Universiteit Leiden

Master Computer Science

Measuring Community Social Capital through the Structure of a Population-scale Social Network

Name:	Bart de Zoete
Student ID:	s2030098
Date:	12-07-2022
Specialisation:	Data Science
1st supervisor:	dr. Frank Takes
2nd supervisor:	dr. Eelke Heemskerk

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Social capital is a social scientific concept defined as the value found in social structures that enable a group of people to function effectively. Concretely, this concept has been linked to numerous outcomes such as better perceived health and lower crime rates. Existing works using social network data to represent social structures and measure social capital suffer from three main problems, namely a) a lack of empirical evidence that what is measured actually is social capital, b) highly specific types of social capital are considered, and c) sampling problems of the considered social network data. We overcome these issues by examining the relation between network structure and social capital using a unique social network of the entire population of the Netherlands. Specifically, we use regression analyses to determine the direct relations between network measures and a broad range of social capital outcomes available from open source data. We propose a new network measure called the ego network growth rate, which captures resources beyond the direct neighborhood. This measure is shown to relate to node centrality and is therefore used to measure social bridging. Social bonding is measured using excess closure. We utilize and extend a network analysis development framework to efficiently compute these measures and perform analyses at the unprecedented scale of millions of people and hundreds of million of connections. The results of our analyses shed new empirical light on the relations between network structure and social capital. Our findings suggest that network structure has a strong relation to social capital. Social bridging seems positively related to social capital, whereas social bonding has a negative effect. Overall, this work establishes the value of network-based measurement of social capital in a population-scale social network.

Contents

Chapter 1

Introduction

Social capital is one of the foundational concepts in modern-day social scientific research. It is roughly defined as the value found in social structures that can facilitate action [Col88, AK02, Lin17, Put01]. Decades of research have shown this notion to relate to a multitude of interesting and important phenomena. For example, higher levels of social capital have been linked to lower violent crime rates [WKK98, P$^+$00], improved perceived health and happiness [Poo06, SKK02, Put01], and more effective regional governments [LNP01]. This makes social capital a valuable to help understand and explain societal phenomena.

Given the use of social capital in explaining a wide range of societal and individual outcomes, it is only natural that its measurement has received a lot of attention from researchers as well. Such measurements could be used, for example, by governments to target interventions as to combat community decline [Sto01]. As the concept is broad however, there is no single agreed upon way to measure social capital. Most "traditional" works measure social capital indirectly by looking at aggregated (survey) data from demographic groups on theoretically expected outcomes of social capital. The assumption is then that if the outcome is present in empirical data, this indicates that social capital itself must be present as well. This approach is interesting as it can have direct practical applications. For example, the government can set up programs in communities with low social capital to give the people a larger say in government institutions, such as the police and schools, as to help improve their community social capital [War01].

One downside of this approach is that social structure is not taken into account, despite this being an important aspect of the definition of social capital. Social networks offer a solutions here, as this type of data directly captures social structures by representing people as *nodes* and the connections between them as *edges*. Networks have been used increasingly to measure social capital in various contexts, see for example [LLY13, BAGADF15, HA13] where scientific co-authorship networks and Twitter follower networks are analyzed. However, there are three main problems with existing works. First, they measure social capital using some network measure which is chosen solely based on theory. There is often no empirical analysis to verify that these network measures are actually related to social capital. In most networks this is not possible at all, because little or no information is available about the people besides the network structure. Second, highly specific kinds of social capital are considered. For example, in a scientific co-authorship network we might consider the social capital of a researcher to be their citation count. When such a specific type of social capital is considered, it is unclear how the knowledge translates into different domains or what it means for general social capital measurement. As a result, such measurements do not have the same practical implications as exist for traditional measurement techniques. That is, the measurement cannot be used to directly target interventions as to increase social capital for certain people or communities. Third, these works often use small-scale sampled data, which can lead to conclusions that are difficult to replicate or place into a broader context, especially in the face of data quality problems. Because of these issues, the precise empirical relations between social structures and social capital have remained largely unknown.

In this thesis we consider the measurement of social capital in a network where the results have valuable real-world applications, and where the measure has been shown to relate to social capital. We determine the relation between network structure and social capital in a population-scale social network. This new kind

of social network has an enormous scale, containing an entire well-defined population of people, in this case the population of the Netherlands. Using a statistical approach, we assess the power of network measures in predicting well-studied social capital outcome measures. Network structural aspects that we consider are social bonding and social bridging. Social bonding captures the extent to which people form tight-knit social circles, whereas social bridging relates to the position that people have in a network which allows them to effectively spread information. Additionally, we consider social capital on a community level. A community here is a certain group of people that are in some way related to one another, such as people of a certain demographic group or that live in the same area. Our analyses will give insights into the true value of social network data in the measurement of social capital.

The unique large-scale network data [VdL22] used in this thesis is made by Statistics Netherlands[1] (CBS) and contains various so-called formal ties among the entire Dutch population. These ties are grouped into layers containing for example colleague or family connections. A total of 17.2 million people are represented in the network, making it a uniquely large social network. Because of its size, working with this data is highly computationally demanding and many commonly used network measures are not feasible to compute. The network also contains additional attributes for each person such as their age and sex, as well as the geographical neighborhood in which they live. Using this attribute, we can define communities to be people living in the same neighborhood. This makes it possible to link existing data on various common social capital outcomes at the neighborhood level that are publicly available. For this we pick four neighborhood-level outcomes: the percentage of people with good perceived health, the number of reported crimes per one thousand people, the percentage of people receiving social assistance benefits, and the percentage of people that do volunteering work. The number of deaths per one thousand people is used as a robustness check.

Our approach is based on interpretable statistical models that estimate the relations between network measures and social capital outcomes. In each model, the dependent variable is one of the social capital outcomes, and the independent variables that are used to model the outcome consist of both control variables and network measures. To measure social bonding we use excess closure [Bok]. This is similar to the clustering coefficient which measures the portion of triangles (a group of three people that are all linked to each other) over the total number of triangles that could exist. Excess closure only considers triangles that consist of at least two different layers. Our other network measure is the ego network growth rate, which looks at the relative change between the direct ego network size and the size one step further away. This measure relates to node centrality and thus to social bridging. We also take the average number of connections into account, which acts as a moderating variable and is needed in order to make sense of the former two network measures. We compute these network measures for each of the roughly 17.2 million people in the network, and aggregate these to the community level on which we perform our analysis. This aggregation step is done by averaging the values from all individuals in the geographical neighborhood. Control variables are derived from the node attributes and contain, for example, the average age and the percentage of highly educated adults in the neighborhoods. Using these models, we can interpret the importance of each independent variable in the prediction of the social capital outcomes. Finally, we can use this knowledge to conclude how the network structural aspects relate to community social capital.

This research was conducted as part of the POPNET project[2]. POPNET (POPulation-scale NETwork analysis) is a research project that aims to advance the methods and applications of population-scale social network data for solving societally relevant problems. Various sub-projects of POPNET focus on matters such as network anonymity [DJ], the creation of a platform where other researchers can responsibly use the network, and various other social scientific research directions. One aim of POPNET is to take initial steps towards creating network-based official statistics that can be used by policy makers. As we set a first step towards network measurement of social capital in population-scale data, this thesis contributes to that goal.

The main research question that we will focus on is:

What is the relationship between network structure and social capital?

[1] `https://www.cbs.nl/en-gb`
[2] `https://www.popnet.io/`

In the remainder of this thesis we propose a methodology and empirically assess its suitability for answering this question. We start by discussing the relevant literature relating to social capital in Chapter 2. We then provide detailed descriptions of the network data and what parts of it are used, as well as the social capital outcome data in Chapter 4. Our research design, network measures and regression analysis are detailed in Chapter 5. We then go on to discuss our experiments and findings in Chapter 6, where we also provide a discussion on the limitations of our work. Lastly, we finish with a conclusion in Chapter 7.

Chapter 2

Related work

In this chapter we discuss work related to our research. We first go into detail on the topic of social capital, after which we look at how it is usually measured.

Social capital is an important concept in social scientific research. Roughly speaking, it relates to the value found in social structures that can facilitate action [Col88, AK02, Lin17, Put01]. It is often considered at either the individual or the community level. *Individual social capital* relates to "how a person can leverage their social ties for personal gain" [Bor98]. This is especially relevant in scenarios where new resources are sought and obtained [Lin17], and is closely tied to financial and human capital [Bur92]. *Community social capital* on the other hand, is considered to be "a property of a group of people that is not only determined by the network structure, but also by trust and shared norms and values" among the group members [P$^+$00]. Below, we only discuss literature relating to community social capital, as this is the type considered in our research.

One of the founding works in social capital research [Col88] lays out the theory for a possible connection between the cultural and network structural aspects of social capital. When two people connected to a certain person and they are also connected to each other, they can combine forces to sanction that person. This is called network closure, and when it is missing, it makes norms and trust less likely to arise in the network.

In another fundamental work [Bou18], it is said that social groups are formed by individuals which invest in the group for their personal gain. However, all members of a community can benefit from the capital held in that community, showing how there is inherent value in communities. Furthermore, it is explained how all forms of capital can essentially be (nearly) reduced to economic capital.

Most significant in the community social capital literature is its measurement through levels of civic participation [LNP01, P$^+$00]. Here it is said that when civic participation rates are higher, people are more involved with their community which improves the social capital. Using data on various forms of civic participation, a convincing case was built that social capital in the United States has decreased drastically since the 1950s [P$^+$00]. On a state level, correlations were found between social capital and many outcome variables such as child welfare, decreased murder rates, and tolerance [P$^+$00].

Community social capital has been linked to various other outcomes, for example: more effective maintaining of law and order [HMB95], lower rates of violent crime [WKK98], better perceived health [Poo06, SKK02], lower overall mortality rates [KKL97], and even lower rates of suicide [Dur05]. Additionally, high levels of societal trust are believed to be necessary for large companies to arise [Fuk96].

Most commonly, community social capital is measured by actually measuring outcomes of social capital as indicators of its existence [Sto01]. Say a certain societal outcome should theoretically arise when social capital is high, and the outcome is observed to be high in practice, then this must mean that social capital is high. Indeed there is circular reasoning here, because we let social capital be defined entirely by the outcome. Therefore saying that social capital is high since the outcome is high, is stating the same twice. It can also be questioned what the directionality of the relationship is: Does social capital cause the outcome, of does social capital arise as a result from the outcome being present? Nevertheless, the consensus in social capital

researchers appears to be that this method of measurement is valuable, despite aforementioned issues.

The two main outcomes that are used in this way to measure social capital are to look at civic participation, and trust and cooperative norms [SS13, LST17]. These approaches often work with samples of a population, created using surveys [LKK99, P$^+$00]. Regarding civic participation, surveys often ask whether or not people are involved with certain organizations such as sports groups or churches. For trust, surveys often ask matters such as how much trust people have in strangers or people in their geographical neighborhood. Individual answers are then aggregated to represent the social capital of the community.

In networks, social capital can in theory be measured using measures of network cohesion [BJE98, SL08]. Examples of this are network density, average path length, and attribute assortativity. Since the clustering coefficient [WS98] measures the portion of triangles in the network that are closed, this closely relates to the closure theory [Col88] described before. As the network which we will use has unique properties (described in Section 4.1.2), many of these measures need to be adjusted to accommodate, for example, the multilayer aspect of our type of network.

Concludingly, although there is a theoretical link between the aforementioned network measures and social capital, little work exists that actually measures social capital in network data. One example is in [AWH14]) where a co-authorship network is used to show the relation between the degree and betweenness centrality of authors and their h-index, which can be seen as a sort of measure for (individual) social capital in that context. Social capital as measured through questionnaires has also been linked to network measures using online social network data, for example, in [EVGL14]. The measurement of social capital in population-scale, real-world networks as we will do in this work is a novel and exciting new avenue of social capital measurement in network data, that promises to alleviate a number of conceptual and methodological challenges of existing approaches.

Chapter 3

Preliminaries

This work uses terminology from the field of network science. In this chapter we explain network concepts that are used throughout this thesis, using a minimal amount of formal notation.

A *network* or *graph* is a mathematical structure consisting of a set of *nodes* (also *vertices*) and *edges* (also *links* or *connections*) between the nodes. Nodes usually represent real-world objects such as people, whereas edges describe the connections between the nodes. When various types of edges are present in a network, this is called a *multilayer network*, since every different kind of edge can be seen as a separate layer of the overall network. We can also consider different edge types to be part of a single layer. The network at hand (further described in Chapter 4) is such a multilayer network. Here we have various layers such as Close Family where we find link types such as sibling, child, and parent. The nodes in social networks represent real people, in our case Dutch citizens.

Formally, the multilayer network that we use can be defined as a graph $G = (V, E, L, A)$, where V is the set of nodes, E is the set of edges, L is the set of available layers, and A is the set of attributes. As each layer consist of various link types, we consider $L = \{L_1, L_2, ..., L_g\}$, where each set L_i contains the link types from that layer. We also consider $A = \{A_1, A_2, ..., A_{|V|}\}$ such that each set A_i contains the attribute values for node $v_i \in V$. Edges are defined as triplets $(v, w, l) \in E$, where $v \in V$ is the starting node of the edge, $w \in V$ is the receiving node, and $l \in L_i \in L$ denotes the link type of the edge.

Some additional important network concepts are:

- *Degree* of a node is defined as the number of edges that are connected to it. In multilayer networks there can be several edges between people, so we consider the degree to be the total number of nodes that are connected to a node.
- *Paths* are sequences of edges such that the ending node of each edge is the starting node of the next edge.
- *Path length* is the number of edges in a path.
- *Shortest path* between two nodes is the path of minimal length which connects them.
- *Distance* between two nodes is the length of a shortest path between those two nodes.

  The formula $distance(v, w)$ denotes the distance between nodes $v, w \in V$. For nodes between which there is no path, we define $distance(v, w) = \infty$. The distance between a node and itself is defined as $distance(v, v) = 0$.
- *Neighbors* of a node at a given distance are all the nodes that are at most at that distance from the considered node. Here we do not include the node itself. This is also called the *open neighborhood*.

  The neighbors of a node $v \in V$ up to distance d are given by: $N_d(v) = \{w : 0 < distance(v, w) \leq d\}$.
- *Ego networks* are the networks that are obtained when isolating a node along with its neighbors up to a given distance from the ego node. This includes all edges among the selected nodes.

The ego network of a node $v \in V$ in a multilayer network for distance d thus contains the set of nodes $V' = N_d(v) \cup \{v\}$ and edges $E' = \{(v, w, l) : v, w \in V' \wedge (v, w, l) \in E\}$ for any existing l.

- *Ego network size* is the number of nodes in a ego network.
- *Connected* nodes have a path between them.
- *Triangles* in networks are groups of exactly three nodes that are all connected to one another.
- *Clustering coefficient* measures the portion of existing triangles in an ego network at $d = 1$ over the theoretically possible number of triangles given the degree of the ego. This is commonly used to measure network closure.
- *Diameter* of a network is the length of its longest shortest path.
- *Complete networks* are networks where there exists an edge between every pair of nodes.
- *Bipartite networks* are networks where the nodes can be split up into two sets, such that all edges lie between these sets and no two nodes within one node set are direct neighbors.
- *Connected components* (CC) are groups of nodes in a network such that each pair of nodes in the component is connected. In many real-world networks the largest CC contains the vast majority of all the nodes, and we speak of a *giant component* (GCC in short).

Chapter 4

Data

In this chapter we discuss the data used in our research. We first discuss the network which is used for measuring social capital. Here we look at the types of relations to which we have access, as well as the rationale behind our data selection process, and we provide an interpretation of how the resulting network is a good reflection a true social network. We then discuss what social capital outcome data we use to see how our network measures relate to social capital.

4.1 Network data and selection

The network used in this research is created by CBS, the Dutch national statistics office. This is a government body that operates independently by law, which creates reliable official statistics on the Dutch population. Recently, CBS created a network of the entire Dutch population [VdL22] where people are connected through various social connections. We call this the *Persons Network*, after its Dutch name of "Personennetwerk". It was derived using official register data from 2018, and as a result it contains every person registered in the Netherlands in that year.

4.1.1 Formal and informal ties

Generally speaking, the Persons Network contains various so called *formal ties*. These are ties that can be derived from government register data and where we are certain that they exist (excluding a negligible number of errors in the registers themselves). For example, there is a government register where the parents of each person are stored, and from this information other family relations can be derived such as whether people are grandparents or cousins. These ties are guaranteed to exist, but that does not mean that they are always social ties to infer social structure from. Consider for example that people can be disconnected from their family after a big fight, or that people might not know everyone at their company.

In contrast to the formal ties in the Persons Network, *informal ties* are not formally registered. A clear example of this are friendships, which are typically captured by online social networks – though often along with additional spurious ties. Since these ties are not present in register data, they are not present in the Persons Network.

4.1.2 Network layers

The specific social ties that are present in the network can be grouped into various layers. Table 4.1 contains a brief overview of the available layers along with some example links and other properties. In the layers where the "Complete components" column is ticked, the network can be split up into components that are all complete networks. These are projections of bipartite networks, with in one node set the people and in the other the affiliations (workplaces, households, classes).

Layer	Clarification	(Example) connections	Complete components	Sampling
Close family	Family members that likely form a household or have done so in the past.	Child, sibling, partner		
Extended family	Family connections one step outside close family.	Niece, aunt, grandchild		
Household	People that manage their lives together.	Housemate	✓	
Work	People working for the same company.	Colleague	✓	✓
School	People in the same education institute, in the same program and that started in the same year.	Classmate in primary education, classmate in higher education	✓	
Neighborhood	People that live close by.	Neighbor		✓

Table 4.1: Characteristics of layers available in the Persons Network.

Figure 4.1 shows a visualization of the ego network of a random node in the Persons Network. We can see how the workplace and school class of the ego create densely connected clusters in the network. The family connections also form dense regions with relatives that live together also standing out.

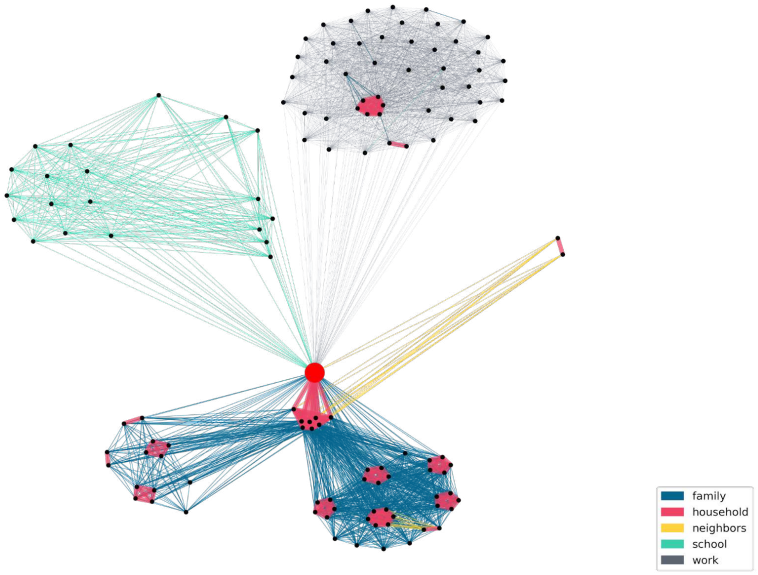


Figure 4.1: Visualization of the ego network of a node in the Persons Network.

We will only briefly discuss some of the quality issues and considerations based on which we select what data to use for this research. An in-depth exploration of each layer and an assessment of the data quality can be found in [BBH+21].

First, sampling has been used to derive the Neighborhood and Work layer, which causes several artifacts such as important links being left out. Second, complete layers are problematic for our research purposes. This is because if people work at a large company or are part of a large study program, then they will have many hundreds or even thousands of connections. In such cases we can realistically never expect everyone

to actually know one another. We thus find a large number of spurious ties in the Work and School layers, that likely do not represent meaningful social ties on which to base a measurement of social structure and ultimately social capital.

With smaller components however, the latter problem can be overcome. Consider the following: at work people are often close to, say, ten colleagues. Each of these also have their circle of colleagues to whom they are close. When the company is sufficiently small, it is quite likely that there is a meaningful link between everyone in that company, albeit a link that in real life is at a distance of two. Similarly in school, an individual might have a handful of friends, who all have friends as well, and through their connections they likely have some connection to classmates given that the class is sufficiently small. Therefore in such small components, we can be reasonably certain that everyone is connected through a real-life link of either one or two steps. And thus in our network, such links of distance two are represented as links of distance one. Such links are still valuable as they do say something about the resources to which a person has access.

In fact, this is similar to selecting both close and extended family. Despite people often being close to their extended family, we believe that is it still common to access the resources that are present there through direct family. Thus extended family connections are at a distance of two in real life. But when both layers are used, such links are shown in our data as being at a distance of one.

4.1.3 Selected layers

Based on these quality considerations, we leave the Neighborhood layer out of our selection. For the Work and School, we choose to only select components that contain at most a hundred nodes. This cutoff is indeed largely arbitrary, but it makes it possible to capture a large number of meaningful ties whilst guaranteeing a certain level of data and interpretational quality. Moreover, it avoid the sampling problems of the Work layer. Finally, all household and all family links are selected.

After making this selection, we obtain a slice of the Persons Network containing the following layers:

- Close family
- Extended family
- Household
- Work for workplaces of at most 100 workers
- School for groups of at most 100 students

4.1.4 Network data descriptives

Table 4.2 shows descriptive statistics of our network (see Chapter 3 for definitions of the statistics).

Layer	# nodes with a link	# edges	Density	% nodes in largest CC	Avg. Excess closure
All	16,879,254	505,920,731	$3.398 \cdot 10^{-6}$	97.990	$8.517 \cdot 10^{-3}$
Close family	16,735,724	79,392,532	$5.332 \cdot 10^{-7}$	91.820	-
Extended family	15,819,843	234,644,115	$1.576 \cdot 10^{-6}$	97.614	-
Household	14,054,443	32,666,006	$2.194 \cdot 10^{-7}$	0.003	-
Work	2,587,661	76,199,234	$5.117 \cdot 10^{-7}$	0.004	-
School	3,048,956	115,363,963	$7.747 \cdot 10^{-7}$	0.003	-

Table 4.2: Descriptive statistics of the slice of the Persons Network that is used in this research.

We observe that the vast majority of all nodes are connected, and that nearly all of these form a giant component together. Within the Close family, Extended family and Household layers, we also find that almost all nodes are connected. Giant components are also present for these layers, except for the Household layers. This is because the household layer consist of complete components, and so the largest CC has a size equivalent to the largest complete component. The same holds true for the Work and School layers. Within the Work and School layers we also find a large portion of nodes having some connection, despite not selecting components of more than a hundred links.

4.1.5 Interpretation of the network structure

In our selection, the Close family, Extended family and Household layers together can be thought of as the backbone of the network. These layers consist of strong ties and the links form a lot of locally dense regions in the network. Additionally, almost every node is part of the giant component when we consider these links, meaning that most people are connected through some path. We thus can view these links as forming the backbone of the network where nearly everyone is present but there are still great distances between people. This is alleviated by adding the Work and School layers. Such links are perhaps intuitively less strong, but they connect different parts of the network that otherwise would appear to be distant. Thus altogether we have a network where some ties facilitate local clustering whereas others connect different parts of the network. In this way, the network has structural properties that we would expect in real-world networks, building further confidence that the network can be used to study the Dutch population.

We believe that the network with the selected links does a good job at mimicking the true (theoretical) social network of the Netherlands. Although we do not capture informal ties which certainly are important, the ones that do not overlap with Family, Work and School connections only make up a relatively small portion of the social ties that people have. And despite leaving out some links, we do capture some of the most important social ties for either all or most people. The quality and completeness of the ties which we use is also high, and they are expected to have true importance in the real world in the vast majority of cases. Because of this, we believe that network measures computed on this network can be used effectively to relate network structure to community social capital.

4.2 Outcome data

In order to relate network measures to social capital, we make use of four main neighborhood-level social capital outcomes, and a fifth outcome that we use as a robustness check. A brief overview of the outcome data can be found in Table 4.3.

These datasets were selected to both cover a wide range of social capital outcomes and also capture outcomes that are most strongly supported by the community social capital literature (see Chapter 2).

Whereas the Crime, SAB, and Mortality outcomes are derived from exact register data, the Health and Volunteer outcomes come from the *Gezondheidsmonitor* of 2016. This is a large-scale questionnaire which is performed every four years in the Netherlands by CBS, the Area Health Authority (GGD) and the National Institute for Public Health and the Environment (RIVM) in order to get an insight into the health of the Dutch population. Results from these questionnaires are used by government agencies to guide policy, but are also of interest to scientists as they each gather diverse information on a large scale. The questionnaire data is used by RIVM to create a model which predicts the answers for the 387,195 people that have filled in the Gezondheidsmonitor in 2016 based on register data such as age, sex and income [vdKZB$^+$17]. This is then used to predict outcomes on the neighborhood-level. The models made by RIVM are of high quality, and they are based on a large number of people – over 2% of the entire Dutch population. This makes it reasonable to assume that if we find relations between network measures and the outcomes, that these are representative of the actual relations that we would find if the exact data existed.

Name	Description	Ages	Source	Year	Related social capital outcome
Health	The percentage of people with good or very good perceived health.	19+	Gezondheidsmonitor[1] GGD, RIVM, CBS	2016	Improved health
Volunteer	The percentage of people that do some form of volunteering work.	19+	Gezondheidsmonitor GGD, RIVM, CBS	2016	Increased civic participation
Crime	The number of reported violent crimes per one thousand residents.	-	Geregistreerde criminaliteit[2] CBS	2018	Lower violent crime rates
SAB	The percentage of people that receive social assistance benefits (SAB) from the Dutch government.	15+	Bijstandsuitkeringenstatistiek[3] CBS	2018	Improved socio-economic position
Mortality	The number of deaths per one thousand residents.	-	Kerncijfers Wijken en Buurten[4] CBS	2018	Improved health

Table 4.3: The used datasets of neighborhood-level social capital outcomes.

An additional problem is that the model made by RIVM makes use of the household composition, and this also is reflected in our network structure. In turn, the household size will have an impact on the network measures which we compute. As a result, some of the information that is used in the RIVM model also effects the network measures. This biases our results for the Volunteer and Health outcomes to a certain extent. For Health we can perform a robustness check of the results by using an additional health-related outcome which is based on register data. To this aim we make use of the Mortality outcome. The exact relations between network measures and Health might be different from those obtained with Mortality, because mortality rates are much more age dependent than perceived health. Nevertheless, if we also find network measures to be important for Mortality, then we can say with greater certainty that network measures have an effect on health outcomes. For Volunteer there sadly is, to the best of our knowledge, not another related dataset available on a neighborhood level which we could use in the same way.

We only use neighborhoods where all of the outcome data is available. This leaves us with data on 2,772 out of the in total 3,086 neighborhoods (in Dutch "wijken") in the Netherlands as per December 2018, covering roughly 90% of all neighborhoods. In terms of inhabitants, we cover 16,468,685 out of 17,257,207 people that were registered in the Netherlands in 2018, which is approximately 95% of the population.

Figure 4.2 shows the distribution of the four social capital outcome measures over the Netherlands. The colors correspond to the quartiles, so the lowest 25% of values are assigned to 'very low', the next 25% are considered 'low', and so on. The upper two outcomes positively relate to social capital, whereas the bottom three associate negatively. Blank spots are caused by the neighborhood being dropped due to missing data.

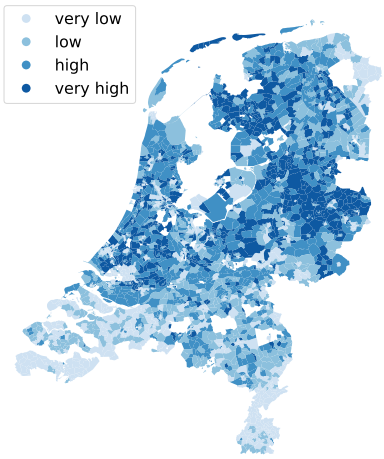
As the east of the country is more rural, we see a clear link between urbanization and the Volunteer outcome. We also observe that many neighborhoods fall into different bins than adjacent neighborhoods, showing that the urbanization level likely is not the sole contributing factor to varying levels in the outcomes.

[1] https://statline.rivm.nl/portal.html?_la=nl&_catalog=RIVM&tableId=50089NED&_theme=85

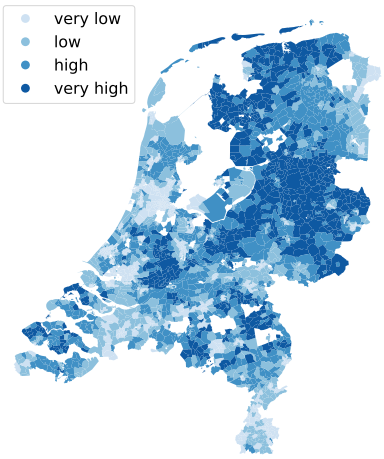
[2] https://www.cbs.nl/nl-NL/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksbeschrijvingen/geregistreerde-criminaliteit/

[3] https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksbeschrijvingen/bijstandsuitkeringenstatistiek--bus--

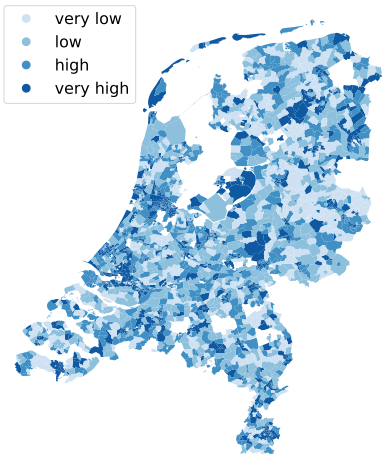
[4] https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksbeschrijvingen/kerncijfers-wijken-en-buurten



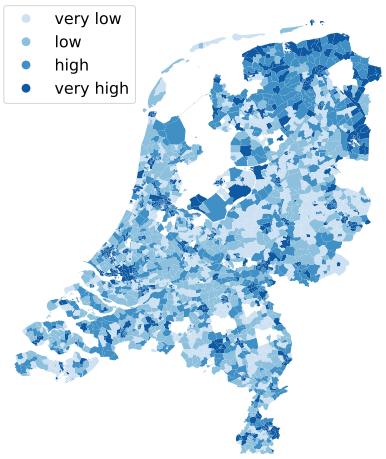
(a) Health – Percentage of people with good perceived health.



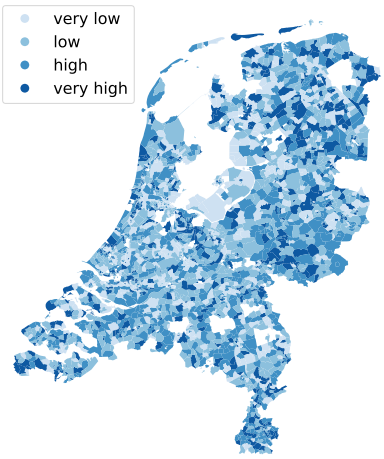
(b) Volunteer – Percentage of people that do volunteer work.



(c) Crime – Number of reported crimes per one thousand residents.



(d) SAB – Percentage of people that receive social assistance benefits.



(e) Mortality – Number of deaths per one thousand residents.

Figure 4.2: Distributions of social capital outcomes over the Netherlands.

Chapter 5

Methodology

In this chapter we describe the methodology of our research. First we describe the overall goal of our analysis and how we designed our research to achieve that goal. We then explain the network measures which we have considered. And lastly, we discuss how we examine the relations of these network measures to the social capital outcomes.

5.1 Research design

The ultimate goal of our analysis is to be able to draw conclusions about the relation between network structure and social capital. We achieve this indirectly through the examination of the relations that different network measures have with social capital outcomes.

An overview of the general design of our research can be found in Figure 5.1. This diagram shows that we first compute the network measures and control variables on an individual level (green box), which form the independent variables in our analyses. Each of these roughly 16.5 million data points are then aggregated to the community level (red box) by averaging over all people in the neighborhood. Data on social capital outcomes are gathered (in blue) and do not need to be processed further as they already exist on the neighborhood level. These outcomes are used as dependent variables. With all data being on the same unit of analysis, we can analyze the relations between the independent and dependent variables (blue box). These analyses are set up such that we can determine how the network measures relate to the social capital outcomes. This knowledge is then used to make more broad conclusions on the overarching relationship between network structure and social capital (in purple).

5.1.1 Network measure selection and computation

As we seek to relate network structure to social capital, we focus on use network measures that capture important and diverse aspects of network structure. Moreover, since the outcome data is on the community level, we also need the network measure data to be on the community level. For that we could use network measures that consider the subnetwork of just the people in a community and the connections among them. However this is problematic since we lose a considerable amount of information by discarding links from people within the community to people outside it. Instead we choose to use network measures that are computed for each individual in the entire network, which we then aggregate to the community level. This way we view the community as the sum of its parts and consider all resources to which individuals have access. In a way, this also allows us to say something about the average individual-level social capital of the individuals. We thus compute the network measures for each of the 16,468,685 individuals which are covered by our outcome data.

We specifically focus on two network measures that we believe capture two important aspect of social structure. *Excess closure* is a measure of social bonding, whereas the *ego network growth rate* measures

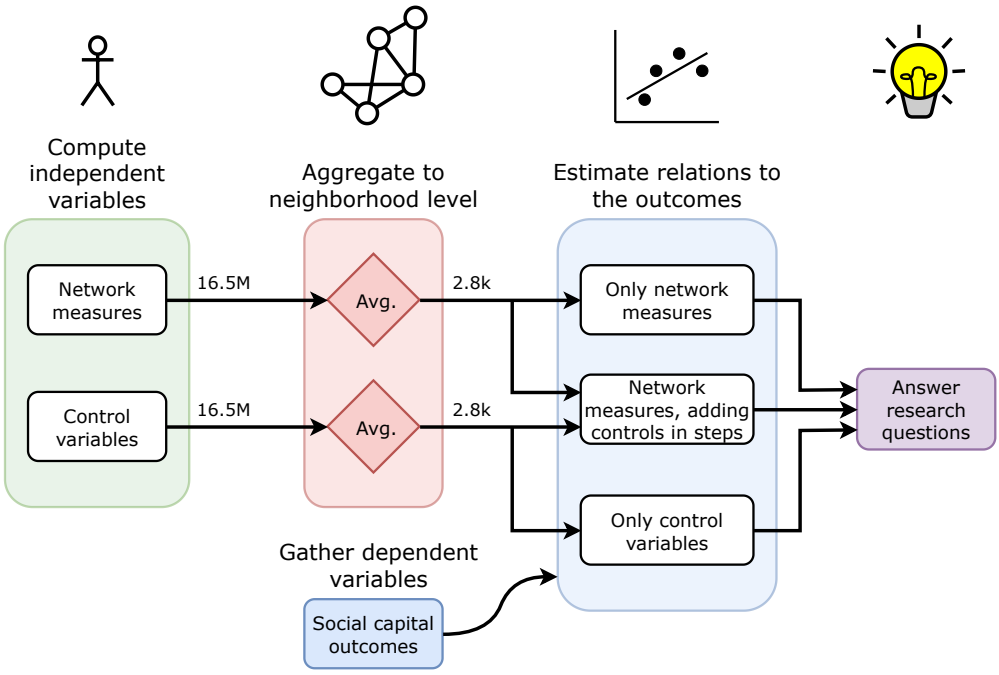


Figure 5.1: Research design overview.

social bridging. Both are moderated by a third network measure, the *degree*. Further explanations of the network measures are provided in Section 5.2.

5.1.2 Community choice and aggregation

Whereas we compute the independent variables on the individual level, our analyses are performed on the community level since we focus on community social capital. We consider communities to be people living in the same geographical neighborhood. There are three reasons for this decisions.

First, it allows us to use high-quality open data on social capital outcomes. There are many open datasets available on the Dutch population which are aggregated to the neighborhood level, some of which relate directly to community social capital. Thus by considering neighborhoods to be communities, we can make use of these datasets to examine how our network measures relate to these aspects of social capital.

Second, other possible choices of communities would be problematic. When using a non-spatial-based unit of analysis it would be difficult if not impossible to find relevant social capital outcome data. This is because such data does not exist for all sorts of demographic groups – certainly not on a population scale. Survey data is also difficult to use as the network data is pseudonymized and quality issues of the network data are more problematic on the individual level.

Third, we believe that neighborhoods can be considered as natural communities in this context. Within neighborhoods there is likely a certain level of connectivity, caused by the connections between people being largely bounded by geographical distance. Additionally, there must be some level of homophily within neighborhoods, as many statistics about the Dutch population vary substantially between neighborhoods, even when they are adjacent (see for example Figure 4.2). For example, if the average education level is higher in a certain neighborhood than in the adjacent neighborhood, then the former has more highly educated people and in that way people are (at least on average) similar.

As the independent variables are computed on the individual level, they must be aggregated to a single number for each neighborhood before the final analyses. This is achieved by averaging the individual values over all people in each neighborhood. The outcome data which we selected are described in Section 4.2.

These data are already available on the neighborhood level, and thus no aggregation is necessary.

5.1.3 Analysis of network measure impact

We perform a set of analyses to see how network measures impact the social capital outcomes. Linear regression models are used that take the network measures along with control variables as independent variables and predict the social capital outcomes. We set these models up so that the coefficients of each independent variable can easily be interpreted, making it possible to determine how each network measure relates to the social capital outcomes. A detailed explanation of the statistical models can be found in Section 5.3.

5.2 Network measures

We consider two network measures of social capital. These are believed to capture different aspects of network structure. A third network measure is used as a moderating variable, which is simply the average degree. As our social capital outcome data covers roughly 16.5 million people, the computation of these measures is a computationally demanding task. For each network measure we explain how it works and how it captures the respective network structural aspects.

5.2.1 Excess closure

The *closure* or *local clustering coefficient* of a node is defined as the portion of closed triangles over the total number of triangles that could exist given the degree of the node. *Excess closure* can be seen as closure for multilayer networks, where only triangles are considered that consist of links from at least two different layers. Figure 5.2 shows excess triangles (highlighted in pink) in a toy ego-network where link type is represented by edge style. The red ego node has 2 excess triangles, as there are just two triangles where several types of edges are present. A detailed description of this measure can be found in [Bok].

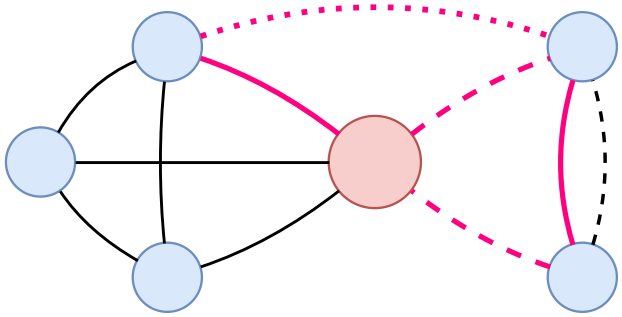


Figure 5.2: Example of excess triangles in a multilayer network.

Only considering triangles that span multiple layers is necessary in the population-scale social network that is considered in this thesis. When we consider all triangles, as is done using the clustering coefficient, we will obtain deceptively large values, since a lot of triangles will always be closed as per the nature of the network data. Namely in the layers that consist of complete components (see Section 4.1.2), all existing triangles are a result of this completeness. Such triangles might be accurate in representing real-world triangles, but they do not provide us with valuable information.

If we have a triangle that spans multiple layers, this signals the presence of more important "triangle closing" ties. Thus when the excess closure of a node is higher, their ego network will be more tightly knit. As a result, excess closure can be seen as a measure of social bonding.

5.2.2 Ego network growth rate

The most common network measure of social bridging is betweenness centrality [Fre77], since it directly measures the extent to which a node bridges densely connected regions of the network. However this measure, as well as other similar centrality measures, are highly computationally expensive and cannot feasibly be computed on our data. Therefore, we require a custom measure that is relatively computationally inexpensive but still has a clear relation to social bridging.

For this we introduce the *ego network growth rate* (ENGR) which we define to be the relative change between the ego network size at a distance of 1 and a chosen distance of d. This is given by the following formula:

$$ENGR_d(v) = \frac{|N_d(v) \cup \{v\}| - |N_1(v) \cup \{v\}|}{|N_1(v) \cup \{v\}|} = \frac{|N_d(v)| - |N_1(v)|}{|N_1(v)| + 1}$$

Here v is the node which we are looking at and $d > 1$ is the distance that we are interested in. Figure 5.3 shows the ENGR in a toy multilayer network using $d = 2$. The red node is the ego, blue nodes are at a distance of 1 from the ego and purple nodes are at a distance of 2. Line style represents the edge type. The ego network size at $d = 1$ here is 6, and at $d = 2$ it is 10. Thus the ego network growth rate for v is $\frac{10-6}{6} = \frac{2}{3}$.

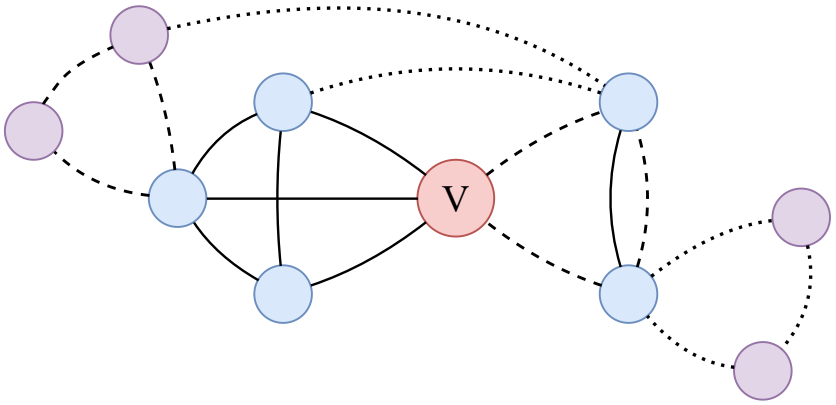


Figure 5.3: Example of the ego network growth rate in a multilayer network.

This measure can be seen as the size of the ego network at distance d normalized to the number of neighbors. As we only need to compute these two values, this measure is perhaps the simplest possible measure of social bridging. This makes it relatively cheap to compute, making it feasible to compute on networks with several hundred millions edges and millions of nodes.

The ENGR relates to social bridging as it considers the rate at which the ego network grows when we look d steps away from the ego. When d is small, high values indicate that the node has a bridging position in the network that allows it to reach a large number of nodes in a small number of steps. Low values indicate that few nodes can be reached outside of the direct ego network in a small number of steps. In this way, the ENGR is an indication of how well the ego network of a person allows them to obtain novel information, which other network measures of social bridging also measure.

Conceptually, the ENGR also relates directly to betweenness centrality. That is, when nodes form bridges between densely connected regions of a network, they will have a large ENGR since they can reach people in all the regions which they connect. Nodes that are central within a more dense regions will also have a high ENGR as they can reach all other nodes in this region in a few steps. In both cases, the betweenness centrality will also be higher. Nodes that are less central will have a low ENGR as their network position does not allow their ego network to grow as fast.

We further examine the relation of ENGR to node centrality in Figure 5.4, which shows correlations between two centrality measures and ENGR for various values of d using random networks. Different values

for d and the number of nodes n in the network are used, and the correlation coefficients are averaged over 25 random graphs for each network model and parameter set. Random network models that are used are the Erdös-Rényi model [ER+60], the Watts–Strogatz model [WS98], and Barabási–Albert model [AB02]. Centrality measures which we use are betweenness centrality and closeness centrality [Bav50]. The random network models that are used here have different properties than our data, and the latter is much larger. However, this figure does show that we can say at least with some certainty that there is indeed a positive correlation between node centrality and ENGR for small values of d, as we see consistently that the correlations here become larger as the network size is increased.

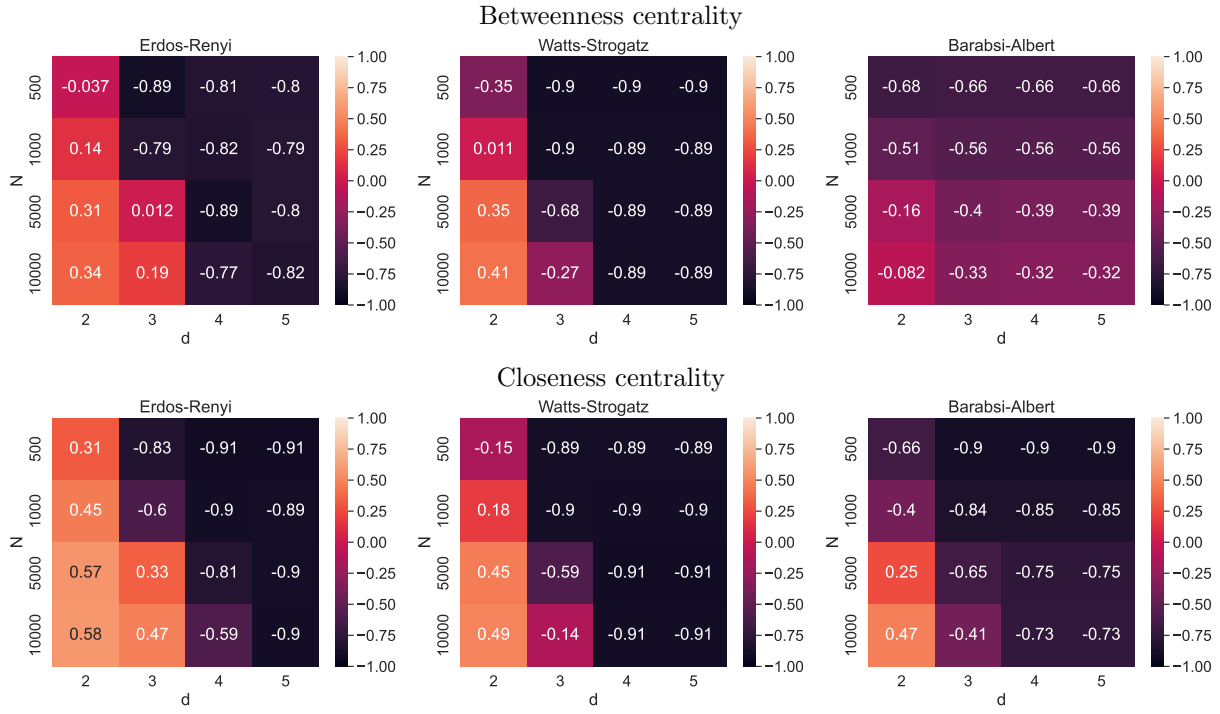


Figure 5.4: Pearson correlation coefficients between ENGR and node centrality.

To choose the value of d that we use in our analyses, we only consider values of $d = 2$ and $d = 3$, as larger values pose several problems. First, it is unlikely that people in the real world can indirectly access others that are more then a few steps away from them in the social network. This is made worse by some links that are of distance two in the real world being represented as distance one links in our network (see Section 4.1.2), because distances can appear larger than they are. Second, using a high value of d might be misleading, as when a node is not central, their ego network size at d will still be large, and if their degree is also low this can lead to a greatly inflated ENGR. Third, as d is increased, the ENGR becomes more computationally complex. Thus considering small values keeps the computation feasible.

Our choice of d is based on the correlations of the measure to the other measures, in order to ensure that we capture as little redundant information as possible. There will expectedly be a strong correlation between the ENGR and the node degree, since degree tends to relate to network centrality measures, since higher degree nodes tend to be more central. Moreover, if the degree of a node is high it is less likely in practice to have a closed ego network, and so the ENGR is expected to be higher.

Figure 5.5 shows the correlations between the aggregated network measures on our data, which confirms that the degree and ENGR are highly correlated. Correlations to excess closure are moderate. We observe that both the correlations are lower when d is smaller, and for that reason we use the ENGR with $d = 2$ in our analyses.

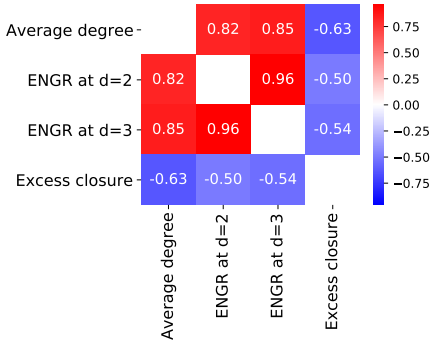


Figure 5.5: Pearson correlations between the network measures, using $d = 2$ and $d = 3$ for ENGR.

Despite the high correlation between the ENGR and the degree, it still makes sense to use the ENGR as a measure of social bridging. First, many network centrality measures correlate highly to each other, but they are still considered to be different as they have a different interpretation, and the correlation is not perfect. This is the case here as well. Second, whereas degree only considers direct resources, the ENGR at $d = 2$ focuses on resources that are one more step away. Such connections can be highly valuable and thus it is important to use a measure that considers them, especially as we use the measure to represent social bridging.

5.3 Statistical models

In order to examine what relations exist between network structure and social capital, we perform regression analyses. We take an approach similar to the one described in Chapter 2, where social capital is measured by looking at expected outcomes. We use four outcomes that are well known to relate to community social capital. An overview of the data which we selected for this can be found in Section 4.2. Relations between these outcomes and the network measures are estimated using linear models, created using ridge regression. We set these models up in such a way that that the coefficients of each variable can intuitively be interpreted, allowing us to conclude how the network measures impact the social capital outcomes.

5.3.1 Multicollinearity

Simply using ordinary least squares estimators would be problematic here as there is multicollinearity, meaning that some independent variables are highly correlated, as can be seen in Figure 5.5. Since the variables are not independent from one another, a change in one will cause a change in the correlated variables as well. This gives outlier data points more influence as they have a great impact on many variables at once. When ordinary least squares is used, coefficients are chosen to optimally fit the data, and so this large impact of individual data points can cause the model to overfit. Alternatively, we can say that the model has high variance as its performance will be highly different if the data is slightly perturbed or if small changes are made to the model itself. Since the model is unstable, we cannot fully trust the coefficients that are found for the correlated variables.

The solution to this is to introduce more bias into the model. Having more bias will cause the model to be less susceptible to small changes, which reduces variance and thus alleviates the problem of unreliable coefficients. Often this can also improve the overall quality of the model, because the reduction in variance is greater than the increase in bias.

5.3.2 Ridge regression

A common way to achieve this is by using ridge regression [HK70] to estimate the coefficients of a linear model. Like ordinary least squares, ridge regression gives an interpretable model that we can use to answer our questions. The difference between these methods is that an ridge regression uses an additional penalty term. Specifically L2 regularization is used to pull the coefficients towards zero. This keeps the model more simple and thus increases bias, which in turn decreases variance. Therefore we can use ridge regression in the face of multicollinearity to obtain more reliable estimates of the coefficients.

In ridge regression there is one parameter to tune: λ, which determines the strength of the penalty term. The higher λ, the greater the regularization and the simpler our model. We use cross-validation to tune this parameter, where we additionally apply the *One Standard Error Rule* [BFOS17, HTFF09]. This is a rule of thumb that is used to select the simplest yet most accurate model, and is further explained in Section 6.1.

Moreover, before performing ridge regression we must standardize the data by subtracting the mean and dividing by the standard deviation, so that the data has a mean of zero and a standard deviation of one. This is needed as the range of the variables impacts the size of the coefficients that are obtained, and thus this might cause too much or too little regularization to be applied.

5.3.3 Interpretation of the coefficients

Coefficients found with ridge regression can be interpreted in the same way as for standard linear models. By preprocessing the data, we obtain coefficients that are intuitive to interpret and compare to each other. For that we apply log-transformations to the independent and the dependent variables. After that we can interpret a coefficient β as the percentage change. That is, if the corresponding independent variable is increased by percent, then in our model the dependent variable will increase by β percent. Furthermore, log-transformations make linear regression models better at handling skewed data.

5.3.4 Control variables

Control variables are used to take factors other than network structure into account and ensure that we can analyze the effects of the network measures in our regression analyses. We selected a broad range of control variables that would be expected to have an effect on the social capital outcomes. Since the data has to be available on the community level, we used information from the available node attributes to compute the control variables. The list of potential control variables, along with a brief explanation on how they might effect social capital, is as follows:

- *Size*: This measures the number of people living in the neighborhood to which individual results are aggregated. Larger neighborhoods might be less close-knit and thus have less social capital.
- *Average age*: Younger neighborhoods might experience less social capital as the inhabitants are yet to build it up.
- *Urbanization*: The urbanization level of the neighborhood on a 5-point scale, where higher means more urban. More urbanized areas might experience less social capital as people are less dependent on each other.
- *Median household income*: More affluent areas likely have more social capital, since social capital is linked to financial success.
- *Percentage native Dutch*: The percentage of people that are ethnically native Dutch. This measure says something about the segregation of the neighborhood, and we expect more segregated communities to have less social capital.
- *Percentage highly educated adults*: As education level is related to socioeconomic status, having a higher portion of highly educated adults could affect social capital.

- *Percentage of people with a partner*: When more people have registered partners, this might indicate stronger family values within the neighborhood, in turn improving social capital.

To keep the regression model simple and interpretable, we eliminate control variables that highly correlate to one another. Figure 5.6 shows a correlation matrix of these variables.

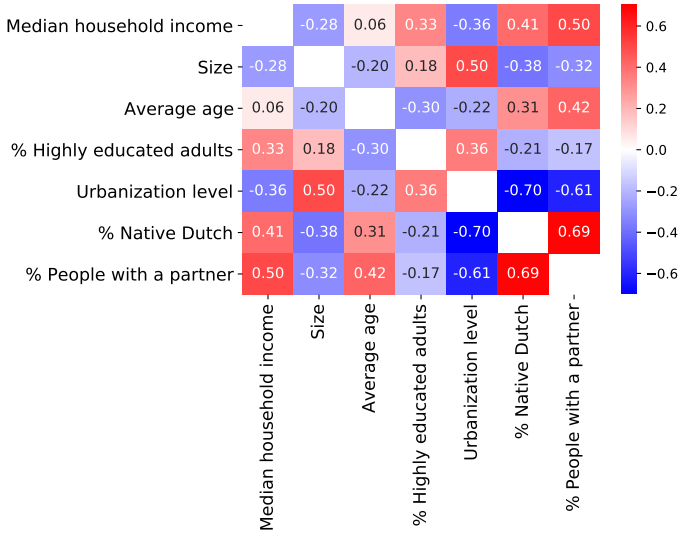


Figure 5.6: Pearson correlation coefficients between control variables.

Most variables are only moderately correlated, meaning that they capture different aspects of society. However, there are a few strong correlations between the Percentage native Dutch, Urbanization, and Percentage of people with a partner variables. We also see that these have nearly identical correlation patterns to the other variables. Therefore it seems that these controls capture more or less the same information. We choose to keep the Urbanization level, as we expect this to be the driving force that influences the other two variables, and not the other way around. This is because cities attract more migrants and people with families might prefer to live away from cities.

Thus we make use of five control variables. Table 5.1 shows the outputs of the regression models where only the control variables are considered for modeling the five outcome variables in Section 4.2. For most social capital outcomes we can already achieve good performance without using network measures, explaining up to 70% of variance for the SAB outcome.

	Dependent variable:				
	crime	SAB	health	volunteer	mortality
	(1)	(2)	(3)	(4)	(5)
med_income	-1.11	-2.02	0.18	0.34	-0.41
size	-0.06	0.10	-0.01	0.00	0.15
urbanization	0.40	0.24	-0.03	-0.21	0.03
average_age	-0.43	-0.05	-0.18	-0.10	2.69
highly_educated	-0.10	-0.02	0.03	0.01	-0.10
Constant	14.85	20.96	3.17	0.46	-4.76
R^2	0.23	0.70	0.58	0.57	0.38
λ	0.29	0.17	0.02	0.03	0.16

Table 5.1: Ridge regression results when using all control variables at once.

Chapter 6

Experiments

In this chapter we describe and discuss the experiments which we have performed. First we explain the experimental setup such that our methods can be reproduced. We then present and interpret our findings, after we finish with some discussion on the limitations of our research.

6.1 Experimental setup

In order to compute network measures on our data, we use the `popnet_mln` library[1]. This is a library developed by the POPNET team specifically for this purpose, though it can work with any multilayer network. It is written in Python[VRDJ95] and makes extensive use of pandas [pdt20], NumPy [HMvdW$^+$20] and SciPy [VGO$^+$20], all of which were also used in this work to extend the library and work with the data gathered through `popnet_mln`.

We store the individual-level network measures in a SQLite[2] database. This way we can easily process the data and export it to a CSV file so that we can import it for the regression analysis.

Here we make use of the R programming language [R C22]. We use the ridge regression implementation of the `glmnet` package[3]. This package implements linear models that use Elastic-Net regularization [ZH05]. Here a combination of LASSO [Tib96] (which uses L1 regularization) and ridge regression can be made. Using the α parameter, we can control the amount of each type of regularization that is used. We set α to 0 such that only ridge regression is used. Furthermore, using `glmnet` we can easily find optimal λ values using cross-validation and the One Standard Error Rule.

We specifically use 10-fold cross-validation, as is recommended in [HTFF09]. In order to apply the One Standard Error Rule, we evaluate the predictive performance, measured as mean squared error (MSE), and the standard error over the different folds for various values of λ. Based on this, we determine which setting of λ has the lowest MSE. The λ which we ultimately choose is the highest λ where the MSE is within one standard error of the MSE of the best model. This allows us to find a model with performance highly similar to the best one but which is (possibly) much simpler, which is especially important to avoid problems caused by multicollinearity.

Our analyses are set up such that we can determine the relations between the network measures and the social capital outcomes. For each social capital outcome (the dependent variable in the regression models), we run ridge regression using various configurations of independent variables. The names of the models, the independent variables that are used and what we use the model for are:

- **(1)**: ENGR and average degree.
  This provides insight into the isolated role of the ENGR.

[1] `https://github.com/bokae/popnet_mln`
[2] `https://www.sqlite.org/index.html`
[3] `https://cran.r-project.org/web/packages/glmnet/glmnet.pdf`

- **(2)**: Excess closure and average degree.
  This provides insight into the isolated role of the excess closure.
- **(3)**: All network measures: excess closure, ENGR and average degree.
  With this we can observe the combined performance of the network measures without control variables.
- **(4) through (7)**: All network measures, adding control variables in steps.
  Here we observe how the coefficients of the network measures change as we control for more variables.
- **final**: All independent variables.
  Using this model we draw determine the relations between network measures and the social capital outcomes while controlling for all control variables.
- **controls**: All the control variables.
  This model allows us to compare how well the models perform when network measures are added.

6.2 Results

The results from our experiments consist of the presentation of the regression results for each social capital outcome along with an interpretation of the coefficients. We then plot the data to get additional insights into the relations between the network measures and the social capital outcomes. An overview of the outcome variables which are examined can be found in Section 4.2. The network measures are described in Section 5.2, and a description of the control variables can be found in Section 5.3.4.

6.2.1 Violent crime rates

Table 6.1 presents the results of the analysis for the Crime outcome, where we look at the number of reported violent crimes in neighborhoods per one thousand people.

Overall we observe that the performance of this model is satisfactory, with the final model explaining 30% of variance in the data. Up to nearly half of this can be achieved using only network measures. In the final model, we find that the most influential control variables are the average age and the median income. When these are independently raised by one percent, we respectively expect a drop of 1.11% and 0.86% in the crime rate per one thousand people. That is, older and more wealthy neighborhoods see lower crime rates, which is to be expected.

We also find the network measures to have a strong impact. In comparison to the model which uses only the control variables, we find that network measures help explain an additional 7 percentage points of variance. The importance of network measures is also evident in the coefficients. If the average degree is increased by one percent, this leads to a drop of 1.31% in crime rates per one thousand people. The ENGR and excess closure also have strong effects. We see that when excess closure is increased, this lowers the crime rate considerably, as this is in fact the most important variable in the final model. Surprisingly, the ENGR has a positive coefficient, suggesting that improved social bridging relates to higher crime rates. It is also interesting how the directionality of excess closure flips when the neighborhood size is taken into account.

6.2.2 Social Assistance Benefits

We present the results for the regression analysis for SAB outcome in Table 6.2. With this social capital outcome we consider the percentage of people in a neighborhood that receive social assistance benefits from the Dutch government.

	Dependent variable:								
	crime								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	final	controls
avg_degree	-1.76	-0.84	-1.36	-0.90	-1.22	-1.17	-1.21	-1.31	
engr	1.91		1.26	0.52	1.03	1.05	0.73	0.64	
excess_closure		0.69	0.57	0.40	-0.04	-1.42	-1.10	-1.47	
med_income				-1.11	-1.27	-1.11	-1.05	-0.86	-1.13
size					0.01	-0.07	-0.07	-0.06	-0.06
urbanization						0.41	0.35	0.39	0.41
average_age							-0.93	-1.11	-0.44
highly_educated								-0.18	-0.10
Constant	1.93	3.96	2.19	14.22	15.56	14.67	18.52	18.42	15.10
R^2	0.14	0.09	0.12	0.17	0.19	0.28	0.29	0.30	0.23
λ	0.16	0.58	0.28	0.46	0.24	0.18	0.22	0.20	0.26

Table 6.1: Ridge regression results for social capital outcome Crime.

	Dependent variable:								
	SAB								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	final	controls
avg_degree	-0.60	-0.41	-0.38	-0.10	-0.20	-0.10	-0.13	-0.15	
engr	-0.30		-0.15	-0.45	-0.38	-0.48	-0.58	-0.62	
excess_closure		2.31	2.19	2.50	1.67	1.10	1.14	1.08	
med_income				-2.54	-2.04	-1.97	-2.00	-1.94	-2.06
size					0.13	0.10	0.10	0.10	0.10
urbanization						0.21	0.19	0.20	0.25
average_age							-0.39	-0.43	-0.04
highly_educated								-0.05	-0.02
Constant	3.81	1.32	1.69	27.48	21.74	21.39	23.55	23.46	21.32
R^2	0.09	0.13	0.12	0.59	0.68	0.71	0.72	0.72	0.70
λ	0.48	0.59	0.65	0.16	0.21	0.19	0.17	0.17	0.16

Table 6.2: Ridge regression results for social capital outcome SAB.

Overall the quality here is high, our final model being able to explain 72% of variance. A large part of this is due to the median income variable. This is to be expected, because there is clearly a direct relation between this and the outcome variable. Nevertheless, we also find that network structure has a strong influence. This time it is excess closure where we find unexpected results. As excess closure is increased by one percent, this leads to a 1.08% increase in the number of people that receive social assistance benefits. We would have expected social bonding to decrease SAB levels, since people with social bonding can get more assistance from their social contacts, decreasing the need for government assistance. The ENGR does behave as expected, it having a large and negative coefficient. We also observe that the models with network measures allow us to explain 2 percentage points more variance.

6.2.3 Perceived health and mortality rates

The results for the regression analysis for the Health outcome can be found in Table 6.2. Here we considered the percentage of people that perceive their health as being good or very good, as determined by the RIVM model which is built using the Gezondheidsmonitor (see Section 4.2).

	Dependent variable:								
	health								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	final	controls
avg_degree	0.09	0.08	0.06	0.05	0.05	0.05	0.03	0.04	
engr	0.10		0.06	0.09	0.08	0.09	0.06	0.08	
excess_closure		-0.43	-0.36	-0.43	-0.37	-0.40	-0.39	-0.32	
med_income				0.21	0.18	0.20	0.20	0.15	0.18
size					0.00	0.00	0.00	-0.01	-0.01
urbanization						-0.01	-0.01	-0.01	-0.03
average_age							-0.13	-0.09	-0.18
highly_educated								0.04	0.03
Constant	3.78	4.26	4.10	2.00	2.33	2.11	2.71	2.86	3.17
R^2	0.21	0.32	0.30	0.60	0.58	0.60	0.64	0.67	0.58
λ	0.07	0.05	0.07	0.02	0.03	0.02	0.02	0.03	0.02

Table 6.3: Ridge regression results for social capital outcome Health.

Again we observe that high R^2 scores are obtained, although now we also see that the network measures by themselves can explain a large percentage of variance. We also note that here coefficients are relatively low compared to the other models. Still, we find that the coefficients of the network measures are fairly high in comparison to most control variables, with excess closure being the most influential variable overall. However, we do observe again that excess closure negatively effects the social capital outcome. This effect is also consistent in models with control variables. Again we also observe that the final model is improved by adding network measures, with the explained variance increasing by 9 percentage points. These findings suggest that network structure has a strong relation to health outcomes.

As the Health data is made using a model and not register data (see Section 4.2), we additionally consider the Mortality outcome as a robustness check, to see if network measures play an important role here as well. The Mortality data comes from registers and results here thus are not dependent on any external models. The results of this analysis are shown in Table 6.4.

	Dependent variable:								
	mortality								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	final	controls
avg_degree	0.00	-0.08	0.00	-0.14	-0.20	-0.30	-0.05	-0.07	
engr	-0.01		-0.01	-0.16	-0.18	-0.11	0.64	0.61	
excess_closure		0.37	0.02	0.65	0.81	1.03	0.87	0.77	
med_income				-0.23	-0.28	-0.36	-0.48	-0.40	-0.40
size					0.06	0.10	0.14	0.15	0.13
urbanization						-0.04	0.00	0.01	0.04
average_age							2.79	2.78	2.45
highly_educated								-0.08	-0.10
Constant	2.11	2.21	2.11	5.10	5.35	5.93	-6.62	-7.14	-3.84
R^2	0.00	0.02	0.00	0.05	0.13	0.16	0.38	0.39	0.37
λ	136.65	4.82	137.30	1.98	0.89	0.35	0.16	0.15	0.23

Table 6.4: Ridge regression results for social capital outcome Mortality.

For this social capital outcome we find quite different patterns than for the Health outcome. Here the average age in the neighborhood clearly has the largest impact on the mortality rate, as is expected given the nature of the outcome. We observe that the ENGR by itself has no value, and only once the average age is considered does it become important. Excess closure also has a high coefficient in the final model. This time we surprisingly find that both network measures negatively impact the outcome. In spite of these effects being different from what we saw for the Health outcome, we still observe that network structure overall has a strong effect for this health-related outcome. This shows that the exact relation between health and network structure is perhaps difficult to predict, but that the relation exists nonetheless.

6.2.4 Volunteer

Lastly we look at the results for the Volunteer outcome, where we consider the percentage of people which do some form of volunteering work. Like the Health data, this is comes from a model by the RIVM that was made using the Gezondheidsmonitor data. In this case we do not have additional data for a robustness check, but we nevertheless believe our results to be accurate. We present the results for the Volunteer outcome in Table 6.5.

	Dependent variable:								
	volunteer								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	final	controls
avg_degree	0.49	0.37	0.30	0.26	0.26	0.23	0.27	0.30	
engr	0.40		0.33	0.41	0.40	0.49	0.55	0.64	
excess_closure		-1.39	-1.26	-1.17	-1.00	-0.65	-0.70	-0.56	
med_income				0.40	0.32	0.29	0.27	0.21	0.34
size					-0.03	-0.01	-0.01	-0.01	0.00
urbanization						-0.15	-0.13	-0.15	-0.21
average_age							0.32	0.40	-0.10
highly_educated								0.06	0.01
Constant	0.63	2.80	2.04	-2.17	-1.14	-1.19	-2.57	-2.82	0.46
R^2	0.34	0.39	0.38	0.49	0.54	0.71	0.73	0.74	0.57
λ	0.06	0.10	0.12	0.10	0.11	0.05	0.05	0.03	0.03

Table 6.5: Ridge regression results for social capital outcome Volunteer.

We see here that both network measures by themselves, along with the average degree, are able to explain over a third of all variance in the data. The quality of the final model is also excellent at an R^2 of 0.74. Compared to the model with just the control variables, we see a great improvement in the overall quality. Specifically, an additional 17 percentage points of variance in the data can be explained by adding network measures. The importance of network measures is further seen in their coefficients. This is quite remarkable as none of these variables were present in the model which was used to create the data. Again the excess closure has an unexpected directionality, relating to decreased levels of people doing volunteering work. The ENGR does behave as expected, and it has the largest overall coefficient in the model.

6.2.5 Data visualization

To help further understand these results, we plot the relations between the network measures and social capital outcomes using scatter plots. This gives us additional insights into how the network measures relate to the outcomes.

Figure 6.1 shows these scatter plots, with on the horizontal axis the social capital outcomes and on the vertical axis the network measures. Overall this figure shows similar relations to those that were found in the regression analyses. There are clear correlations between the network measures and social capital outcomes, and the directionality of the relations is also conform with the regression results. However, this figure does show us that the relations are not always linear. For example the relation between the ENGR and Health appears to have an exponential nature. We also find that there are outliers in the outcome data, especially for Crime and Mortality. These outliers might be problematic for the regression models, although this effect should be minimized by using ridge regression to estimate the coefficients.

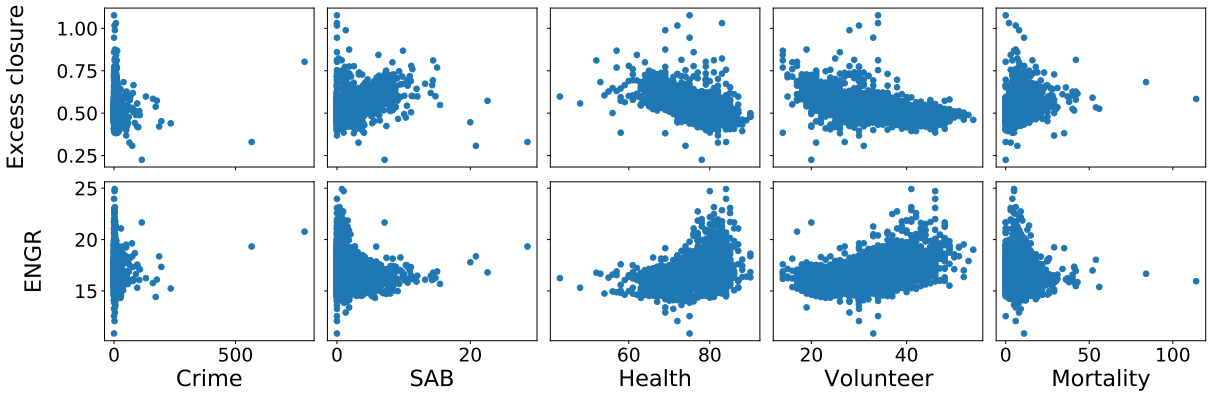


Figure 6.1: Visualization of the relations between network measures and social capital outcomes.

6.3 Discussion

Overall we find that network measures consistently improve the accuracy of the models. Network measures additionally often have higher coefficients than most if not all control variables. These findings suggest that network structure is strongly related to social capital.

Table 6.6 shows a brief summary of the results from the final models of our analyses. The plus signs indicate that there was a positive relation between the network measure and the outcome variable, whereas the minus sign indicates a negative relation.

Outcome	Excess closure	ENGR
Crime	-	+
SAB	+	-
Health	-	+
Mortality	+	+
Volunteer	-	+

Table 6.6: Summary of the relations between network measures and the social capital outcomes.

In most cases we found excess closure to be the most influential network measure, and often it was among the most important predictors overall. However, in nearly all cases we find that excess closure has a relationship in the opposite direction from what we expect, since increased excess closure generally was found to relate to the worsening of the social capital outcome. As excess closure considers only ties of multiple different types, it appears that when the social circles of a person greatly overlap, that this is bad for social capital. One likely explanation is that this overlap indicates a lack of diverse resources, which can trap people in low-opportunity social bubbles. It thus seems that having a more varied social network is important in terms of access to social capital. In this way, social bonding can have a negative influence. Yet, the negative relation to crime does suggest that social bonding can have a positive effect as well.

The results for the ENGR were in line with our expectations, as increased values here related to substantial improvements in most of the social capital outcomes. This suggests that increased social bridging relates positively to social capital.

6.4 Limitations

One of the limitations of this work is that two of our social capital outcomes are based on data created by a model. It would be preferred to use the raw data that was used to train the model, since this should still give accurate estimations for a large number of neighborhoods. Alternatively, register-based data would be ideal, but this is not available for all outcomes.

Our results are also limited by the choice of model. More complex models for example, might help to improve the estimation of the relations between the network measures and social capital outcomes. However, this would come at a cost of interpretability. On the other hand simpler models can have advantages as well. For example when using ordinary least squares to train a linear model we would be able to estimate p-values of each coefficient. This would allow us to ascertain if the effects of the network measures were statistically significant. However due to the multicollinearity problem we opted to rather use ridge regression and focus on the coefficients themselves to draw conclusions. Another reason for this is that the use of p-values in this context is somewhat controversial. It can be argued that in population data p-values are meaningless as coefficients are not estimates but precise values, although it is also argued that the population is part of a greater superpopulation. By focusing on the coefficients we avoid this discussion and we can still draw meaningful conclusions. However, not being able to speak of significance is a limiting factor.

Further limitations include the small selection of network measures. A virtually endless number of network measures can be created that might capture some aspect of social capital. We only focused on two for which there is a clear theoretical reasoning as to what they say about network structure and how this relates to social capital. However it is certainly possible that better choices exist. Additionally, a broader range of network measures could be considered, either to have several measures for one network structural aspect, or to cover a wider range of network structural aspects altogether.

Of course our findings are highly dependent on the network that we used. As discussed in Section 4.1.1, the network only contains formal ties, missing some highly valuable connections between people. And as we did not select large classrooms and workplaces, this also makes us miss out on a fair number of important formal ties. More experimentation using different selections of links could lead to better results. Moreover, our data in its current state is just a snapshot of the Dutch population in 2018. Having snapshots of different years could be valuable to test the robustness of these results.

Lastly we have limited the scope of this work to community social capital, but similar analyses using our data can also be performed on an individual level. Here too, it would be highly valuable to have longitudinal data. Another approach could be to link individuals to their outcomes from large-scale surveys such as the Gezondheidsmonitor.

Chapter 7

Conclusion

In this thesis we have made initial steps towards the validated measurement of social capital using network data. For this we have indirectly analyzed the relation between network structure and social capital. We achieved this by, for the first time, bringing traditional and network-based measurement techniques of social capital together. This is achieved by analyzing the connection between two network measures and a wide range of well-studied social capital outcomes on the geographical neighborhood level.

Ridge regression was used to estimate coefficients of linear models, using the network measures along with several control variables as independent variables, and the social capitals as dependent variables. The network measures were chosen to capture various aspects of network structure. Excess closure measures the level of social bonding in a community, whereas the ego network growth rate (ENGR) relates to social bridging. With this analyses we were able to draw conclusions on the relations between network structure and social capital.

Our contribution overcomes the following three problems with existing literature that measures social capital in networks: a) often no empirical evidence is provided that the used network measures actually relate to social capital, b) works that do not have this problem measure highly specific types of social capital, and c) small sample sizes can lead to weaker and less broad conclusions. By finding the relation to network measures and a broad range of social capital outcomes, we overcome the first two issues, and the third is overcome by analyzing a population-scale social network. Additionally, we have proposed a new network measure that can efficiently be computed at a large scale, and we have shown it to relate to network centrality.

Overall we find that network structure has a strong relation to social capital. The network measures are consistently among the most important predictors, and the regression models are always improved by adding the network measures, in some cases by a great amount. Excess closure often related to decreased levels of social capital. This suggests that social bonding can constrain people in their access to opportunities. However, we also find that it can be beneficial to a community because it relates to lower crime rates. The results for the ENGR were in line with our expectations, as increased values related to substantial improvements in most of the social capital outcomes. This suggests that increased social bridging relates positively to social capital and comes with opportunities for people.

Our research question has been: *What is the relationship between network structure and social capital?* Altogether our findings show a strong relation between network measures and social capital outcomes, and thus between network structure and social capital. We find that social bonding can constrain people to low-opportunity social networks, although it can also be beneficial to a community. Social bridging has a positive relation to social capital, relating to improvements for nearly all social capital outcomes.

Some ways to extend the work presented here are to make use of more social capital outcome data and consider a wider array of network measures. This can improve the robustness of the results or reveal new insights altogether. Specifically the usage of the raw data from the Gezondheidsmonitor has the potential

to reveal additional and more robust insights on the connection between network structure and health. Many other extensions are also possible, such as the addition of more control variables to improve the robustness. Further analyses on the behavior of excess closure and why its effect is negative could also be valuable. Another possibility is to dive deeper into the results and for example see what neighborhoods are problematic to predict for our models.

Future work could also focus on individual-level social capital. There, survey data can be used to relate network measures to the outcomes found in the surveys. The income of people can also be used as a proxy for their social capital. Lastly, this work could be extended to form the first network-driven official statistic. This statistic could capture network structure with either a single or a combination of network measures to say something about, for example, the total social resources available to people in a neighborhood.

Acknowledgments

We would like to sincerely thank everyone that helped in the process of making this thesis. This research would not have been possible without the help of the supervisors dr. Frank Takes and dr. Eelke Heemskerk, who have provided invaluable guidance and advice throughout the entire process. Our thanks also go to the POPNET team, Eszter Boyányi, Yuliia Kazmina and Rachel de Jong, for the many great discussions, feedback and ideas, along with creating the library upon which we build to perform our analyses. We also thank Platform Digitale Infrastructuur for funding POPNET, and Statistics Netherlands for the internship that this thesis was part of and granting us access to the data. We are also grateful to the researchers at Statistics Netherlands, Gert Buiten, Mark van der Loo, Marjolijn Das, Edwin de Jonge, Frank Pijpers, Jan van der Laan and Maarten van Rossum for providing valuable feedback on the work presented in this thesis. Lastly we would like to thank all our friends and family for the love and support.

Bibliography

[AB02] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47, 2002.

[AK02] Paul S Adler and Seok-Woo Kwon. Social capital: Prospects for a new concept. *Academy of Management Review*, 27(1):17–40, 2002.

[AWH14] Alireza Abbasi, Rolf T Wigand, and Liaquat Hossain. Measuring social capital through network analysis and its influence on individual performance. *Library & Information Science Research*, 36(1):66–73, 2014.

[BAGADF15] María Bordons, Javier Aparicio, Borja González-Albo, and Adrián A Díaz-Faes. The relationship between the research performance of scientists and their position in co-authorship networks in three fields. *Journal of Informetrics*, 9(1):135–144, 2015.

[Bav50] Alex Bavelas. Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, 22(6):725–730, 1950.

[BBH+21] Eszter Bokányi, Gert Buiten, Eelke Heemskerk, Edwin de Jonge, Yuliia Kazmina, Jan van der Laan, and Frank Takes. Popnet data quality exploration report. Internal technical report, 2021.

[BFOS17] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 2017.

[BJE98] Stephen P Borgatti, Candace Jones, and Martin G Everett. Network measures of social capital. *Connections*, 21(2):27–36, 1998.

[Bok] Eszter Bokányi. The anatomy of a population-scale social network.

[Bor98] Stephen P Borgatti. A socnet discussion on the origins of the term social capital. *Connections*, 21(2):37–46, 1998.

[Bou18] Pierre Bourdieu. *The forms of capital*. Routledge, 2018.

[Bur92] Ronald S Burt. *Structural holes*. Harvard University Press, 1992.

[Col88] James S Coleman. Social capital in the creation of human capital. *American Journal of Sociology*, 94:S95–S120, 1988.

[DJ] Rachel De Jong. Algorithms for efficiently computing exact structural anonymity in complex networks.

[Dur05] Emile Durkheim. *Suicide: A study in sociology*. Routledge, 2005.

[ER+60] Paul Erdős, Alfréd Rényi, et al. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5(1):17–60, 1960.

[EVGL14] Nicole B Ellison, Jessica Vitak, Rebecca Gray, and Cliff Lampe. Cultivating social resources on social network sites: Facebook relationship maintenance behaviors and their role in social capital processes. *Journal of Computer-Mediated Communication*, 19(4):855–870, 2014.

[Fre77] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[Fuk96] Francis Fukuyama. *Trust: Human nature and the reconstitution of social order*. Simon and Schuster, 1996.

[HA13] Matthias Hofer and Viviane Aubert. Perceived bridging and bonding social capital on twitter: Differentiating between followers and followees. *Computers in Human Behavior*, 29(6):2134–2142, 2013.

[HK70] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[HMB95] John Hagan, Hans Merkens, and Klaus Boehnke. Delinquency and disdain: Social capital and the control of right-wing extremism among east and west berlin youth. *American Journal of Sociology*, 100(4):1028–1052, 1995.

[HMvdW+20] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

[HTFF09] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[KKL97] Ichiro Kawachi, Bruce P Kennedy, and Kimberly Lochner. Long live community: social capital as public health. *American Prospect*, pages 56–59, 1997.

[Lin17] Nan Lin. Building a network theory of social capital. *Social Capital*, pages 3–28, 2017.

[LKK99] Kimberly Lochner, Ichiro Kawachi, and Bruce P Kennedy. Social capital: a guide to its measurement. *Health & Place*, 5(4):259–270, 1999.

[LLY13] Eldon Y Li, Chien Hsiang Liao, and Hsiuju Rebecca Yen. Co-authorship networks and research impact: A social capital perspective. *Research Policy*, 42(9):1515–1530, 2013.

[LNP01] Robert Leonardi, Raffaella Y Nanetti, and Robert D Putnam. *Making democracy work: Civic traditions in modern Italy*. Princeton University Press Princeton, NJ, 2001.

[LST17] Karl V Lins, Henri Servaes, and Ane Tamayo. Social capital, trust, and firm performance: The value of corporate social responsibility during the financial crisis. *The Journal of Finance*, 72(4):1785–1824, 2017.

[P+00] Robert D Putnam et al. *Bowling alone: The collapse and revival of American community*. Simon and schuster, 2000.

[pdt20] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

[Poo06] Wouter Poortinga. Social relations or social capital? individual and community health effects of bonding social capital. *Social Science & Medicine*, 63(1):255–270, 2006.

[Put01] Robert D. Putnam. Social capital: Measurement and consequences. *Isuma: Canadian journal of Policy Research*, 2(Spring 2001):41–51, 2001.

[R C22] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[SKK02] Subu V Subramanian, Daniel J Kim, and Ichiro Kawachi. Social trust and self-rated health in us communities: a multilevel analysis. *Journal of Urban Health*, 79(1):S21–S34, 2002.

[SL08] Joonmo Son and Nan Lin. Social capital and civic action: A network-based approach. *Social Science Research*, 37(1):330–349, 2008.

[SS13] Katherine Scrivens and Conal Smith. Four interpretations of social capital: An agenda for measurement. *Organisation for Economic Cooperation and Development (OECD)*, 2013.

[Sto01] Wendy Stone. Measuring social capital. *Australian Institute of Family Studies, Research Paper*, 24, 2001.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[vdKZB+17] Jan van de Kassteele, Laurens Zwakhals, Oscar Breugelmans, Caroline Ameling, and Carolien van den Brink. Estimating the prevalence of 26 health-related indicators at neighbourhood level in the netherlands using structured additive regression. *International Journal of Health Geographics*, 16(1):23, Jul 2017.

[VdL22] Jan Van der Laan. A person network of the netherlands. Technical report, Centraal Bureau voor de Statistiek, 2022.

[VGO+20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[VRDJ95] Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[War01] Mildred Warner. Building social capital:: The role of local government. *The Journal of Socio-Economics*, 30(2):187–192, 2001.

[WKK98] Richard G Wilkinson, Ichiro Kawachi, and Bruce P Kennedy. Mortality, the social environment, crime and violence. *Sociology of Health & Illness*, 20(5):578–597, 1998.

[WS98] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440–442, 1998.

[ZH05] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.