting to know
d3_4							contact to foreigners: don't care from which country a person comes
d4_1	D4	D4	D4	D5	D5	D5	Proud of Germany: economic success
d4_2							Proud of Germany: success in sports
d4_3							Proud of Germany: cultural success
d4_4							Proud of Germany: armed forces
d4_5							Proud of Germany: educational system
d5	D5	D5	D5	D4	D4	D4	acculturation: importance of preserving German customs and traditions
d6_1	D6	D6	D6			D6	attitudes towards foreigners: commit criminal offences more frequently
d6_2							attitudes towards foreigners: can be trusted like Germans
d6_3							attitudes towards foreigners: should go back to their countries of origin
d6_4							attitudes towards foreigners: should be treated like Germans

Variable name	W1	W2	W3	W4	W5	W6	Label
d7_1	D7	D7	D7	D6	D6	D7	Germany: public regard: respect
d7_2							Germany: public regard: like
d7_3							Germany: public regard: positive view
d8_1	D8	D8	D8	D7	D7	D8	discrimination: Germany: spoken badly
d8_2							discrimination: Germany: insulted or offended
d8_3							discrimination: Germany: treated badly or unfair
d9_1	D9	D9	D9	D8	D8	D1	permeability: can be German, even if born in other country
d9_2							permeability: can be German, if born in Germany, but foreign parents
d10_1	D10	D10					number of best friends: total
d10_2							number of best friends from country other than Germany: total
d11_1	D11	D11					number of friends: total
d11_2							number of friends from country other than Germany: total
e1_w1	E1						religion
e1_w2		E1	E1	E1	E1	E1	
e1_coded	E1						religion: other (coded)
e2	E2	E2	E2	E2	E2	E2	religiosity: frequency of praying
e3	E3	E3	E3	E3	E3	E3	religiosity: visiting place of worship
e4	E4	E4	E4	E4	E4	E4	religiosity: celebrating religious holidays within family
e5_1	E5	E5	E5	E5	E5	E5	religious identity: importance
e5_2							religious identity: troubles me, if sb. speaks ill
e5_3							religious identity: close to my heart
e5_4							religious identity: feel as part
e6_1	E6	E6	E6	E6	E6	E6	religious identity: public regard: respect
e6_2							religious identity: public regard: like
e6_3							religious identity: public regard: positive view
e7_1	E7	E7	E7	E7	E7	E7	discrimination: religion spoken badly
e7_2							discrimination: religion insulted or offended
e7_3							discrimination: religion treated badly or unfair
e8_1	E8	E8	E8	E8			parents: religious identity: troubles them, if sb. speaks ill
e8_2							parents: religious identity: close to their heart
e8_3							parents: religious identity: feel as part
e8_4							parents: prefer friends of same religion
e9_1					E8	E8	permeability: difficult to be seen as a German (religion)

Variable name	W1	W2	W3	W4	W5	W6	Label
e9_2							permeability: impossible to be seen as a German (religion)
f1_1	F1	F1	F1	F1	F1	F1	German language: speaking
f1_2							German language: understanding
f1_3							German language: reading
f1_4							German language: writing
f2	F2	F2	F2	F2	F2	F2	other spoken language in household
f2_coded							other spoken language in household (coded)
f3_1	F3	F3	F3	F3	F3	F3	other language: speaking
f3_2							other language: understanding
f3_3							other language: reading
f3_4							other language: writing
f4_1	F4	F4	F4			F4	spoken language: mother
f4_2							spoken language: father
f4_3							spoken language: parents among themselves
f4_4							spoken language: siblings
f4_5							spoken language: other relatives
f4_6							spoken language: friends
g1_1	G1	G1	G1	G1	G1	G1	spending time in neighborhood with: Germans
g1_2							spending time in neighborhood with: Turks
g1_3							spending time in neighborhood with: Italians
g1_4							spending time in neighborhood with: Poles
g1_5							spending time in neighborhood with: Muslims
g1_6							spending time in neighborhood with: Christians
g1_7							spending time in neighborhood with: Jews
g2_1	G2	G2	G2	G2	G2	G2	liking: Germans
g2_2							liking: Turks
g2_3							liking: Italians
g2_4							liking: Poles
g2_5							liking: Muslims
g2_6							liking: Christians
g2_7							liking: Jews
g3_1	G3	G3	G3			G3	parents: national identity: troubles them, if sb. speaks ill
g3_2							parents: national identity: close to their heart
g3_3							parents: national identity: feel as part
g3_4							parents: support having German friends
g3_5							parents: support having friends from countries other than Germany

Variable name	W1	W2	W3	W4	W5	W6	Label
g4_1	G4	G4	G4	G3	G3	G4	parents: tell that it is important what friends to have
g4_2							parents: tell that should not have contact with certain people
g4_3							parents: tell when they don't like friends
g4_4							parents: encourage to do activities with friends they like
g4_5							parents: tell always to tell them what one is doing
g4_6							parents: want to know parents their child is hanging out with
g5	G5	G5	G5	G4	G4	G5	mother: how often seen
g6	G6	G6	G6	G5	G5	G6	mother: occupation
g7_isco08	G7	G7	G7	G6	G6	G7	mother: occupation: ISCO-08
g7_isei08							mother: occupation: ISEI-08
g7_siops08							mother: occupation: SIOPS-08
g8	G8	G9	G9	G8	G8	G9	father: how often seen
g9	G9	G10	G10	G9	G9	G10	father: occupation
g10_isco08	G10	G11	G11	G10	G10	G11	father: occupation: ISCO-08
g10_isei08							father: occupation: ISEI-08
g10_siops08							father: occupation: SIOPS-08
g11_mc_1	G11	G13	G13	G12	G12	G13	persons in household: biological mother
g11_mc_2							persons in household: biological father
g11_mc_3							persons in household: adoptive/step-/foster mother
g11_mc_4							persons in household: adoptive/step-/foster father
g11_mc_5							persons in household: brother/s
g11_mc_6							persons in household: sister/s
g11_mc_7							persons in household: grandparents/other family members
g12	G12	G14	G14	G13	G13	G14	size of household
g13_1	G13	G15				G15	younger sisters: number
g13_2							younger brothers: number
g14_1	G14	G16				G16	older sisters: number
g14_2							older brothers: number
g15	G15						family owns car
g16_1	G16	G17	G15	G14	G14	G17	pocket money: frequency
g16_2							pocket money: weekly amount
g16_3							pocket money: monthly amount
g16_4		G17	G15	G14	G14	G17	pocket money: amount
g17	G17	G18	G16				last year family vacation
g17_coded							last year family vacation (coded)
g18	G18					G19	cultural capital: number of books in household
g19_1	G19	G20				G20	elementary school: mention 1
g19_2							elementary school: mention 2

Variable name	W1	W2	W3	W4	W5	W6	Label
g19_3							elementary school: mention 3
g20_1	G20	G21	G20	G16	G16	G21	neighborhood district: mention 1
g20_2							neighborhood district: mention 2
g20_3							neighborhood district: mention 3
g21	G21	G22				G22	neighborhood district: percentage of Germans
g22_str	G22	G25	G27			G29	comments to the questionnaire
g23		G8	G8	G7	G7	G8	appreciation of mother
g24		G12	G12	G11	G11	G12	appreciation of father
g25_1		G19	G17	G15	G15	G18	trust: people in general
g25_2							trust: cannot rely on anyone
g25_3							trust: better to be careful with strangers
g25_4							trust: neighborhood
g25_5							trust: school
g25_6							trust: grade
g25_7							trust: class
g26		G23	G21	G17	G17	G23	way to school
g27_1		G24	G26	G22	G22	G28	Big Five (extraversion): quiet character, do not talk a lot
g27_2							Big Five (openness): fantasy and imagination
g27_3							Big Five (openness): many different interests
g27_4							Big Five (tolerance): frequently in conflict with family and peers
g27_5							Big Five (extraversion): shy and reserved
g27_6							Big Five (agreeableness): steadily considerations of others
g27_7							Big Five (tolerance): starts arguments/fights
g27_8							Big Five (extraversion): like to be surrounded by many people
g27_9							Big Five (agreeableness): sometimes a bit too rude to others
g27_10							Big Five (extraversion): like to be in focus
g27_11							Big Five (openness): like to think about issues
g27_12							Big Five (extraversion): easy to make me laugh
g28			G18				football world champion: glad over German success
g29_coded			G19				football world champion: preferred next champion
g30_coded			G22	G18	G18	G24	elective subject (coded)
g30_1							network: liking (elective subject)
g30_2							
g30_3							
g30_4							

Variable name	W1	W2	W3	W4	W5	W6	Label
g30_5							
g31			G23	G19	G19	G25	subjects with separated lessons
g32_10_coded			G24	G20	G20	G26	math course (coded)
g32_11							network: liking (math course)
g32_12							
g32_13							
g32_14							
g32_15							
g32_20_coded							English course (coded)
g32_21							network: liking (English course)
g32_22							
g32_23							
g32_24							
g32_25							
g32_30_coded							German course (coded)
g32_31							network: liking (German course)
g32_32							
g32_33							
g32_34							
g32_35							
g33_coded			G25	G21	G21	G27	religion course (coded)
g33_1							network: liking (religion course)
g33_2							
g33_3							
g33_4							
g33_5							

Variables may feature four types of *missing codes*. The standard (Stata) missing code “.” signals that a student did not provide information on a particular item or question. The missing code “.u” indicates that the information the students provided remained unclear, for example, because it was not readable or because the student made multiple ticks at the same time or ticks in between the scale. While there was information, it was not attributable to the questions at hand. An example of this would be ticking right in the middle of scale levels of the pictures of the books in the household with no inclination to either bookshelf. The missing code “.v” indicates the response “Don’t know”. The missing code “.e” indicates that a student provided redundant information, which was then edited. An example of this would be, when asked about music styles, having written down a musician such as “Kendrick Lamar” or “Kool Savas” and at the same time ticking “HipHop, Rap”. The missing code “.a” indicates that the variable had been anonymized for the download version of the dataset, and the missing code “.w” indicates if the variable was not part of the survey wave.



## 6.4. Coding of open questions

We checked and coded string variables as follows:

- **a12\_coded\_1 – a12\_coded\_2** provide the *favorite music* the student indicated in the open response category for music genres other than those listed in the questionnaire. We summarized similar music genres in broader categories. If a student mentioned specific songs or artists, we classified the answer in the music genre they represent.
- **a13\_coded\_1 – a13\_coded\_5** show the *sports* the student indicated in the open response category for sports other than those listed the questionnaire. We coded the answers following the coding scheme for sports provided by the NEPS Data Center (National Educational Panel Study; for further information, see <https://www.neps-data.de>), which corrected for different spellings of specific sports. The five items present up to five different indications.
- **a14\_coded** indicates the school subject students like the most, categorized from an open response.
- **a15\_coded** indicates the subject students like the least, categorized from an open response.
- **a17\_coded\_1** contains the *trade-in value of the students' smartphones*. This information is generated from searching online for the trade-in price of the smartphone (model name provided by the students) at the time of the interview and is available from wave 3 onwards.
- **a17\_coded\_2** contains the *value in new of the students' smartphones*. This information is generated from searching online for the value in new price of the smartphone (model name provided by the students) at the time of the interview and is available from wave 3 onwards.
- **b4\_coded** indicates the *country of birth of the student*. We defined countries using the ISO-3166-1 coding scheme<sup>3</sup>. For instance, Germany has the value "276." We applied this scheme to all following variables referring to countries.
- **b6\_coded** presents the *citizenship of the student* mentioned in the open response category for citizenships other than those listed in the questionnaire. If more than one country was mentioned, the code of the variable shows the different citizenships combined one after another using the ISO-3166-1 coding scheme. For instance, "Germany, Turkey" is coded as "276792."
- **b7\_coded** and **b8\_coded** show the *country of birth of the parents*, applying the coding rule as described for b4\_coded.
- **c1\_coded** indicates the *family's country of origin*, which was assessed by an open question. If more than one country was mentioned, the code shows the different countries combined one after another. For an example, see the description of b6\_coded.
- **c7\_coded** presents the *self-categorized identity* a student indicated in the open response category for identities other than those listed in the questionnaire. Again, if more than one identity was mentioned, the code shows the different identities combined one after another. The code of each identity follows the ISO-3166-1 list for countries.
- **c17\_coded** and **c18\_coded** give the *country of birth of the paternal grandparents*. This information is available from wave 2 onwards.

---

<sup>3</sup> For more information about the ISO-3166-1 coding scheme see:  
[http://www.iso.org/iso/country\\_codes.htm](http://www.iso.org/iso/country_codes.htm).

- **c19\_coded** and **c20\_coded** show the *country of birth of the maternal grandparents*. This information is available from wave 2 onwards.
- **e1\_coded** presents the *religion of the student* mentioned in the open response category for religions other than those listed in the questionnaire. We assigned religious subgroups to broader categories (e.g., Shia and Sunni both are coded as “Islam”).
- **f2\_coded** shows the *language spoken in the household* if a student indicated that another language than German was spoken as well.
- **g7\_isco08**, **g7\_isei08** and **g7\_siops08** indicate the *occupation of the mother*. First, 4-digit ISCO-08 was coded on the basis of the open response<sup>4</sup>. ISCO includes two dimensions: the skill level of the occupation and the skill specialization regarding the professional activity. Furthermore, ISEI-08 and SIOPS-08 are derived from ISCO-08. Coding was carried out by external service provider Manuel Munz of STR Coding (Nuremberg, Germany).
- **g10\_isco08**, **g10\_isei08** and **g10\_siops08** present the *occupation of the father*. Like in case of the mother, first the 4-digit ISCO-08 is coded on the basis of the open response of the student. ISEI-08 and SIOPS-08 are derived from ISCO-08.
- **g17\_coded** presents the *country of vacation of the family in the last year*. We used the ISO-3166-1 coding scheme to indicate the country.
- **g29\_coded** shows the *desired next soccer world champion*. This information is given in wave 3. The code relies on the ISO-31661-1 list for countries (see **b4\_coded**).
- **g30\_coded** indicates the *elective subject*. This information is available from wave 3 onwards. This variable is not labelled due to data confidentiality.
- **g32\_10\_coded**, **g32\_20\_coded**, **g32\_30\_coded**, and **g33\_coded** represent the respective *course for math, English, German and religion* if the student indicated that there were separate lessons in school. This information is available only for wave 3 and is to be viewed experimentally. The reason is that the exact course structure and students’ attribution to these within different schools was unknown and self-disclosure about the course structure by students was very ambiguous due to naming courses in non-transparent ways.

## 6.5. Edited variables

To ease working with the data, we revised or extended some variables of the questionnaire. This especially concerns variables with inconsistent information or ambiguous values. Furthermore, we constructed variables capturing the ethnic origin and generational migration status including information from different variables.

- **ed\_a6** gives information on student gender that has been harmonized over waves.
- **ed\_a7\_2** gives information on the month of birth, harmonized over waves.
- **ed\_a7\_3** gives information on the year of birth, harmonized over waves.
- **ed\_age** gives information on age, using harmonized year and month of birth.
- **ed\_gradesize** contains the number of students in the grade.
- **ed\_classsize** contains the number of students in class.
- **ed\_response\_grade** contains the response rate at the grade-level.

---

<sup>4</sup> For more information about ISCO-08 see the webpage of the International Labour Organization: <http://www.ilo.org/public/english/bureau/stat/isco/isco08/>.

- **ed\_response\_class** contains the response rate at the classroom-level.
- **ed\_oq** indicates the accidental application of an outdated version of the questionnaire in the first wave. The response scale of the variables regarding grades in school (a8\_1\_w1, a8\_2\_w1, and a8\_3\_w1) was designed in a different way. In the normal version, the question offered three spaces for the students' response. In contrast, the outdated version showed a scale from 1 to 6 representing school grades, and students were asked to check the respective grade. In total, 310 students of the first wave received the outdated version. Importantly, during fieldwork, this accidental application was noticed and 246 old questionnaires were changed by hand to comply with the normal version. Therefore, in a strict sense, only 64 students filled out the old version.

The data set includes numerous edited variables for both ethnic origin and generational status. As these concepts are defined in different ways in the literature, we provide different variables. Our procedure concerning missing values and/or inconsistent information follows the approach proposed by Dollmann et al. (2014).

- **ed\_countorig\_cils** defines the *student's country of origin* by drawing on the information about the country of birth of the student as well as the student's parents and grandparents. The first relevant information to construct the country of origin is provided by the grandparents. It takes the birth countries into account and follows the majority rule, but only if the respective grandparents are born in a foreign country. Accordingly, the country of origin represents the birth country of the majority of the grandparents born abroad. If there is no majority of a foreign country, it gives most importance to the female ancestries. Thus, the maternal grandmother has the first priority, followed by the maternal grandfather. If the grandparents are born in Germany, the variable is based on the countries of birth of the parents. We applied the same rule at the parental level. Again, we only consider foreign countries of birth. If the parents were born in different countries, the country of the mother has priority. Whenever grandparents as well as parents are born in Germany, the country of birth of the student defines the country of origin. For further information about the construction of the variable, especially regarding the treatment of missing values and non-trivial cases, see Dollmann et al. (2014).
- **ed\_countorig\_cils\_group** summarizes the information provided by the variable *ed\_countorig\_cils*. It still represents the countries with a substantial number of observations as separate categories, such as Turkey or Poland. However, the coding collapses countries with fewer observations into broader groups representing different world regions, such as Southern Europe or Northern Africa.
- **ed\_miggen\_cils\_narrow** defines the *migration generational status of the student* following a fine-grained approach. It differentiates between the 1<sup>st</sup>, 1.25<sup>th</sup>, 1.5<sup>th</sup>, 1.75<sup>th</sup>, 2<sup>nd</sup>, 2.5<sup>th</sup>, 2.75<sup>th</sup>, interethnic 2<sup>nd</sup>, 3<sup>rd</sup>, 3.25<sup>th</sup>, 3.5<sup>th</sup>, interethnic 3<sup>rd</sup>, and 3.75<sup>th</sup> generations and natives (Dollmann et al. 2014). First, it takes the country of birth of the student into account, followed by the countries of birth of the parents, and then those of the grandparents. If a student is foreign born and migrated him- or herself to Germany, it defines this student as *1<sup>st</sup> generation* immigrant. For further differentiation we take the age of the student at the time of migration into account. Children who migrated after the age of 10 are indicated as 1.25<sup>th</sup>, between the ages of 6 and 10 as 1.5<sup>th</sup>, and before the age of 6 as 1.75<sup>th</sup> *generation*. Students born in Germany with both parents foreign-born are 2<sup>nd</sup> *generation* immigrants. If one of the parents is born in Germany, the student is

classified as *2.5<sup>th</sup> generation*. If the parent born in Germany has a parent (grandparent of the student) born abroad, the student is defined as *2.75<sup>th</sup> generation*. If the parent born in Germany has parents born in Germany (grandparents of the student), the student is considered to be *interethnic 2<sup>nd</sup> generation*. The student is classified as *3<sup>rd</sup> generation* if the student as well as the parents are born in Germany, but the grandparents are born abroad. Students who are born in Germany with both parents born in Germany but three grandparents born abroad are defined as *3.25<sup>th</sup> generation*. The *3.5<sup>th</sup> generation* includes students who are born in Germany, with both parents also born in Germany, but have each one parent (grandparents of the student) who is born abroad. The *interethnic 3<sup>rd</sup> generation* consists of students who are born in Germany, and with parents also born in Germany, but both parents of one parent (grandparents of the student) born abroad. Students born in Germany with both parents born in Germany, but one grandparent born abroad, are defined as *3.75<sup>th</sup> generation*. *Natives* include students who are born in Germany with all parents and grandparents also born in Germany. Additional categories regarding missing information are provided. For a detailed description, illustrations of the procedure, and information about the treatment of missing values, see Dollmann et al. (2014).

- **ed\_miggen\_cils\_wide** pools the information of the variable *ed\_countorig\_cils* in broader categories.
- **ed\_countorig\_ext** defines the *student's country of origin*, and is based on *ed\_countorig\_cils* constructed following the approach of the CILS4EU study (Dollmann et al. 2014). However, to reduce missing values we used additional information from the questionnaire, in particular family's country of origin (*c1\_coded*) as well as self-categorized identity (*c7\_coded*). Therefore, we applied three rules. First, we replaced missing by family's country of origin (*c1\_coded*) if this is not missing and not Germany. Second, we replaced remaining missing by identity (*c7\_coded*) if this is not missing and not German. Third, Germany is seen as country of origin, and thus the student as native, whenever country of birth of the student, parents and grandparents is Germany or missing, and on the variables regarding family's country of origin (*c1\_coded*) and identity (*c7\_coded*) are also missing.
- **ed\_countorig\_ext\_group** summarizes the variable *ed\_countorig\_ext* in broader categories. Countries with the most cases are indicated as separate categories, the remaining countries are pooled in groups of the different world regions.
- **ed\_miggen\_ext\_wide** defines the *migration generational status of the student*. We constructed the variable, as well as *ed\_miggen\_cils\_wide*, following the approach proposed in the CILS4EU study (Dollmann et al. 2014). However, it is based on *ed\_countorig\_ext* with reduced missing values.

For the construction of the variables regarding the students' ethnic origin as introduced above, we used several additional edited variables.

- **ed\_migage**: age of migration of the student
- **ed\_cob**: country of birth of student
- **ed\_cobm** and **ed\_cobf**: country of birth of parents
- **ed\_cobmm** and **ed\_cobmf**: country of birth of maternal grandparents
- **ed\_cobfm** and **ed\_cobff**: country of birth of paternal grandparents

- **ed\_cobfm\_group** and **ed\_cobff\_group**: summarizes the variables *ed\_cobmf* and *ed\_cobff* in broader categories. Countries with the most cases are indicated as separate categories, the remaining countries are pooled in groups of the different world regions.
- **ed\_cobmm\_nat** and **ed\_cobmf\_nat**: maternal grandparents born in Germany
- **ed\_cobfm\_nat** and **ed\_cobff\_nat**: paternal grandparent born in German

## 6.6. Data anonymization

DeZIM.fdz offers the collected data of the study “Friendship and Identity in School” as anonymized Scientific Use Files (SUF). The degree of anonymization depends on the modes of data access (download, remote desktop, on-site). The more the data access is technically controlled, the less the variance of the data has to be reduced by aggregation and the greater its analytical potential remains.

In the study “Friendship and Identity in School”, direct identifiers (such as names of the students) were recorded in a separate data set and are not included in the data sets. Instead, the data were pseudonymized. With regard to indirect identifiers, the following were assessed for the study: day and month of birth, all country information (including the respondent's country of birth, the respondent's mother's or father's country of birth, the respondent's grandfather's or grandmother's country of birth), entry age to Germany, mother's and father's occupation, the spoken language, the respondent's religious affiliation, and the household size. The corresponding variables were - depending on the modes of data access - either aggregated or partially deleted. Open questions are also considered indirect identifiers and were either coded or deleted. Aggregated questions may be identified in the data ending with the suffix “\_group”. All variables subject to anonymization by deletion are still contained in the data but their values are removed and replaced with the missing code “a”.

Variables subject to different forms of anonymization are: *a7\_2*, *a7\_3*, *a12\_coded\_1*, *a12\_coded\_2*, *a13\_coded\_1*, *a13\_coded\_2*, *a13\_coded\_3*, *a13\_coded\_4*, *a13\_coded\_5*, *b4\_coded*, *b5*, *b6\_coded*, *b7\_coded*, *b8\_coded*, *c1\_coded*, *c7\_coded*, *c17\_coded*, *c18\_coded*, *c19\_coded*, *c20\_coded*, *e1\_w1*, *e1\_coded*, *f2\_coded*, *ed\_age*, *g17\_coded*, *g29\_coded*, *g7\_isco08*, *g10\_isco08*, *g12*, *g13\_1*, *g13\_2*, *g14\_1*, *g14\_2*, *ed\_a7\_2*, *ed\_a7\_3*, *ed\_cob*, *ed\_migage*, *ed\_cobm*, *ed\_cobf*, *ed\_cobmm*, *ed\_cobmf*, *ed\_cobfm*, *ed\_cobff*, *ed\_countorig\_cils*, *ed\_countorig\_ext*.

## 6.7. Response rates

In Table 8, we provide an overview of students' participation in the survey across grades, waves and classes. Readers may find this information helpful both to get an overview of the comprehensiveness of the data and to select networks for their statistical analysis.

Table 8: Response rates by grade, class, and wave

School	Grade	Wave	Grade size	Share participated	Number of classes	Share participated within classes				
						Class 1	Class 2	Class 3	Class 4	Class 5
1	1	1	60	90.0	3	100.0	85.0	85.7		
1	1	2	62	85.5	3	95.0	95.0	68.2		
1	1	3	68	86.8	3	95.5	82.6	82.6		
1	1	4	63	57.1	3	100.0	78.9	0.0		
1	1	5	66	78.8	3	90.5	61.9	83.3		

1	1	6	70	85.7	3	87.5	86.4	83.3	
1	2	1	38	73.7	2	77.8	70.0		
1	2	2	48	77.1	2	75.0	79.2		
1	2	3	47	91.5	2	91.7	91.3		
1	2	4	53	86.8	2	88.9	84.6		
1	2	5	51	70.6	2	73.1	68.0		
1	2	6	42	88.1	2	79.2	100.0		
1	3	1	63	79.4	3	52.4	100.0	85.0	
1	3	2	63	77.8	3	61.9	95.5	75.0	
1	3	3	60	90.0	3	100.0	90.9	78.9	
1	3	4	54	85.2	3	86.7	88.2	81.8	
1	3	5	54	70.4	3	68.8	58.8	81.0	
2	4	1	63	87.3	2	90.6	83.9		
2	4	2	61	86.9	2	87.5	86.2		
2	4	3	61	86.9	2	93.5	80.0		
2	4	4	60	95.0	2	96.7	93.3		
2	4	5	60	81.7	2	83.3	80.0		
2	5	1	60	63.3	2	73.3	53.3		
2	5	2	61	95.1	2	96.7	93.5		
2	5	3	63	96.8	2	100.0	93.8		
2	5	4	59	98.3	2	100.0	96.6		
2	5	5	61	82.0	2	83.9	80.0		
2	6	1	55	80.0	2	92.9	66.7		
2	6	2	65	92.3	2	93.9	90.6		
2	6	3	54	88.9	2	96.0	82.8		
2	6	4	50	84.0	2	66.7	100.0		
2	6	5	49	83.7	2	73.9	92.3		
3	7	1	77	81.8	3	73.9	85.7	84.6	
3	7	2	79	86.1	3	92.0	85.2	81.5	
3	7	3	80	83.8	3	79.2	89.3	82.1	
3	8	1	85	91.8	3	89.3	96.4	89.7	
3	8	2	86	89.5	3	92.9	82.8	93.1	
3	8	3	83	89.2	3	82.1	89.3	96.3	
3	9	1	80	88.8	3	100.0	88.5	77.8	
3	9	2	70	88.6	3	95.5	95.0	78.6	
3	9	3	70	80.0	3	83.3	81.0	76.0	
4	10	1	103	56.3	4	64.3	77.3	46.4	40.0
4	10	2	114	82.5	4	90.0	77.3	77.4	83.9
4	10	3	112	83.9	4	90.0	77.3	90.0	76.7
4	11	1	115	76.5	4	87.5	89.3	64.0	63.3
4	11	2	115	87.0	4	90.3	93.1	88.5	75.9
4	11	3	115	91.3	4	93.1	87.1	92.9	92.6
4	12	1	122	61.5	4	53.6	69.7	77.4	43.3
4	12	2	122	86.1	4	73.3	93.9	86.7	89.7
4	12	3	124	87.1	4	80.0	100.0	80.6	86.7
5	13	1	112	73.2	4	76.0	88.5	83.3	48.4
5	13	2	112	84.8	4	73.1	88.5	86.7	90.0
5	13	3	109	83.5	4	88.5	76.0	83.3	85.7

5	13	4	107	86.0	4	100.0	84.6	82.8	78.6	
5	13	5	111	75.7	4	80.8	73.1	66.7	82.8	
5	13	6	115	92.2	4	96.3	92.6	80.6	100.0	
5	14	1	119	68.1	4	82.8	50.0	86.7	53.3	
5	14	2	116	73.3	4	86.2	51.7	82.1	73.3	
5	14	3	117	86.3	4	96.6	75.9	86.7	86.2	
5	14	4	116	77.6	4	90.0	75.0	75.9	69.0	
5	14	5	119	79.8	4	71.0	83.3	80.0	85.7	
5	14	6	117	86.3	4	83.9	73.3	88.9	100.0	
5	15	1	117	62.4	4	89.3	58.1	35.7	66.7	
5	15	2	120	77.5	4	93.3	76.7	63.3	76.7	
5	15	3	119	71.4	4	67.7	73.3	66.7	78.6	
5	15	4	111	90.1	4	89.7	86.2	92.3	92.6	
5	15	5	113	76.1	4	69.0	83.3	65.4	85.7	
6	16	1	133	87.2	5	93.3	53.8	95.7	96.0	96.6
6	16	2	129	83.7	5	96.6	68.0	81.8	84.0	85.7
6	16	3	129	91.5	5	93.1	92.6	95.8	73.9	100.0
6	16	4	130	93.1	5	96.6	92.6	91.3	82.6	100.0
6	16	5	132	93.2	5	92.9	100.0	81.0	96.4	92.9
6	16	6	142	93.0	5	93.1	96.7	88.9	88.9	96.6
6	17	1	117	74.4	4	96.8	96.8	0.0	96.4	
6	17	2	115	72.2	4	93.3	100.0	17.2	76.9	
6	17	3	115	77.4	4	96.7	100.0	24.1	88.5	
6	17	4	117	92.3	4	90.3	93.1	96.6	89.3	
6	17	5	120	86.7	4	90.3	83.9	82.8	89.7	
6	17	6	116	90.5	4	93.8	93.1	85.2	89.3	
6	18	1	121	82.6	4	90.0	76.7	87.1	76.7	
6	18	2	120	90.0	4	96.7	80.0	90.0	93.3	
6	18	3	115	94.8	4	96.7	92.9	92.9	96.6	
6	18	4	120	95.8	4	89.7	96.7	100.0	96.7	
6	18	5	119	89.1	4	86.2	96.8	90.0	82.8	
7	19	1	26	80.8	1	80.8				
7	19	2	49	67.3	2	64.0	70.8			
7	19	3	49	81.6	2	73.9	88.5			
7	19	4	52	80.8	2	80.0	81.5			
7	19	5	52	76.9	2	76.0	77.8			
7	19	6	69	75.4	3	87.0	73.9	65.2		
7	20	1	47	55.3	2	77.3	36.0			
7	20	2	51	60.8	2	76.0	46.2			
7	20	3	54	81.5	2	77.8	85.2			
7	20	4	54	81.5	2	80.8	82.1			
7	20	5	53	77.4	2	76.0	78.6			
7	20	6	40	75.0	2	69.2	85.7			
7	21	1	45	84.4	2	90.9	78.3			
7	21	2	51	72.5	2	80.0	65.4			
7	21	3	45	91.1	2	82.6	100.0			
7	21	4	37	89.2	2	88.9	89.5			
7	21	5	38	86.8	2	83.3	90.0			

8	22	1	102	79.4	4	81.5	66.7	83.3	87.5
8	22	2	107	91.6	4	89.7	96.3	80.8	100.0
8	22	3	113	92.0	4	89.7	96.3	83.3	100.0
8	22	4	117	91.5	4	92.9	93.1	93.1	87.1
8	22	5	117	90.6	4	96.4	79.3	96.6	90.3
8	22	6	112	88.4	4	96.7	82.1	96.2	78.6
8	23	1	108	85.2	4	92.6	92.9	88.5	66.7
8	23	2	109	89.9	4	92.3	92.9	100.0	74.1
8	23	3	108	91.7	4	92.0	88.9	96.4	89.3
8	23	4	100	92.0	4	100.0	96.2	100.0	74.1
8	23	5	101	87.1	4	82.6	96.2	88.5	80.8
8	23	6	101	88.1	4	100.0	80.8	80.0	92.0
8	24	1	101	86.1	4	89.3	70.8	88.0	95.8
8	24	2	104	88.5	4	88.9	87.5	92.0	85.7
8	24	3	104	92.3	4	95.8	91.7	100.0	83.3
8	24	4	109	96.3	4	96.2	100.0	96.3	93.3
8	24	5	105	73.3	4	80.0	76.9	80.8	57.1
9	25	1	41	80.5	2	77.3	84.2		
9	25	2	51	76.5	3	78.9	80.0	70.6	
9	25	3	49	75.5	2	64.0	87.5		
9	26	1	75	65.3	4	73.7	76.5	40.0	73.7
9	26	2	70	62.9	4	77.8	75.0	35.0	68.8
9	26	3	56	78.6	4	93.8	100.0	55.0	66.7
10	27	1	44	65.9	2	56.5	76.2		
10	27	2	65	73.8	3	72.7	76.2	72.7	
10	27	3	66	74.2	3	79.2	85.7	57.1	
10	28	1	76	68.4	3	84.6	60.0	60.0	
10	28	2	79	72.2	3	69.2	73.1	74.1	
10	28	3	65	83.1	3	90.9	81.8	76.2	
10	29	1	78	65.4	3	70.4	64.0	61.5	
10	29	2	78	78.2	3	84.0	73.1	77.8	
10	29	3	62	79.0	3	81.8	88.9	68.2	

## 7. Examples of data use

Given the rich longitudinal information on social relations, the data at hand are particularly well-suited for social network analysis and therefore we expect the majority of researchers working with this dataset to be keen on employing techniques of social network analysis. To ease the process of getting familiar with the data, we provide a number of examples that show how the data can be formatted and used for network analysis, allowing researchers to conveniently adapt these scenarios to their own applications. In doing so, we mainly employ the statistical software package R. However, Section 6.3. also demonstrates an application in Stata. For convenience, the code for all examples is provided with the data.



## 7.1. Plotting networks

In our first example, we create what frequently is the first step in social network research: a network plot. In this case, we show how to generate a single grade-level network plot, with the color of the actor nodes representing students' gender.

We start by clearing the workspace and loading the required packages:

```
rm(list=ls())

packages <- c("sna", "tidyverse", "haven")
lapply(packages, require, character.only=T)
rm(packages)
```

```
path_in <- "PATH_TO_DATA"
```

Then we load both the composition data (representing the network boundaries) and the survey data for all grades and all waves.

```
fis.orig <- read_dta(paste(path_in, "/fis_long.dta", sep=""))
fis.comp <- read_dta(paste(path_in, "/fis_composition.dta", sep=""))
```

We specify the variables constituting the network. In this case, the network consists of the students' friendship nominations captured in a1\_0101 to a1\_0110. We specify the covariates of interest – students' gender (ed\_a6) – and the time-constant individual-student identifier (id\_p). Finally, we specify the grade network we want to plot (grade 1) and the wave we are interested in (wave 1).

```
# Specify Set of Network Characteristics (10 friendship nominations)
netw <- c("a1_0101", "a1_0102", "a1_0103", "a1_0104", "a1_0105",
         "a1_0106", "a1_0107", "a1_0108", "a1_0109", "a1_0110")

# Specify Covariate for Node Color: ed_a6 (gender)
cova <- c("ed_a6")

# Specify Individual Identifier
iden <- c("id_p")

# Specify Grade and Wave
g <- 1
w <- 1
```

From the composition file and the survey data file, we extract information on all students that were in the grade of interest in the relevant wave. In the next step, we merge the two data sets, keeping information on all students that were part of the grade-level network. Thus, students who did not provide survey data were included as well (variable `right` takes the value NA).

```
comp.g.w <- filter(fis.comp, wave == w, id_g == g)
fis.g.w <- filter(fis.orig, wave == w, id_g == g) %>%
  select(iden, netw, cova) %>%
  mutate(right = 1)

fis.g.w <- left_join(comp.g.w, fis.g.w, by="id_p") %>% arrange(id_p)
```

We next restrict the data set to the network-related variables. To facilitate working with the R network packages, we generate new student identifiers that start with "1". The relation between the original identifiers and the new identifiers is saved in `ids.tr`.

```
fis.fr.g.w <- select(fis.g.w, iden, netw)
ids <- unlist(select(fis.fr.g.w, "id_p")) # extract all IDs
```

```
ids.s      <- sort(ids) # sorts unique IDs
ids.tr     <- cbind(ids.s,1:length(ids)) # generates matching IDs from 1 onwards
```

We identify unit non-response, as indicated above, from the “right” variable taking the missing value NA.

```
miss.w <- ids[unlist(is.na(select(fis.g.w,right)))]
```

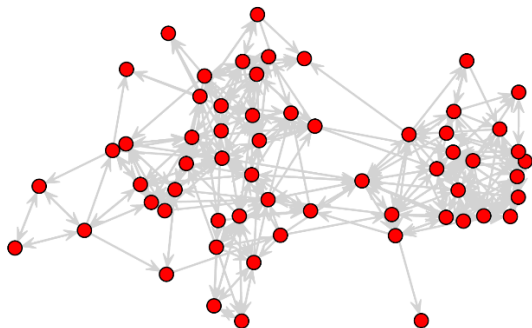
We overwrite the original identifier information in the network data set and identify the new ids of units of non-response.

```
f <- 0
for (i in ids.s) {
  f <- f + 1
  fis.fr.g.w[fis.fr.g.w == i] <- f
}
miss.w.tr <- ids.tr[,2][ids.s %in% miss.w]
```

Now, we are ready to generate the network information. We first generate an edgelist from the students’ friendship nominations. Then we generate an empty adjacency matrix and fill it with the relations provided by the edgelist. Out-going nominations of students with unit non-response are set to be missing. With these preparations, we are now ready to plot the network.

```
edge.w <- cbind(unlist(fis.fr.g.w[, 1]), unlist(c(fis.fr.g.w[, -1]))) # generate edgelist
mat.w <- matrix(0,nrow=length(ids),ncol=length(ids)) # generate empty adjacency matrix
mat.w[edge.w] <- 1 # fill adjacency matrix
mat.w[miss.w.tr,] <- NA # fill missing actors with NAs
mat.adj.w <- as.network(mat.w,matrix.type="adjacency") # transform to R adjacency matrix
```

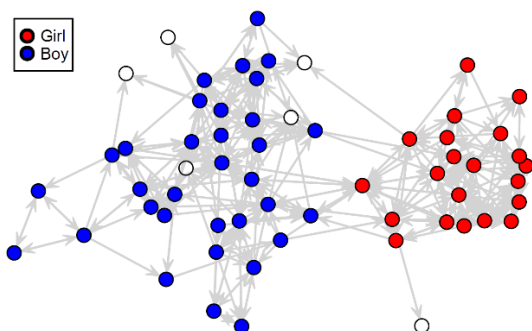
```
netw.plot <- gplot(mat.adj.w,edge.col="lightgrey")
```



Finally, we create a vector capturing students’ gender. Note that we have to make sure that the identifiers are identically sorted in the network data and in the gender vector. We arrive at the final colored network plot:

```
gen.col <- fis.g.w %>% arrange(id_p) %>% transmute(gen.col = ifelse(ed_a6 == 2, "blue",
  ifelse(is.na(ed_a6)==TRUE,NA,"red"))) %>% unlist()

gplot(mat.adj.w,vertex.col=gen.col,coord=netw.plot,edge.col="lightgrey")
legend(min(netw.plot[,1]), max(netw.plot[,2]), legend=c("Girl", "Boy"),pch=21,
  col="black",pt.bg=c("red","blue"),pt.cex=1.3,pt.lwd=2.5,cex=.8)
```



## 7.2. Longitudinal network analysis (RSiena): Same-sex homophily

In this example, we show how stochastic actor-oriented models (SAOM) can be run with the Friendship and Identity in School data. Substantively, we focus on gender segregation of students' friendship networks. We recommend to study the network plot example (6.1.) before turning to this example because the former clarifies many of the components used in this example. Here, we demonstrate how the composition and the survey data can be used to generate grade-level networks over time, build SIENA objects, analyze the SIENA objects, and aggregate the results using a meta-analysis.

We start by clearing the workspace and loading the required packages:

```
rm(list=ls())

lapply(c("doParallel", "RSiena", "sna", "mvmeta",
        "tidyverse", "haven", "rlang"), require, character.only=T)

path_in <- "PATH_TO_DATA"
path_out <- "PATH_TO_OUTPUT_FOLDER"
```

Then we load both the composition data (representing the network boundaries) and the survey data for all grades and all waves.

```
fis.orig <- read_dta(paste(path_in, "/fis_long.dta", sep=""))
fis.comp <- read_dta(paste(path_in, "/fis_composition.dta", sep=""))
```

We specify the variables constituting the network. In this case, the network consists of the students' friendship nominations captured in a1\_0101 to a1\_0110. We specify the covariates of interest – students' gender (ed\_a6) – and the time-constant individual-student identifier (id\_p). We specify the networks we want to analyze (grades 4, 5, and 6) and the waves we are interested in (waves 1, 2, and 3). Finally, we create empty lists of the SIENA objects and the effects included in the SIENA analysis, which will be filled by grade-level information later.

```
# Specify Set of Network Characteristics
netw <- c("a1_0101", "a1_0102", "a1_0103", "a1_0104", "a1_0105",
        "a1_0106", "a1_0107", "a1_0108", "a1_0109", "a1_0110")

# Specify Set of Covariates
cova <- c("ed_a6")

# Specify Individual Identifier
iden <- c("id_p")

# Specify Waves
waves <- c(1,2,3)
```

```
# Specify Grades
grades <- c(1,4,6)

# initialize lists of SIENA and effect objects
siena.data.g <- list()
effects.specific.g <- list()
gr1 <- 0 # counter to fill lists
```

This initialization is followed by a loop that generates the SIENA objects separately for all grades specified (i.e., grade 4, 5, and 6 in the example at hand). For each grade under investigation, we first extract the composition file data for all three waves. This data has to be transformed into wide format: This ensures that all networks constructed below contain information on all students that have ever been part of the grade, which constitutes the longitudinal network boundary. We later account for unit non-response and absence from the grade in specific waves.

```
for (g in grades) {

  gr1 <- gr1 + 1

  ## Specify Full (Network) Data ##
  ## ----- ##

  # composition over time in wide format
  comp.g <- filter(fis.comp,wave %in% waves, id_g == g) %>% spread(key="wave",value="wave")
```

We next extract the survey data for the given grade for all waves in long format. For each grade, we then loop across all waves under analysis (i.e., waves 1, 2, and 3 in the example). Before doing so, however, we initialize a list and a vector that capture the wave-specific networks and initialize a covariate vector.

```
# survey data over time in long format
fis.g <- filter(fis.orig,wave %in% waves, id_g == g) %>%
  select(iden,wave,netw,cova) %>% mutate(right = 1)

# Extract Wave-Specific Networks
mat.str.w <- list()
mat.adj.w <- list()
mat.str.v <- NULL
gender <- rep(NA,dim(comp.g)[1]) # per-grade covariate
```

We next loop across waves. For each wave, we merge the wide composition data with the wave-specific survey data. We extract a subset of the data that captures the network only.

```
for (w in waves) {

  fis.g.w <- left_join(comp.g,filter(fis.g,wave == w),by="id_p") %>% arrange(id_p)
  fis.fr.g.w <- select(fis.g.w,iden,netw)
```

We identify all students that could not be matched via the composition file (NA value on the `right` variable; i.e., students who have ever been part of the network but do not provide survey data in that wave). We differentiate two types of missingness: First, we consider structural missings. Structural missings are students who were not part of the network in the wave under consideration (because they joined or left the grade over time). These are identifiable from missing values on the wave identifiers in the composition data. Second, we consider unit non-response – the remaining missing cases.

```
# all actors not matched through the composition file
all.nomatch <- filter(fis.g.w,is.na(right)==TRUE) %>%
  select("id_p") %>% unlist()

# structurally missing in the wave (not in classroom)
```

```
miss.w.s <- filter(fis.g.w, is.na(!!!parse_exprs(paste("`", w, "`", sep="")))) == TRUE) %>%
  select("id_p") %>% unlist()

# the remainder: unit non-response
miss.w <- all.nomatch[all.nomatch %in% miss.w.s == FALSE]
```

To facilitate working with the R network packages, we generate new student identifiers that start with “1”. The relation between the original identifiers and the new identifiers is saved in `ids.tr`.

```
ids <- unlist(select(fis.fr.g.w, "id_p")) # extract unique IDs
ids.s <- sort(ids) # sorts unique IDs
ids.tr <- cbind(ids.s, 1:length(ids)) # generates matching IDs from 1 onwards
```

We overwrite the original id information in the network data set and identify the units of non-response and structural missingness in the new student identifiers.

```
f <- 0
for (i in ids.s) {
  f <- f + 1
  fis.fr.g.w[fis.fr.g.w == i] <- f
}

miss.w.tr <- ids.tr[, 2][ids.s %in% miss.w]
miss.w.s.tr <- ids.tr[, 2][ids.s %in% miss.w.s]
```

Now we are ready to generate the network information. We first generate an edgelist from the students’ friendship nominations. Then we generate an empty adjacency matrix and fill it with the relations provided by the edgelist. Out-going nominations of students with unit non-response are set to missing. After transforming to a matrix, we replace incoming and outgoing nominations of structurally missing actors to structural zeroes (code 10). Because these actors were not part of the network in the given wave, they could neither nominate others nor be nominated. Finally, we put all grade-specific network information over time into a single vector because the data can be read into the SIENA package easily this way.

```
edge.w <- cbind(unlist(fis.fr.g.w[, 1]), unlist(c(fis.fr.g.w[, -1])))
mat.w <- matrix(0, nrow=length(ids), ncol=length(ids))
mat.w[edge.w] <- 1
mat.adj.w[[w]] <- as.network(mat.w, matrix.type="adjacency")
mat.str.w[[w]] <- as.matrix(mat.adj.w[[w]])
mat.str.w[[w]][miss.w.s.tr,] <- 10
mat.str.w[[w]][, miss.w.s.tr] <- 10
mat.str.v <- c(mat.str.v, mat.str.w[[w]])
```

Still in the loop for the wave-specific data preparation, we generate wave-specific information on the covariate, students’ gender (`gender.w`). Note that we have to make sure that the identifiers are identically sorted in the network data and in the gender vector. The final variable capturing students’ gender (`gender`) is dynamically updated across waves: if information on students’ gender is not provided in a given wave, it is replaced by the information provided in later waves. This concludes the wave-specific data preparation.

```
# Covariate Preparation ##
# ----- ##

gender.w <- fis.g.w %>% arrange(id_p) %>% select(ed_a6) %>% unlist()

# create gender variable from the longitudinal data
gender <- ifelse(is.na(gender), unlist(gender.w), gender)
}
```

To allow for SAOM analysis, the next paragraph converts the data to the SAOM format. Furthermore, we specify effects for the network models, focusing on a parsimonious model specification that only considers outdegree, reciprocity, transitivity, and gender ego, alter, and same effects. Note that all these steps of the data preparation are still grade-specific (i.e. within the grade loop). We save the data in SIENA format and the model specifications in the corresponding lists for all grades under analysis. Finally, we print descriptive information on the networks. This concludes the loop across grades.

```

friends <- sienaDependent(array( mat.str.v ,dim=c(dim(mat.str.w[[1]]),length(waves)) ))
gender <- coCovar(gender,centered=FALSE)

# SIENA Data
siena.data <- sienaDataCreate(friends,gender)

# Effects to Include: Outdegree, Reciprocity, Transitivity, Ego, Alter, Same Gender
effects.specific <- getEffects(siena.data)
effects.specific <- includeEffects(effects.specific,gwespFF)
effects.specific <- includeEffects(effects.specific,egoX,interaction1="gender")
effects.specific <- includeEffects(effects.specific,altX,interaction1="gender")
effects.specific <- includeEffects(effects.specific,sameX,interaction1="gender")

# Saving Data for Each Grade
siena.data.g[[grl]] <- siena.data
effects.specific.g[[grl]] <- effects.specific

# Printing Model Information for Each Grade
setwd(path_out)
print01Report(siena.data.g[[grl]],modelName=paste("Model_",grl,sep=""))
}

```

In the analysis part, we use the `doParallel` package, which facilitates running grade-specific SIENA models on multiple processors simultaneously. The results from the grade-specific SIENA models are saved as list elements of the `siena.models` object

```

registerDoParallel(cores=2) # Specify number of cores to run analysis on
setwd(path_out)

# Parallel Processing Analysis
siena.models <- foreach(grl=icount(length(siena.data.g)), .packages="RSiena") %dopar% {
m.base <- sienaAlgorithmCreate(projname=paste("Model_",grl,sep=""),
                             cond=FALSE,MaxDegree=c(friends=10),nsub=2,n3=100)
  m.base.results <- siena07(m.base,data=siena.data.g[[grl]],
                           effects=effects.specific.g[[grl]],batch=TRUE,silent=FALSE)
  m.base.results
}

```

Finally, we aggregate the results from the grade-level SIENA models in a fixed-effects meta-analysis:

```

estimates.siena <- NULL
covariances.siena <- list()
for (x in 1:length(siena.models)) {
  estimates.siena <- rbind(estimates.siena,siena.models[[x]]$theta) # Parameter estimates
  covariances.siena[[x]] <- siena.models[[x]]$covtheta # (Co)variances
}
colnames(estimates.siena) <- siena.models[[1]]$effects$effectName
rm(x)

meta.res <- mvmeta(estimates.siena,covariances.siena,method="fixed")

# Results from the Meta-Analysis
summary(meta.res)

```

The exemplary analysis provides the following results, suggesting reciprocity and transitivity in friendship relationships, and suggesting that same-sex friendships are more frequent than cross-sex friendships.

	Estimate	Std. Error	z	Pr(> z )	95%ci.lb	95%ci.ub	
constant friends rate (period 1)	11.6578	0.6884	16.9352	0.0000	10.3086	13.0070	***
constant friends rate (period 2)	11.1947	0.7493	14.9407	0.0000	9.7261	12.6632	***
outdegree (density)	-2.5544	0.1143	-22.3469	0.0000	-2.7785	-2.3304	***
reciprocity	1.3091	0.0711	18.4134	0.0000	1.1698	1.4485	***
GWESP I -> K -> J (69)	1.0988	0.0525	20.9407	0.0000	0.9959	1.2016	***
gender alter	-0.0273	0.0617	-0.4426	0.6580	-0.1482	0.0936	
gender ego	0.0597	0.0721	0.8279	0.4077	-0.0817	0.2011	
same gender	0.2695	0.0596	4.5208	0.0000	0.1527	0.3864	***

---  
 Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### 7.3. Friendship similarity in Stata

In this example, we show how the Friendship and Identity in School data can be used to investigate similarity among peers in Stata. Substantively, we address gender similarity in friendships and variation in the tendency of students to “party” depending on their friendship networks’ gender composition.

We start with a cross-sectional investigation of the data from the first wave. We restrict the data set to the friendship variables, gender, and students’ frequency of partying. We recode the variables.

```
global path_in "PATH_TO_DATA"

use "${path_in}/fis_long.dta", clear

keep id_g id_p wave a1_01* ed_a6 a9_3
keep if wave == 1

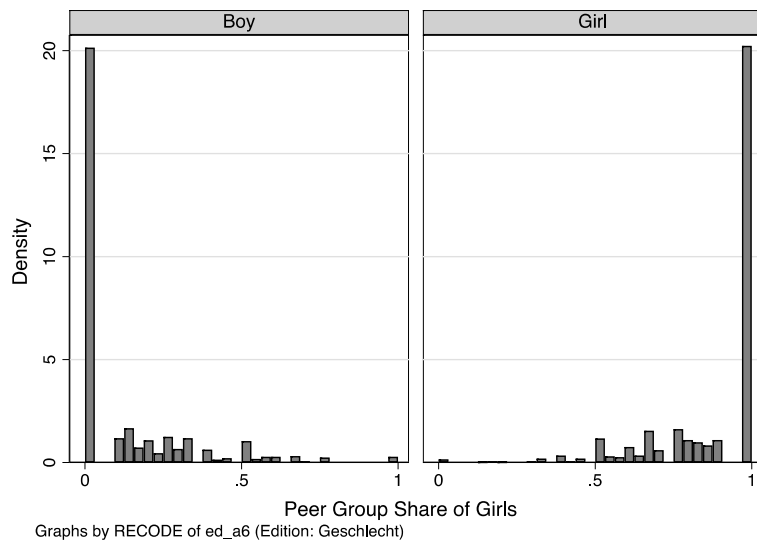
recode ed_a6 (2 = 0 "Boy") (1 = 1 "Girl"), gen(girl)
recode a9_3 (1 = 0) (2 3 4 5 6 = 1), gen(party)
```

Then we use the user-provided command `npinfo` to generate variables that capture the gender of all of a student’s nominated friends `a1_0101 – a1_0110`, from which we calculate the share of female friends. The command `npinfo` may be installed typing “`ssc install npinfo`”.

```
npinfo a1_01??, id(id_p) npcov(girl) replace
egen sh_girl = rowmean(a1_01??_girl)
```

Plotting the proportion of female friends by gender, we see that friendship networks are heavily gender-segregated.

```
hist sh_girl, by(girl) xtitle("Peer Group Share of Girls")
```



We next run three linear probability models to investigate how partying behavior depends on individuals' gender and on the gender composition of the friendship network:

```
egen miss = rowmiss(girl sh_girl)

reg party i.girl if miss == 0, robust
est store m1
reg party i.girl c.sh_girl, robust
est store m2
reg party i.girl c.sh_girl i.girl#c.sh_girl, robust
est store m3

esttab m1 m2 m3
```

	(1) party	(2) party	(3) party
1.girl	-0.0469* (-1.97)	-0.0346 (-0.66)	0.223** (2.66)
sh_girl		-0.0162 (-0.26)	0.207* (2.48)
1.girl#c.s~1			-0.485*** (-3.94)
_cons	0.593*** (36.01)	0.595*** (32.81)	0.569*** (29.41)
N	1721	1721	1721

t statistics in parentheses, \* p<0.05, \*\* p<0.01, \*\*\* p<0.001

In Models 1 and 2, the analyses do not suggest significant differences in partying behavior between boys and girls or according to the gender composition of the friendship network. Including an interaction effect between individual gender and gender composition in Model 3, we find that girls with very few female friends are more likely to party than comparable boys. The pattern reverses at a high share of female friends: Girls with mostly female friends are less likely to party than comparable boys are.

Next, we consider whether this variation also holds up in a longitudinal analysis:

```
use "${path_in}/fis_long.dta", clear
```



```
keep id_g id_p wave a1_01* ed_a6 a9_3

recode ed_a6 (2 = 0 "Boy") (1 = 1 "Girl"), gen(girl)
recode a9_3 (1 = 0) (2 3 4 5 6 = 1), gen(party)
```

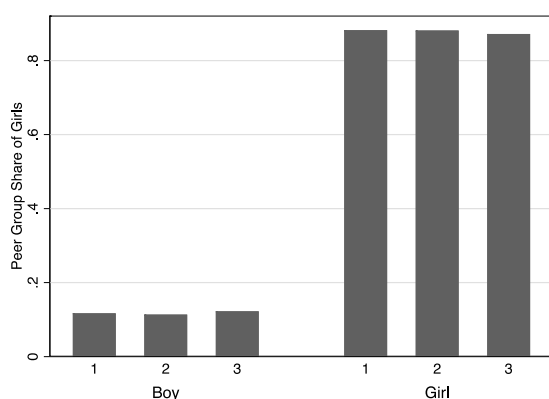
For the longitudinal analysis, we generate wave-specific data files that capture the gender composition of students' friendship networks, and then combine the wave-specific data files:

```
foreach w in 1 2 3 {
    preserve
    keep if wave == `w'
    npinfo a1_01*, id(id_p) npcov(girl) replace
    egen sh_girl = rowmean(a1_01??_girl)
    tempfile wave_`w'
    save wave_`w', replace
    restore
}

* Append Wave-Specific Data Sets *
clear all
use wave_1.dta, clear
append using wave_2.dta
append using wave_3.dta
```

We first consider how the gender composition of friendship networks varies across waves:

```
graph bar sh_girl, over(wave) over(girl) ytitle("Peer Group Share of Girls")
```



We also find strong gender segregation in friendship networks in the longitudinal analysis, though networks seem to be slightly more integrated in the third wave. Now, we replicate the analysis on partying behavior with linear fixed-effects models on the basis of the longitudinal data:

```
egen miss = rowmiss(girl sh_girl)

xtset id_p wave

xtreg party c.sh_girl if miss == 0, fe robust
est store m1
xtreg party c.sh_girl i.girl#c.sh_girl if miss == 0, fe robust
est store m2

esttab m1 m2
```

	(1) party	(2) party
sh_girl	-0.0130 (-0.29)	0.142* (2.43)

1.girl#c.s~1		-0.324*** (-3.73)
_cons	0.548*** (25.38)	0.609*** (21.82)
-----		
N	5639	5639
-----		
t statistics in parentheses		
* p<0.05, ** p<0.01, *** p<0.001		

As in the cross-sectional analysis, we find that the gender composition of the friendship network has different consequences for boys and girls. For boys, making more female friends tends to be associated with more partying behavior. For girls, on the other hand, acquiring more female friends over time is associated with a decrease in partying.

## 8. References

- Dollmann, J., Jacob, K., & Kalter, F. (2014). Examining the Diversity of Youth in Europe. A Classification of Generations and Ethnic Origins Using CILS4EU Data (Technical Report). *MZES Working Paper No. 156*, Mannheimer Zentrum für Europäische Sozialforschung.
- Huisman, M. E., & Snijders, T. A. B. (2003). Statistical Analysis of Longitudinal Network Data with Changing Composition. *Sociological Methods & Research*, 32, 253–287.
- Leszczensky, L., & Gräbs Santiago, A. (2014b). Ethnische und nationale Identität von Kindern und Jugendlichen, in: D. Danner, & A. Glöckner-Rist (Eds.), *Zusammenstellung sozialwissenschaftlicher Items und Skalen*. <http://zis.gesis.org/ZisApplication/DoId/zis158>
- Leszczensky, L., & Gräbs Santiago, A. (2015). The Development and Test of a Measure of Young Immigrants' Ethnic and National Identity. *Methods, data, analyses – mda*, in press.
- Leszczensky, L., & Pink, S. (2015). Ethnic Segregation of Friendship Networks in School: Testing a Rational-Choice Argument of Differences in Ethnic Homophily between Classroom- and Grade-Level Networks. *Social Networks*, 42, S. 18-26.
- Snijders, T. A. B., van de Bunt, G. G., & Steglich, C. E. G. (2010). Introduction to Stochastic Actor-Based Models for Network Dynamics. *Social Networks*, 32, 44–60.
- Valente, T. W., Fujimoto, K., Unger, J. B., Soto, D. W., & Meeker, D. (2013). Variations in Network Boundary and Type: A Study of Adolescent Peer Influences. *Social Networks*, 35, 309–316.
- Veenstra, R., Dijkstra, J. K., Steglich, C., & van Zalk, M. H. W. (2013). Variations in Network Boundary and Type: A Study of Adolescent Peer Influences. *Journal of Research in Adolescence*, 23, 399–412.

### 8.1. Publications based on the data

- Leszczensky, L., & Gräbs Santiago, A. (2015). The Development and Test of a Measure of Youth's Ethnic and National Identity. *methods, data, analyses*, 9, 87-110.
- Leszczensky, L., & Pink, S. (2015). Ethnic Segregation of Friendship Networks in School: Testing a Rational-Choice Argument of Differences in Ethnic Homophily between Classroom- and Grade-Level Networks. *Social Networks*, 42, 18-26.
- Leszczensky, L. (2016). *Tell Me Who Your Friends Are? Disentangling the Interplay of Young Immigrants' Host Country Identification and Their Friendships with Natives*. Mannheim: University of Mannheim. (Dissertation)
- Leszczensky, L., & Pink, S. (2017). Intra- and Inter-group Friendship Choices of Christian, Muslim, and Non-religious Youth in Germany. *European Sociological Review*, 33, 72-83.
- Stark, T. H., Leszczensky, L., & Pink, S. (2017). Are there Differences in Ethnic Majority and Minority Adolescents' Friendship Preferences and Social Influence with Regard to their Academic Achievement? *Zeitschrift für Erziehungswissenschaft*, 20, 475-498.
- Jugert, P., Leszczensky, L., & Pink, S. (2018). The Effects of Ethnic Minority Adolescents' Ethnic Self-Identification on Friendship Selection. *Journal of Research on Adolescence*, 28, 379-395.
- Kretschmer, D., Leszczensky, L., & Pink, S. (2018). Selection and Influence Processes in Academic Achievement - More Pronounced for Girls? *Social Networks*, 52, 251-260.

- Leszczensky, L. (2018). Young Immigrants' Host Country Identification and their Friendships with Natives: Does Relative Group Size Matter? *Social Science Research*, 70, 163-175.
- Fleischmann, F., Leszczensky, L., & Pink, S. (2019). Identity Threat and Identity Multiplicity among Minority Youth: Longitudinal Relations of Perceived Discrimination with Ethnic, Religious and National Identification in Germany. *British Journal of Social Psychology*, 58, 971-990.
- Leszczensky, L., Jugert, P., & Pink, S. (2019). The Interplay of Group Identifications and Friendships: Evidence from Longitudinal Social Network Studies. *Journal of Social Issues*, 75, 460-485.
- Leszczensky, L., & Pink, S. (2019). What Drives Ethnic Homophily? A Relational Approach on How Ethnic Identification Moderates Preferences for Same-Ethnic Friends. *American Sociological Review*, 84, 394-419.
- Jugert, P., Leszczensky, L., & Pink, S. (2020). Differential Influence of Same- and Cross-Ethnic Friends on Ethnic-Racial Identity Development in Early Adolescence. *Child Development*, 91, 949-963.
- Pink, S., Kretschmer, D., & Leszczensky, L. (2020). Choice Modelling in Social Networks using Stochastic Actor-Oriented Models. *Journal of Choice Modelling*, 34, 100202.
- Jugert, P., Pink, S., Fleischmann, F., & Leszczensky, L. (2020). Changes in Turkish- and resettler-origin adolescents' acculturation profiles of identification: A three-year longitudinal study from Germany. *Journal of Youth and Adolescence*.
- Leszczensky, L., & Pink, S. (2020). Are Birds of a Feather Praying Together? Assessing Friends' Influence on Muslim Youth' Religiosity in Germany. *Social Psychology Quarterly*, 83, 251-271.

## 8.2. Field reports

- Leszczensky, L. (2012). Dokumentation des Kognitiven Pretests im Rahmen des Projektes "Soziale Netzwerke und ethnische Identifikationen von jugendlichen Migranten". Mannheimer Zentrum für Europäische Sozialforschung.
- Leszczensky, L., & Pink, S. (2012). Dokumentation des Pretests im Rahmen des Projektes "Soziale Netzwerke und ethnische Identifikationen von jugendlichen Migranten". Mannheimer Zentrum für Europäische Sozialforschung.
- Leszczensky, L., & Pink, S. (2012). Dokumentation des Instrumententests im Rahmen des Projektes "Soziale Netzwerke und ethnische Identifikationen von jugendlichen Migranten". Mannheimer Zentrum für Europäische Sozialforschung.
- Leszczensky, L., & Gräbs Santiago, A. (2014a). Die Messung ethnischer und nationaler Identität von Kindern und Jugendlichen. *MZES Working Paper No. 155*, Mannheimer Zentrum für Europäische Sozialforschung.
- Leszczensky, L., Kalter, F., & Pink, S. (2014). Freundschaft und Identität in der Schule: Feldbericht zu Welle 1 und Welle 2 (Technical Report). *MZES Working Paper No. 157*. Mannheimer Zentrum für Europäische Sozialforschung.
- Leszczensky, L., Kalter, F., & Pink, S. (2015). Friendship and Identity in School - Field report on Wave 1, Wave 2, and Wave 3 (Technical Report). *MZES Working Paper No. 161*. Mannheimer Zentrum für Europäische Sozialforschung.