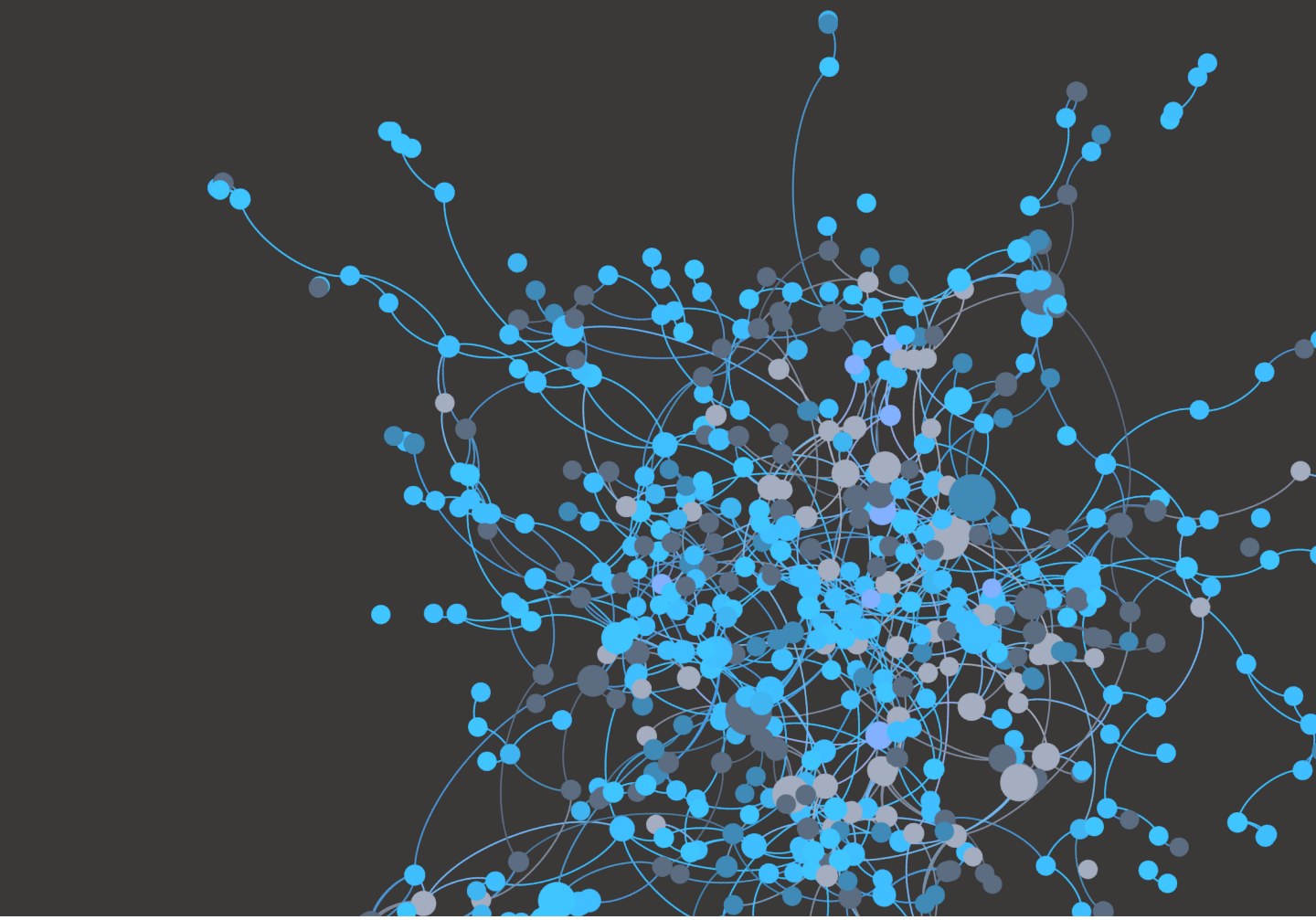
Joost Bouten

Influence Optimization in Interorganizational Healthcare Networks



MSC ECONOMICS THESIS

2018

Tilburg University

Nederlandse Zorgautoriteit

Joost Bouten

Influence Optimization in Interorganizational Healthcare Networks

Master Thesis

Tilburg School of Economics and Management

Supervision

prof. dr. M. C. Mikkers
prof. dr. J. Boone

Coordinator

dr. G. S. Verhoeven

August 23, 2018

Tilburg School of Economics and Management

Title of work:

Influence Optimization in Interorganizational Healthcare Networks

Thesis type and date:

Master Thesis, August 23, 2018

Supervision:

prof. dr. M. C. Mikkers
prof. dr. J. Boone

Student:

Name:	Joost Bouten
Student number:	1265889
ANR:	132978
E-mail:	j.bouten@tilburguniversity.edu
Semester:	Spring 2018

Abstract

Sometimes, regulators wish to obtain support for a particular new policy from various stakeholders, such as all healthcare providers in a particular sector. One way of obtaining such support is by organizing meetings with an ad-hoc selection of representatives. However, by exploiting the underlying network structure of these representatives, we might be able to improve on this selection. To do so, we assume word-of-mouth information spreading and use algorithms designed to maximize influence given a capacity constraint. We apply these techniques to a real network of board members of Dutch healthcare organizations. The network is formed through social links which arise due to joint membership of boards. Recently, algorithms were developed with the aim of selecting a set of persons that maximize influence, given a model for influence spreading. We model influence spreading from one person to another with a one-shot fixed probability. Here, we implement these algorithms in R and compare them based on their expected total influence and computation cost. We find that out of the three compared algorithms, the greedy algorithm has the greatest expected influence, although it is most costly in terms of computation. For our network, the degree discount heuristic approximates the results of the greedy algorithm while being computationally cheap.

Contents

1 Introduction

1.1 Research Question

Viral marketing techniques have long been used to effectively distribute information using the word-of-mouth principle. The concept of viral marketing refers to techniques that aim to spread trends and product popularity using word-of-mouth principles and given an existing social network. Similar to viral marketing agencies, regulators regularly try to target individuals such that information, e.g. regarding proposed policies, will be shared among as many of their stakeholders as possible. Targeting the set of individuals that maximizes the expected influence spread has been shown to be NP-hard. The concept of the NP-hardness, or non-deterministic polynomial-time hardness, is rooted in computational complexity theory and requires extensive explanation. Here, we provide a short example to describe the concept of NP-hard decision problems in order to achieve a basic understanding. For a more elaborate explanation, we refer to Wikipedia contributors (2018). An example of an NP-hard problem is the *subset sum problem*, which is a variation of the *knapsack problem*. The problem is as follows. Given a set of integers, is there a non-empty subset that sums to zero? For small subsets, this problem is easy to solve. For example, given the initial set $\{1, -2, 4, -5\}$, we can easily see that the answer to the subset sum problem is "yes", because the subset $\{1, 4, -5\}$ sums to zero. A computer could find the solution to this problem by calculating the sum of all subsets and checking whether any subset sums to zero. What makes this problem NP-hard is that if we have an initial set with many integers, searching for a subset which sums to zero becomes a lengthy process, as the number of subsets is exponentially related to the length of the initial set. Many problems that are NP-hard require exponential time or even longer, i.e. it becomes exponentially harder for an algorithm to solve it as the size of the input increases. Since the influence maximization problem is NP-hard, heuristics and algorithms have been proposed to distinguish influential individuals and approximate the optimal set of individuals to target. To use such algorithms to the benefit of regulators, such as the Dutch Healthcare Authority (NZa), we must make assumptions regarding the underlying social network and the progression of information flows throughout such a network.

To construct a social network of stakeholders in the Dutch healthcare markets, we study the inter-organizational network that arises upon accounting for the occurrence of interlocking directorates. Directorates are said to interlock if they have a shared member, that is, if there is at least one individual which is seated on a board of both of the firms. A large part of Dutch healthcare governance consists of a two-tier board structure, which is typified by a separation of boards. In

the two-tier model, inside and outside directors are legally separated by a supervisory board and a management board. Inside and outside directors are legally prohibited from simultaneously being an inside or outside director of a competitor (Brancheorganisaties Zorg (BoZ), 2017). Nevertheless, directors often hold positions on the boards of several healthcare providers, causing links between healthcare providers to arise. In this thesis, we ignore the distinction between inside and outside directors. Based on the interaction between board members we construct a network model for information flows. Among other uses, this network may be used for viral marketing techniques employed by the NZa. The aim of this thesis is to maximize the influence of the NZa among its stakeholders. Using social network analysis, we aim to answer the following main research question.

How can we target the most influential board members in healthcare markets?

In order to implement network analysis in this decision-making process, we assume a particular process of information diffusion through the relevant network. It should be noted that the NZa does not necessarily always have the intent to maximize influence. In certain cases, we might want to target the most influential individuals in order to extract information from a network, this distinction is further explained in Chapter 6.

1.2 Literature

This thesis builds on two primary sources of literature. First, the work of Heemskerk, Hendriks, Wats, et al. (2010) opens up the exploration of interlocking directorates in Dutch healthcare as of the introduction of the Health Insurance Act (HIA) in 2006. Second, Kempe, Kleinberg, and Tardos (2003) formulates the influence maximization problem for various influence diffusion models and provides a greedy algorithm which can be used to approximate the optimal solution. In this section, we will first address the work of Heemskerk et al. and Westra (2017) with regard to interlocking directorates. Secondly, we discuss the relevant literature regarding influence maximization in social networks.

1.2.1 Interlocking directorates

To our knowledge, network analysis of Dutch healthcare markets was first introduced when Heemskerk et al. (2010) questioned the impact of *indirect interlocks*[1] on market forces in healthcare. Heemskerk et al. visualize the network of indirect interlocks between Dutch hospitals and show that 95% of Dutch hospitals are connected to each other in this manner. However, they are unable to provide any insight into the extent to which competition among hospitals is affected by interlocking directorates.

Chapter 3 of Westra (2017) further explores inter-organizational relationships in healthcare using

[1] When board members of two organizations both hold a board position at a third organization.

network analysis techniques. Contrary to Heemskerk et al., Westra explores *direct interlocks*. Westra hypothesizes that the presence of numerous healthcare reforms caused the formation of board interlocks as a mechanism for coping with uncertainty following from these reforms. Firstly, the paper seeks to assess the prevalence of direct board interlocks. Secondly, it aims to determine how this prevalence has changed over time. Lastly, Westra studies the interlocks that exist between market entrants and incumbent organizations. Westra finds that interlocking directorates are most common within similar geographical regions. Furthermore, organizations seem to be primarily connected to organizations within similar sectors. In addition, a significant increase in the average number of interlocks between 2007 and 2012 suggests that there is a growing trend in the number of interlocks between organizations. Westra finds that half of all boards are connected by direct board interlocks. Westra advises policymakers and researchers to not only consider the board positions that an individual can hold, but should also consider the interlocks of an organization.

1.2.2 Influence maximization

The influence maximization problem was first algorithmically formulated by Domingos and Richardson (2001), who recognized that the cost-benefit tradeoff of marketing actions should depend on the network value of the targeted individual. Further developments towards solving the influence maximization problem have been introduced by Kempe et al. (2003), who first posed the problem as a discrete optimization problem as analyzed in this thesis. Kempe et al. outline models of a spread of influence in networks and discuss their uses in understanding the dynamics of adoption of information. One such model is the independent cascade model, which is used in this thesis and further elaborated upon in Section 2.3.1. They show that for some models, such as the independent cascade model, finding an algorithm to maximize influence is NP-hard. For models such as these, approximation algorithms are formulated. Kempe et al. compare their algorithms to heuristics based on major concepts from social network studies and find that the algorithm provides significant influence gains.

Chen, Wang, and Yang (2009) recognize that the algorithms proposed by Kempe et al. are insufficiently scalable due to their lengthy computation time. Therefore, Chen et al. seek to provide scalable solutions to influence maximization in social networks. The first direction that the authors take is to improve on the greedy algorithm that was developed by Kempe et al.. The second is to devise new degree discount heuristics that may be used to solve the influence maximization problem relatively well without the need for long running times of algorithms. Among other models, Chen et al. test their algorithm given an independent cascade model of influence diffusion. They find that their improved greedy algorithm establishes better running time and that their degree discount heuristic is able to achieve much greater influence spread than traditional degree and centrality-based heuristics. As part of this thesis, the algorithms that were developed by Chen et al.

have been implemented in R. The correct implementation of these algorithms has been tested using simulations based on the NetHEPT dataset to reproduce Figure 1 of Chen et al.. The underlying processes of the implemented algorithms are further elaborated upon in Section 2.

After the improved algorithms were posed by Chen et al., others have suggested alternative algorithms. Existing algorithms differ in terms of their quality of influence spread, running time efficiency, and main memory footprint. While some algorithms generally outperform others in all aspects, there does not seem to be a single best influence-maximization algorithm (Arora, Galhotra, & Ranu, 2017). Arora et al. compare different algorithms and argue that the choice is to be made between four algorithms. Firstly, IMM is the fastest of all compared algorithms when a weighted-cascade model of information diffusion is used (Tang, Shi, & Xiao, 2015). Secondly, TIM^+, is the fastest option in case a linear threshold model is used (Tang, Xiao, & Shi, 2014). Thirdly, if the independent cascade model is used, the fastest option is to use the PMC algorithm (Ohsaka, Akiba, Yoshida, & Kawarabayashi, 2014). Lastly, a slightly inferior option to IMM, TIM^+, and PCM in terms of efficiency and influence spread is the $EaSyIM$ algorithm (Galhotra, Arora, & Roy, 2016). However, the $EaSyIM$ algorithm is arguably the best performing option when working with low memory. While their benchmarking study considers many of the existing algorithms, Arora et al. do point out that the field of influence maximization is evolving. More recently, Nguyen, Thai, and Dinh (2016) developed *Stop-and-Stare* algorithms as scalable approximation algorithms for influence maximization. Nguyen et al. show that their algorithms, SSA and $D\text{-}SSA$, achieve influence spreads similar to those of the IMM and TIM^+ algorithms while running several orders of magnitude faster. While these recently-developed algorithms have been found to significantly outperform versions of the greedy algorithm in terms of memory and running time, no algorithm has been found to consistently outperform the greedy algorithm in terms of influence spread quality (Arora et al., 2017). For this reason, we expect implementation of these more recently-developed algorithms in the analysis of the board member network only to improve running times and memory usage.

2 Methods

In this chapter, we discuss the methods which are used in further chapters to answer our main research question. First, we clarify a number of terms related to network science. Second, we elaborate on the approach we take in order to construct the board member network. Third, we discuss the most prominent information diffusion models and substantiate the choice of the independent cascade model. Lastly, we describe in detail the approximation algorithms that we implemented in R as part of this thesis.

2.1 Terminology

Within the context of network science, there exists a great deal of terminology that should be clarified before we proceed to the analysis. This section clarifies a number of terms that are used in this thesis. As network science is closely related to graph theory, a number of terms are often interchangeably used in the scientific literature (Barabási & Pósfai, 2016). After some terms, the terminology from graph theory is included between parentheses. Figure 2.1 provides an illustration to support the explanation of some of the terminology using a randomly generated network.

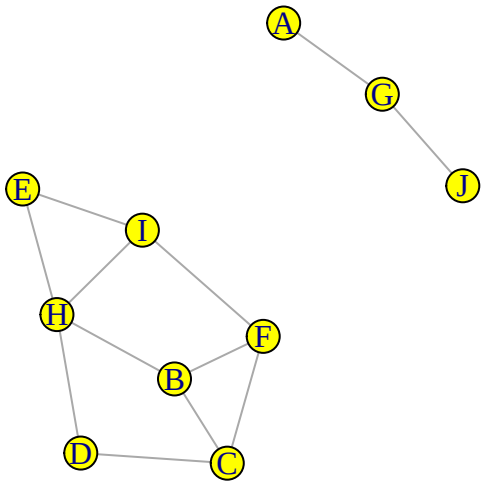


Figure 2.1: Example network

Networks (*graphs*) generally consist of two basic types of elements. On the one hand, *nodes* (*vertices*) represent the connection points of the network. In the network of board members, nodes are used to represent board members. In Figure 2.1, nodes are represented by the yellow circles, which are labeled using the letters A to J. On the other hand, *links* (*edges*) are used to portray a connection between a set of nodes. That is, the notation (A, G) may be used as the representation of a link between nodes A and G. Links can be of two types. Either a link is *directed*, in which case it

has a known source and target, or it is *undirected*. In the social network of board members, we assume no directionality, since we expect communication to be able to go both ways along a social relationship between two board members. Nodes can carry a number of attributes, one of which is its *degree*. The degree of a node represents the number of links it has to other nodes (Barabási & Pósfai, 2016). In Figure 2.1, node H has the highest degree, since it is the only node that is connected to at least four other nodes. The goal of this thesis is to target the individuals with the greatest expected influence spread. The people that are initially chosen to transmit the relevant information are called the *seeds*. Influence diffusion models are developed to model the information flows from one node to another. Stochastic influence models often assume a particular *infection probability* of information successfully flowing across a link and thereby *infecting* the receiving node with the provided information. Contrary to the negative connotation of the word, "infection" is desirable if we aim to maximize influence. Nodes can only infect other nodes that are in the same *component*. A *network component* is a disconnected part of a network, i.e. there is no path that reaches from a node in one component to a node in another component. The example network in Figure 2.1 consists of two components. The smallest component contains nodes A,G, and J. The other nodes are contained in the other component.

2.2 Network Structure Approach

Upon construction of the network, a decision has to be taken regarding the structure of the network. That is, the basis on which something or someone is regarded as a node and the conditions for the establishment of a link between two nodes have to be determined. The most complete network, the *bipartite network*, of interlocking directorates would include two types of nodes, namely firms and individuals. Each firm would be represented by the first type of node that is adjacent[1] to each of its board members, which serve as the second type of nodes. A representation of a bipartite network is shown in Figure 2.2. The bipartite network structure may be used to visualize the network of board members, as is demonstrated in Section 6.1.3. This structure enables the observer to easily observe board membership of firms.

[1] Two nodes are adjacent if there is a link between them.

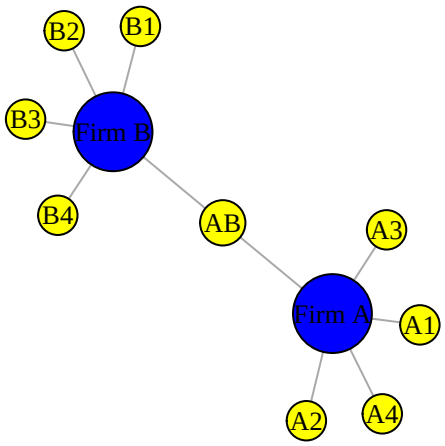


Figure 2.2: Bipartite network example

The example network of Figure 2.2 can be interpreted as follows. The two blue nodes represent firms A and B. Both firms have 5 board members, which are represented by the yellow nodes. One individual is a board member of both firms, namely person AB. Others have discarded the board members from the network of interlocking directorates in their final representation and used shared membership as a condition for a link existing between two organizations (Stokman, Van der Knoop, & Wasseur, 1988; Heemskerk et al., 2010; Westra, 2017). An example of this network structure is shown in Figure 2.3.

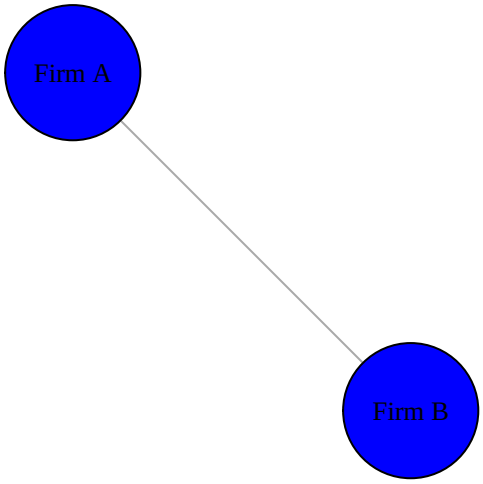


Figure 2.3: Firm network example

From the bipartite network displayed in Figure 2.2, the network in Figure 2.3 is constructed. Nodes "Firm A" and "Firm B" are connected as these firms had a common board member, namely the person "AB". The remaining network merely consists of firms. This representation is particularly useful when analyzing possible strategic relations between firms and the underlying social network is not of great importance for further interpretation. In this thesis, an alternative methodology is used. In our final network structure, firms are no longer in the network as nodes. Instead, we

assume that two board members are linked if they are positioned in a board of the same healthcare provider. A network of this structure is shown in Figure 2.4. This methodology is used since we aim to target influential individuals rather than firms. Furthermore, this approach aims to display social ties between board members as links between them. In the bipartite network structure, this feature is less apparent, as no two individuals are directly linked. For the purpose of network visualization, this approach is likely less desirable. As each board member of a firm is linked to any other board member of the same firm, the number of links in this approach is much greater than the number of links in a bipartite network. Therefore, the bipartite network approach arguably displays the board member network more clearly, while also providing information regarding the relevant firms.

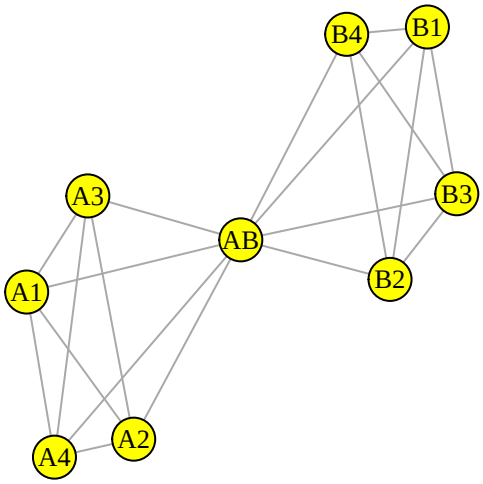


Figure 2.4: Member network example

2.3 Models for Influence Spreading

In this section, we discuss influence diffusion models for social networks that appear most prominently in the relevant literature. Many influence models exist, however, we will not discuss explanatory models, which include epidemic models such as the Susceptible-Infected models, as these models are not commonly used for influence maximization (Li, Wang, Gao, & Zhang, 2017). Instead, we will focus on predictive influence models, in particular, cascade and threshold models. Furthermore, we will not delve into the class of Game Theory cascade models (Camerer, 2011; Hang, Zhu, Song, & Zhang, 2014), as these models generally make use of cost-benefit analyses and strategies which are not apparent in our application. Out of the existing models, we have chosen to implement the independent cascade model as we find that this model is scalable and intuitive.

2.3.1 The Independent Cascade Model

The influence model we implement in this thesis is the independent cascade model with constant infection probability as it was first described by Kempe et al. (2003). One key strength of the independent cascade model is its simplicity. The influence process of the independent cascade model

is as follows. We start with an initial set of nodes, i.e. the seeds. At the first step t, each node v that is in the initial set has a one-shot chance of infecting each currently non-infected neighbor w. The infection succeeds with a given probability p that is independent of v and w, and thus the same for every node. In the following step, i.e. step $t+1$, all nodes that have been infected in t follow the same process. That is, each node infected in t is given a single chance to infect each currently non-infected neighbor. Again, the success is dependent on a random process that takes a given infection probability p. This process of infection continues until no more nodes can be infected. In the implementation we have constructed in R, we simulate the independent cascade process by simultaneously deleting links within the network at random. Each link in the network is deleted from the network with probability $1-p$. All nodes that are reachable from a seed node in the remaining network are infected.

2.3.2 The Weighted Cascade Model

The weighted cascade model of information diffusion is a variation of the independent cascade model in which infection is not determined by a constant infection probability. Instead, the probability of a node being infected by another node is negatively dependent on its degree. This means that a board member less likely to be infected by a given other board member if the board is of greater size. This feature seems to be plausible, more heavily-connected people may have less time to be influenced by each individual connection. On the other hand, a counteracting factor might be that more heavily connected people may be more social by their nature. We suggest for further research to address the relevant probability weights in board member networks. In this thesis, we do not adopt the weighted cascade model, as the independent cascade is simpler and it is unclear which of the discussed factors dominates in the board member network.

2.3.3 Threshold Models

In addition to the cascade model, threshold models were developed to model information diffusion in networks. One such model is the linear threshold model as described by Granovetter (1978). Contrary to the independent cascade model, infection in the linear threshold model linearly depends on the number of nodes that have attempted to infect it. If a certain threshold is surpassed, the node is infected. To account for heterogeneity in the likelihood of information adoption, each threshold may be randomly drawn from a frequency distribution of thresholds. Alternatively, thresholds may be determined according to a constant parameter value.

2.4 Influence Maximization Algorithms

2.4.1 Basic Heuristics

Although the greedy algorithm provides approximation guarantees to the influence maximization problem, one might wish to use heuristics, as running the greedy algorithm on large networks can be a lengthy process. A number of heuristics have been developed to approximate influence spread quality of the greedy algorithm while greatly improving on running times. In this thesis, we evaluate two heuristics after implementing them in R. The first heuristic is the *highest-degree heuristic*, which is arguably one of the easiest heuristics to employ. The highest-degree heuristic states that the nodes with the highest degree should be selected as our seeds. The implementation of this heuristic in my thesis orders all nodes by their degree and takes the top k nodes as seeds, where k is the number of seeds to choose. If the k'th node in the list has a degree that is equal to the degree of node at index $k+1$, the first $k-x-1$ nodes are definitely chosen into the set of seeds, where x is the number of nodes that have a degree equal to the degree of the k'th node but appear in the list before the k'th node does. The remaining $x+1$ seeds are followingly chosen randomly from the nodes that have a degree equal to that of node k. In addition to the highest-degree heuristic, several heuristics may be used to approximate the set of seeds that maximizes influence spread. We explore several additional heuristics in appendix A. Other heuristics are often based on other well-studied centrality measures from graph theory. However, as pointed out by Zhang, Zhu, Wang, and Zhao (2013) these heuristics exhibit common limitations. These limitations include the inability to account for the distance between seeds, the spreading mechanism, and the infection probability. Therefore, we report on two other centrality heuristics and analyze only the performance of the highest-degree heuristic. Firstly, betweenness centrality measures how many of the shortest paths cross through the node. The shortest path is the path across links between two nodes that passes the minimum number of nodes along the path. Secondly, closeness centrality is calculated as the sum of the length of the shortest paths between the node and all other reachable nodes in the network. We are not able to calculate the closeness centrality of nodes in the board member network as the network consists of multiple connected components. Others have proposed heuristics with similar definitions to closeness centrality, while accommodating the presence of multiple components. For example, Chen et al. (2009) provides the *distance heuristic*, which selects nodes with the smallest average shortest-path distances to all other nodes. In the distance heuristic, the shortest-path distance of two disconnected nodes is set to the number of nodes in the network.

2.4.2 Degree Discount Heuristic

As Chen et al. show, a more sophisticated approach is to use the *degree discount heuristic*. The flaw of the highest-degree heuristic that the degree discount heuristic aims to correct is its inability

to account for the inefficiency of choosing seeds that are close to each other in the network. The inefficiency of choosing proximate seeds can be explained as follows. If the distance between seeds is small, there is a chance that the nodes they infect overlap. This flaw is easily recognized in Figure 3.1, as is demonstrated in Chapter 3. The degree discount heuristic first chooses the node with the highest degree as a seed. Based on this seed, it then consecutively picks the other seeds in a similar fashion to the highest-degree heuristic. However, after choosing the first seed, the degree discount heuristic bases its choice on a value ($f(t_v, d_v)$) that is calculated based on the degree of a node and whether the node is a neighbor of nodes that have already been chosen as a seed. It discounts the degree of node v (d_v) based on the formula displayed in equation 2.1, where t_v equals the number of neighbors of v that have already been selected into the set of seeds.

$$f(t_v, d_v) = d_v - 2t_v - (d_v - t_v)t_v p \tag{2.1}$$

As the degree discount heuristic only accounts for directly neighboring seeds, it does not fully resolve the flaw that it aims to correct. Nodes that are close to each other in the network, however not direct neighbors, may still be targeted by the degree discount heuristic where it might be undesirable.

2.4.3 Greedy Algorithm

As has been shown by Kempe et al. (2003), finding the optimal set of seeds in a network given an independent cascade model is NP-hard. The greedy approximation algorithm, which was developed by Kempe et al. and later sped up by Chen et al. (2009) and Cheng, Shen, Huang, Zhang, and Cheng (2013), is able to achieve the optimal approximation guarantee within a reasonable amount of time. Greedy algorithms are a technique commonly used to provide approximations of solutions to problems that are computationally costly or impossible with current technology. This section will further elaborate on the improved greedy approximation algorithm as enhanced by Chen et al.. Greedy algorithms comprise a class of algorithms that attempt to find a globally optimal solution to a problem by consecutively finding local optima. In our example, this means that the greedy algorithm finds the optimal first seed by repeatedly simulating the infection process. Given that this first seed is selected, it then finds a second seed that adds the greatest amount of expected influence based on further simulations. The greedy algorithm generally fails to find the globally optimal solution as it never revisits solutions to previous stages. However, it is due to this characteristic that the algorithm can produce a locally optimal solution in a reasonable amount of time.

The greedy algorithm, as we have implemented it in R for the analysis in this thesis, takes two main inputs, the network G and the number of seeds to be chosen k. The algorithm starts by assigning a value of zero to a newly designated attribute s_v of each node. Following, it runs 20000

simulations of the independent cascade model for each of the k seeds that are selected. In each of these simulations, each link in the network is deleted with probability $1 - p$. Upon deletion of a random subset of all links, the algorithm checks the number of nodes reachable from each node and adds this amount to the newly created attribute s_v. After the 20000 simulations and division of the s_v attribute by 20000, the value of s_v denotes each node's average influence over 20000 simulations if it were to be chosen as the only seed. The node with the greatest value of s_v is followingly selected into the set of seeds. Following, the value of s_v is reset to zero for each node.

From the second seed onwards, the value of s_v is adjusted differently. Within each of the 20000 simulations, the algorithm again checks the number of nodes reachable from each node. However, this number is added to the s_v attribute only if any of the already selected seeds is not reachable from the given node. Again, the s_v value is divided by 20000. The s_v attribute now denotes the average additional influence from adding the given node to the set of already selected seeds. The node with the greatest expected marginal influence is selected into the seed set and this process is continued until the required number of seeds is reached. To better grasp the process of the greedy algorithm, the algorithm is displayed using pseudocode displayed in algorithm 1 (Chen et al., 2009). A number of abbreviations in the pseudocode need further explanation. Table 2.1 provides a short description of each variable used in the greedy algorithm.

Table 2.1: Greedy Algorithm Variable Description

Variables	Description
G	The network
S	Set of seeds
R	Number of simulations per seed
k	Number of seeds to be selected
s_v	Value of attribute s of node v
v	A node in the network
V	All nodes in the network
$R_{G'(X)}$	The set of nodes reachable from node(s) X

Note that we can save on running time by initially saving 20000 simulations to use for the computation of marginal gains for each of the k nodes to select. We can thereby save the computation of $20000 \cdot (k - 1)$ influence simulations. This has been recognized in the development of the StaticGreedy algorithm of Cheng et al. (2013). Saving this amount of networks does require a significant amount of memory. Nevertheless, we recommend for further research to analyze running times of the StaticGreedy algorithm in comparison to the algorithms that were implemented in this thesis.

Others have shown that the greedy algorithm closely approximates the highest achievable approximation guarantee for expected influence spread (Kempe et al., 2003). Therefore, we do not expect other algorithms to greatly improve upon the greedy algorithm in terms of expected influence.

Algorithm 1 GreedyAlgorithm(G, k) from Chen et al. (2009)

initialize $S = \emptyset$ and $R = 20000$
for $i = 1$ to k **do**
    set $s_v = 0$ for all $v \in V \setminus S$
    for $i = 1$ to R **do**
        compute G' by removing each edge from G with probability $1 - p$
        compute $R_{G'}(S)$
        compute $|R_{G'}(\{v\})|$ for all $v \in V$
        for each $v \in V$ **do**
            if $v \notin R_{G'}(S)$ **then**
                $s_v += |R_{G'}(\{v\})|$
            end if
        end for
    end for
    set $s_v = s_v / R$ for all $v \in V \setminus S$
    $S = S \cup \{\arg\max_{v \in V \setminus S}\{s_v\}\}$
end for
output S

3 Synthetic Graph Analysis

To further showcase the potential performance gains from choosing seeds based on the degree discount heuristic rather than traditional centrality heuristics, we design a synthetic network from which we can easily deduce the expected influence spread. Furthermore, analyzing the algorithms in this small-scale network allows us to check the correctness of our implementations. Figure 3.1 shows the synthetic network upon which the analysis in this chapter is based. The network consists of 26 nodes that are connected through a total of 29 links.

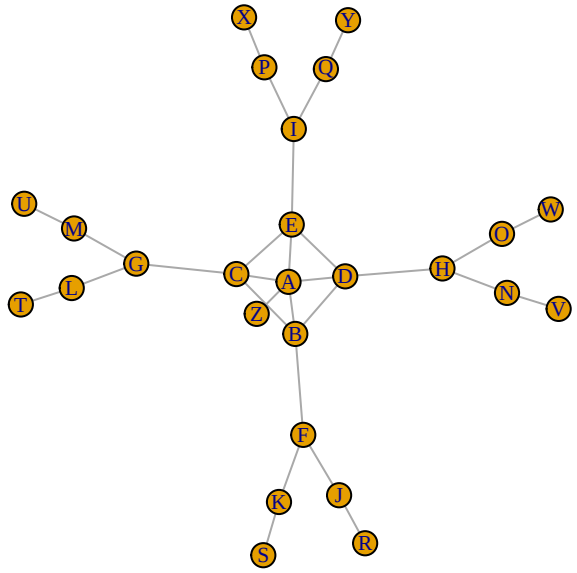


Figure 3.1: Synthetic network

3.1 Degree Heuristics

Due to the small size of the synthetic network, the expected performance of a given set of seeds can be calculated analytically as a function of the infection probability p. Within the network, there is one node with a degree of five, namely node A, other nodes all have a lower degree. The result of applying both degree heuristics to this network, given a budget of five seeds, is easily deduced. As both the highest-degree heuristic as well as the degree discount heuristic firstly choose the node with the highest degree, node A is included in the chosen sets of seeds of both heuristics. The remaining four seeds of the highest-degree heuristic are nodes B, C, D, and E, which have a degree equal to four. Unlike the highest-degree heuristic, the degree discount heuristic accounts for the fact that nodes B, C, D, and E are all connected to node A. Following from equation 2.1 and after already having chosen node A, the discounted degree of the nodes with a degree of four becomes $f(1,4) = 4 - 2 \cdot 1 - (4-1)p = 2 - 3p$, which is below the degree of nodes F, G, H, and I for any

value of p. From this simple analysis we can thus conclude that in any case, the highest-degree heuristic would choose nodes A, B, C, D, and E as seeds while the degree discount heuristic always chooses nodes A, F, G, H, and I. Simulation of both algorithms shows that this is indeed the case, indicating that the algorithms were correctly implemented.

3.2 Influence of the Heuristics

As the synthetic graph is relatively small, we are able to derive a formula of the expected influence of a given set of seeds for the independent cascade model. Let us analyze the expected influence of seeds A, B, C, D, and E. Due to the structure of this particular network, none of the remaining nodes can ever be infected via more than one link. This allows for a relatively easy calculation of expected influence. Our influence model assumes that all targeted nodes are guaranteed to be infected. The expected influence of a set of seeds is equal to the sum of all probabilities of infection over all nodes in the network. As we start by selecting five seeds, the predicted influence will at least be equal to five. Given that five seeds are within one link's reach of the seed set, we add five times the infection probability p to the expected influence. Nodes that are further from the seeds are accounted for in the following manner. Conditional on infection successfully flowing through a link that connects a seed node to another node, further infection again occurs with probability p, meaning that nodes J to K are all infected with probability p^2. In a similar fashion, the remaining eight nodes are infected with probability p^3. This then yields the expected influence of the set of seeds chosen by the highest-degree heuristic to be described by equation 3.1.

$$Influence_{deg}(p) = 8p^3 + 8p^2 + 5p + 5 \tag{3.1}$$

The analysis of the seeds chosen by the degree discount heuristic is significantly more involved. To calculate the total expected influence of seeds A, F, G, H, and I, we again take the sum of each nodes' probability of being infected over all nodes. For some nodes in the synthetic graph, i.e. nodes J to Z, these probabilities are fairly easily calculated. However, within the remaining part of the network, which is visualized in Figure 3.2, nodes B to E can be infected via a great number of routes.

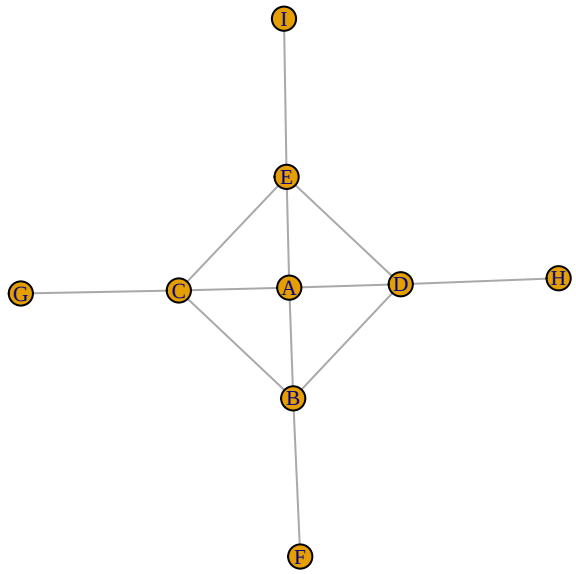


Figure 3.2: Complex part of the synthetic network

One way to determine the expected influence for these nodes is to list all possible combinations of *successful* and *unsuccessful* links. A link is successful if it allows information to be successfully spread to one of the nodes that it connects. A link does thus not actually have to spread infection in order to be deemed successful. Since each of the twelve remaining links is either successful with probability p or unsuccessful with probability $1-p$, we must analyze a number of $2^{12} = 4096$ cases. The probability of a particular case occurring is dependent on the number of links x that succeed to transfer information in case E. The corresponding probability of event E to arise is governed by equation 3.2.

$$P(E) = p^x + (1-p)^{12-x} \tag{3.2}$$

The expected influence within the remaining part of the network is equal to the list of 4096 probabilities following from equation 3.2 multiplied by their respective number of infected nodes. To solve for the expected influence within the remaining part, we have constructed a 4096 by 12 matrix listing all cases as binary combinations indicating for each link whether it is successful or not. Using this matrix, we have used a logical rule for each of the remaining non-infected nodes B, C, D, and E to determine whether they are infected in a given event. Following, we list the counts of all combinations of the number of infected nodes and the number of successful links. Applying equation 3.2 to all of these counts, multiplied by their number of infected nodes and substituting x for the number of successful links, we obtain the expected influence of this complex part of the synthetic network. We add to this the expected influence of the more easily analyzed other part of the synthetic network to obtain the total expected influence of the seeds chosen by the degree

discount heuristic as a function of p. This formula is represented by equation 3.3.

$$Influence_{dd}(p) = 12p^{12} - 112p^{11} + 452p^{10} - 1024p^9 + 1400p^8 - 1128p^7 + 440p^6 + 8p^5 \\ -40p^4 - 24p^3 + 20p^2 + 17p + 5 \tag{3.3}$$

Using equations 3.1 and 3.3, we can plot the expected performance of both degree heuristics as a function of p. In Figure 3.3, both equations are plotted as lines. We run random influence simulations across a number of values for p to check the accurateness of the predicted influence along with the R implementation of the influence process according to the independent cascade model. Each point in Figure 3.3 represents the average influence across 1000 simulations for a given value of p.

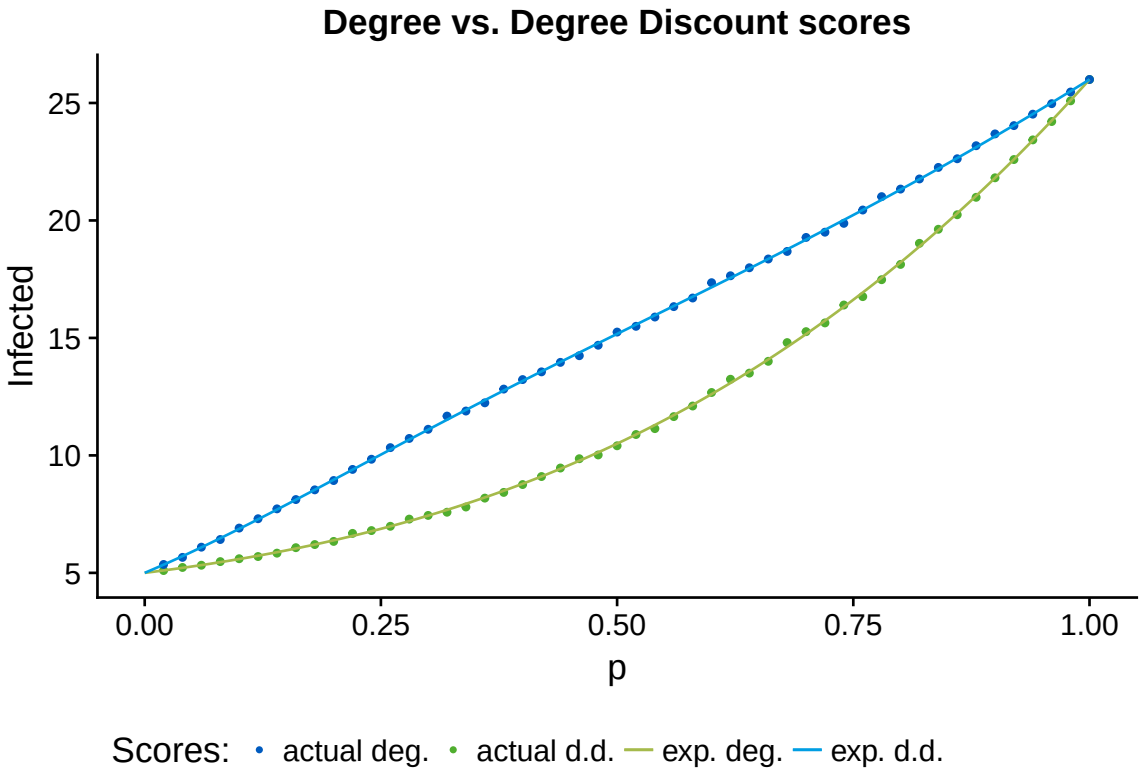


Figure 3.3: Synthetic graph performance simulations

We can see that for any value of p, the seeds chosen by the degree discount heuristic perform at least as well as the seeds chosen by the highest-degree heuristic. Naturally, the two algorithms perform equally well if information flows successfully through every link in the network ($p = 1$), or if information fails to be transferred across every link ($p = 0$). To get a better representation of how the performance of the degree discount heuristic compares to the performance of the highest-degree heuristic, Figure 3.4 shows the expected relative influence difference of both heuristics. Both influence scores have been adjusted downwards by their initial number of seeds to get a better representation of the performance difference. Equation 3.4 describes the performance difference

ratio that is graphically shown in Figure 3.4.

$$Perf\_Ratio_{dd/deg.} = \frac{Influence_{dd}(p) - 5}{Influence_{deg}(p) - 5} \tag{3.4}$$

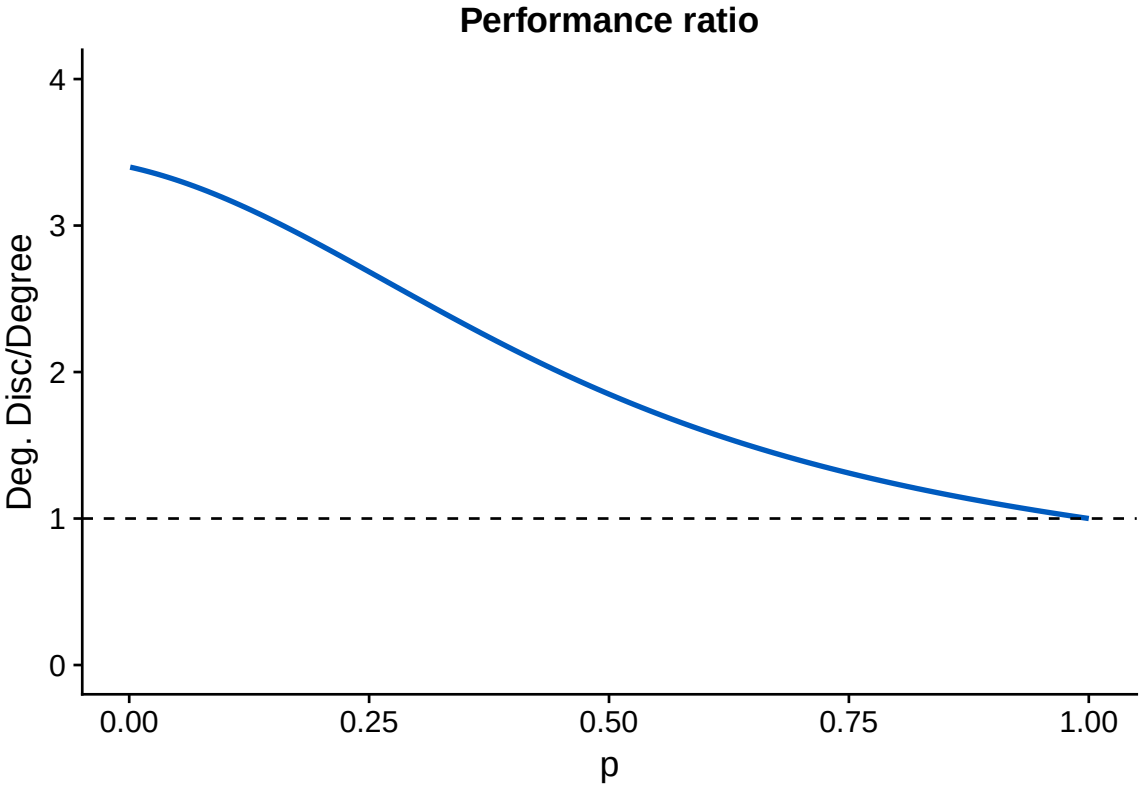


Figure 3.4: Degree discount versus degree: average influence difference

3.3 Conclusions

Although the synthetic network exhibits a structure in which the benefits of the degree discount heuristic are clearly identified, the analysis has clarified a number of factors. First of all, the degree discount heuristic shows its ability to outperform the highest-degree heuristic in terms of influence by more than 200%. Second, the influence difference is negatively related to p. This corroborates footnote 3 of Chen et al. (2009), which states that in the independent cascade model with relatively large infection probability, influence spread is not very sensitive to different algorithms and heuristics. Third, the performance of the highest-degree heuristic never exceeds the performance of the degree discount heuristic in this particular network.

4 The Network of Interlocking Directorates in Dutch Healthcare

This chapter explores the network of healthcare providers using network analysis techniques. Firstly, we introduce the data from which our network representations are constructed. Secondly, we discuss the process of constructing the board member network in R. Thirdly, we examine the characteristics of the network. In particular, we study the degree distribution of the network, as this has been found to assist the interpretation of the network structure. Lastly, we explore the construction of the firm-side network with sector attributes. As policies are typically sector specific, we explore the appearance of links between different sectors.

4.1 Data

Network analysis techniques rely on data regarding nodes in the network and the links between them. The execution of governance of healthcare providers in the Netherlands consists of two councils. On the one hand, the board of directors is concerned with the provider's governance. According to Jeroen Bosch Ziekenhuis (2015), the main aim of the board of directors is to safeguard and expand the policy, mostly aimed at increasing the quality and safety of the provided care and services. On the other hand, the supervisory board is concerned with the supervision of the board of directors. The supervisory board is informed by the board of directors and is focused on the inspection of the operations regarding, strategy and policy, financial and economic approach, quality of care, administrative and legal procedures, and training and scientific research. Data is gathered from a questionnaire that Dutch healthcare providers are annually required to complete in order to comply with Dutch semipublic-sector regulations in the context of social responsibility (CIBG, 2018). The correctness and completeness of the data are therefore dependent upon providers themselves. These documents consist of two main components. On the one hand, the DigiMV contains detailed information regarding the accountability of firms. On the other hand, WNT data comprises data regarding board members and top earners of each firm. Names of the members of the supervisory board and the board of directors are included in a dataset constructed from a combination of the DigiMV and WNT datasets. The data includes information regarding the gender, name, and role of each board member, as well as various data regarding the firm. The DigiMV dataset includes 6042 entries which all describe a member of a supervisory board. Before the start of writing my thesis, a script was written to prepare this dataset for the analysis. The data preparation script goes through the process of providing a unique identification to each person based upon their names while taking account of prefixes and suffixes. The dataset is then manually checked for errors using online resources such as LinkedIn profiles as a backup. To combine this

DigiMV dataset with the WNT dataset, a number of steps have been followed using a script that was written using Python. The WNT dataset consists of 15613 entries describing medical specialists and members of both the supervisory as well as the executive boards. The Python script drops all entries regarding medical specialists, aligns the WNT data with the DigiMV data and merges them. We drop duplicate entries that are completely similar across the two datasets. Following, we try to identify each person in the newly merged file as accurately as possible. That is, if two entries are describing the same person on different boards, we aim to tag them with the same identification number. To do this, the script first deletes all dots, commas, suffixes, prefixes, and information that is provided between parentheses. We use the first appearing space character to isolate initials and we assume two entries to be the same person if their gender, last name, and at least their first two initials are the same. After this procedure, 2009 entries are identified as potentially being duplicate, meaning their combination of gender, first initial, and last name appears somewhere else in this list. These 2009 entries are then manually identified using sources such as LinkedIn. The resulting dataset consists of 11844 entries describing 10137 individuals. This indicates that, on average, a board member sits on 1.17 boards.

4.2 Construction of a Network of Board Members

To construct a network from the data that was described in Section 4.1, we use the *igraph* package in R (Csardi & Nepusz, 2006). To create a network object in R, the data is first transformed into an undirected edge list. An edge list describes each link by identifying source nodes and target nodes. Each row in the edge list describes a link, the first two columns describe the nodes that are involved in the link. Any further columns are used to display descriptive attributes. To create the edge list, we consider each firm separately. The board members of each firm are added to a list. Each possible combination of length two is extracted from the firm-level list of board members and added to the edge list along with the name of the firm as an attribute of the link. The edge list is transformed into a network object. This network object is plotted in Figure 4.1.

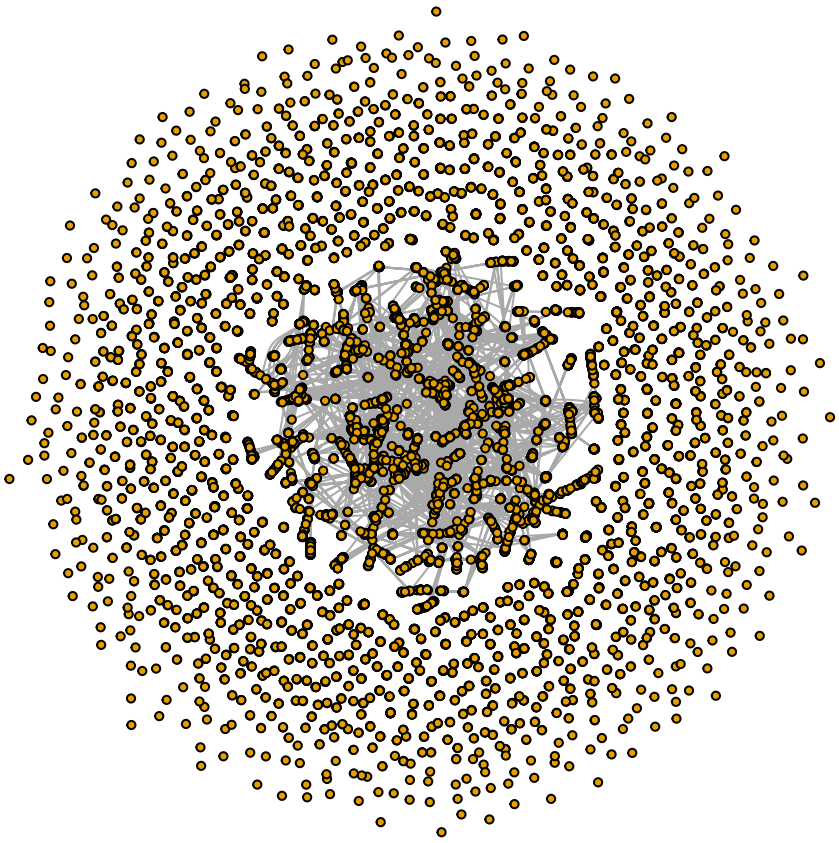


Figure 4.1: Entire network plot

Although the observed network is arguably too large for a plot of the whole network to be informative, there are a few key points to be made about Figure 4.1. Firstly, we observe one giant connected component in the middle of the plot, surrounded by a great number of nodes that are not part of a greater subnetwork. Note that while many of the nodes that are present in the outer ring seem not to be connected, they are still connected to all other board members of the same firm.

4.3 Network Descriptives

The board member data describes the membership of 2296 healthcare providers. The total dataset includes 3631 directors and 6506 supervisory board members. Due to a large number of nodes and links, the overall network characteristics might be best interpreted using descriptive statistics. In the observed network, there are 10137 nodes. Between these nodes, there are 37648 links. The network density, which captures the extent to which the nodes in the network are connected, is equal to 0.73%, meaning that of all possible links, 0.73% is realized. There are a total of 1352 components and 47.41% of the nodes are present in the greatest component. The greatest component contains 4806 nodes, while the second-greatest component already contains merely 28 nodes. The average degree in the network is 7.43.

4.3.1 The Degree Distribution

Degree distributions of networks are commonly used to make inferences regarding the generation and structure of a network. Since the network structure is important for the diffusion of information, we plot the degree distribution that the observed network is characterized by. It should be noted that the degree distribution does not allow for the identification of a key characteristic of nodes in the network. Namely, the degree of a node can be affected by two features. On the one hand, a node might have a great number of links due to the number of boards it is serving on. On the other hand, the number of links might merely be a result of the size of the board it is serving on. The degree distribution of the board member network is shown in Figure 4.2.

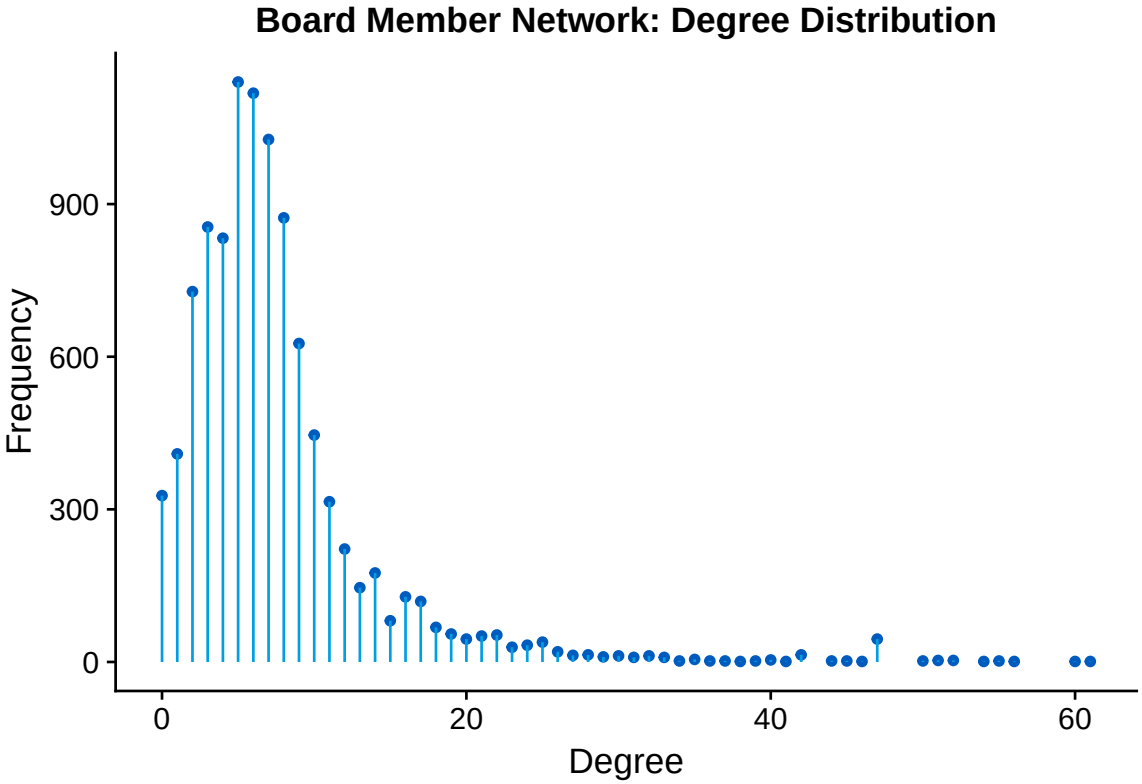


Figure 4.2: Board member network: degree distribution

We may conclude from Figure 4.2 that it is most common for board members to be linked to five other board members. Due to pooling of the two board tiers, having zero links requires a one-man board without supervisory board members and having no membership in the boards of other firms.

Two types of degree distributions are commonly distinguished. First, we have random networks. In random networks, we assume all links to be formed according to a given probability. Random networks are expected to exhibit degree distributions that approximate a binomial distribution. For large networks, the binomial distribution becomes indistinguishable from a Poisson distribution, according to the Poisson limit theorem (Papoulis & Pillai, 2002). Second, we have scale-free

networks. In scale-free networks, the probability of a link successfully being formed is proportionally dependent on the degree of the nodes that are involved in the link. This phenomenon is known as preferential attachment, which is a concept that was, to our knowledge, first posed by Yule et al. (1925) to explain power-law relationships appearing in evolution science. The rationale behind preferential attachment is that a new node is more likely to link to a well-known other node than to less-known nodes precisely because that node is well known. As explained by Barabási and Pósfai (2016), we expect the degree distribution of scale-free networks to be well approximated by a power-law distribution, which is characterized by equation 4.1. In equation 4.1, k characterizes the degree and γ is the *degree exponent* of the given degree distribution. On a log-log scale, equation 4.1 should be linear with a slope equal to $-\gamma$.

$$p_k \sim k^{-\gamma} \tag{4.1}$$

A power law distribution explains the relationship between two variables if the change of one variable is relatively proportional to the change in the other variable. Many real-life networks have been reported to have a degree distribution that follows a power law. Networks that exhibit the scale-free property generally have hubs and spokes. This is what sets apart scale-free networks from random networks. In scale-free networks, as opposed to random networks, only a few nodes have to be taken away from the network in order to disrupt the entire network and break it up into multiple components. However, in scale-free networks with degree exponents greater than three ($\gamma > 3$), hubs become too small to have a significant impact on the average distance between nodes in the network, and it becomes difficult to distinguish the structure of scale-free networks from the structure of random networks.

To learn about the generation of the network of board members, we need to remind ourselves of the multi-layer structure of firms and individuals. We expect the board member network to change in a particular manner. The board member network changes if a board member is replaced or added to a firm. The new board member immediately links to all other board members of the firm. This generational phenomenon contradicts random network generation and the independence between the degree of a node and its probability of forming links. In fact, if a new board member is added, it is likely that it only links to a few nodes that are all present in a specific part of the network. Furthermore, the degree of the nodes that a new board member links to is heavily dependent on the number of board members of the firm in question. In the current board member network setting, it is difficult to imagine preferential attachment to be present. Figure 4.3 corroborates the absence of a power law in the network of board members as we do not observe a linear relationship. However, using only the degree distribution, we might not be able to make inferences regarding the existence of preferential attachment or random processes, as the underlying mechanism in the board member network is arguably too complex.

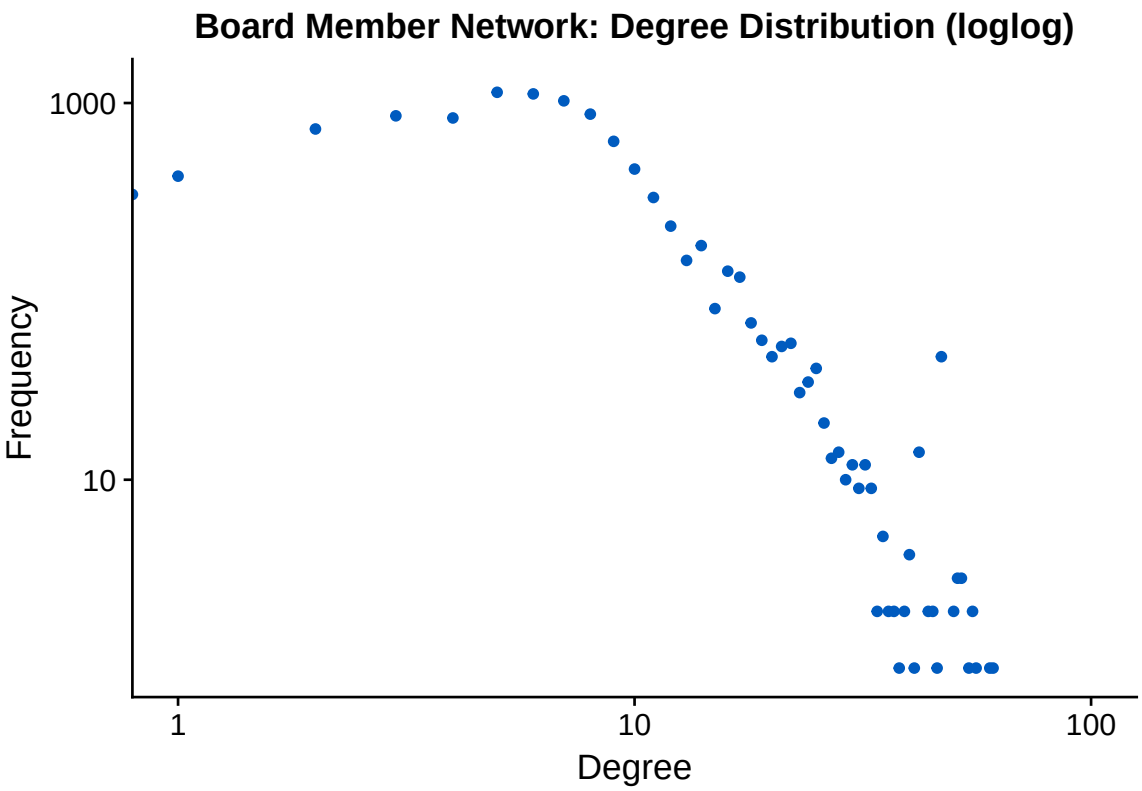


Figure 4.3: Board member network: degree distribution (loglog)

We might be able to learn more about the generation of the observed network structure by merely considering the individuals that establish interfirm relationships. We, therefore, construct an alternative network from our original board member network by removing the board members that serve only one board. In other words, we construct a social network of people that serve on the boards of multiple firms. The greatest connected component of the resulting network is displayed with sector-colored links in Figure 4.4. The degree distribution of this network is plotted using a log-log scale in Figure 4.5. The line plot in Figure 4.5 shows a poisson distribution with a γ equal to the average degree in the observed network.

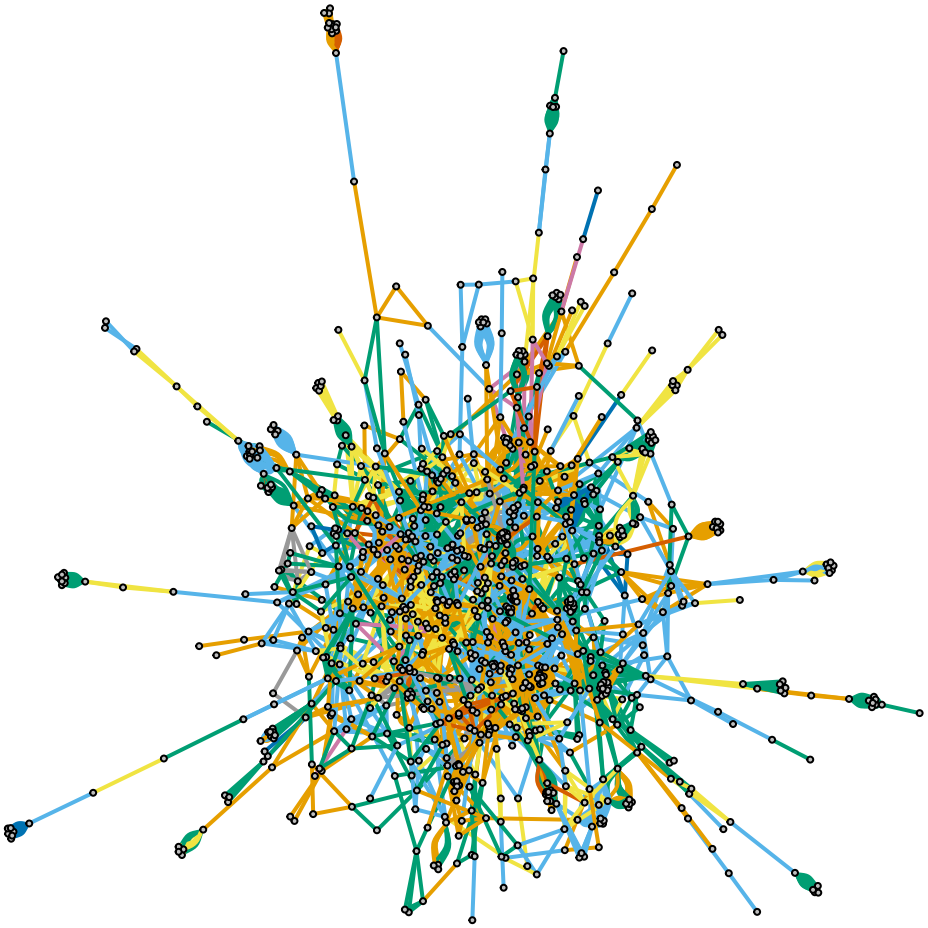


Figure 4.4: Network of individuals serving multiple boards

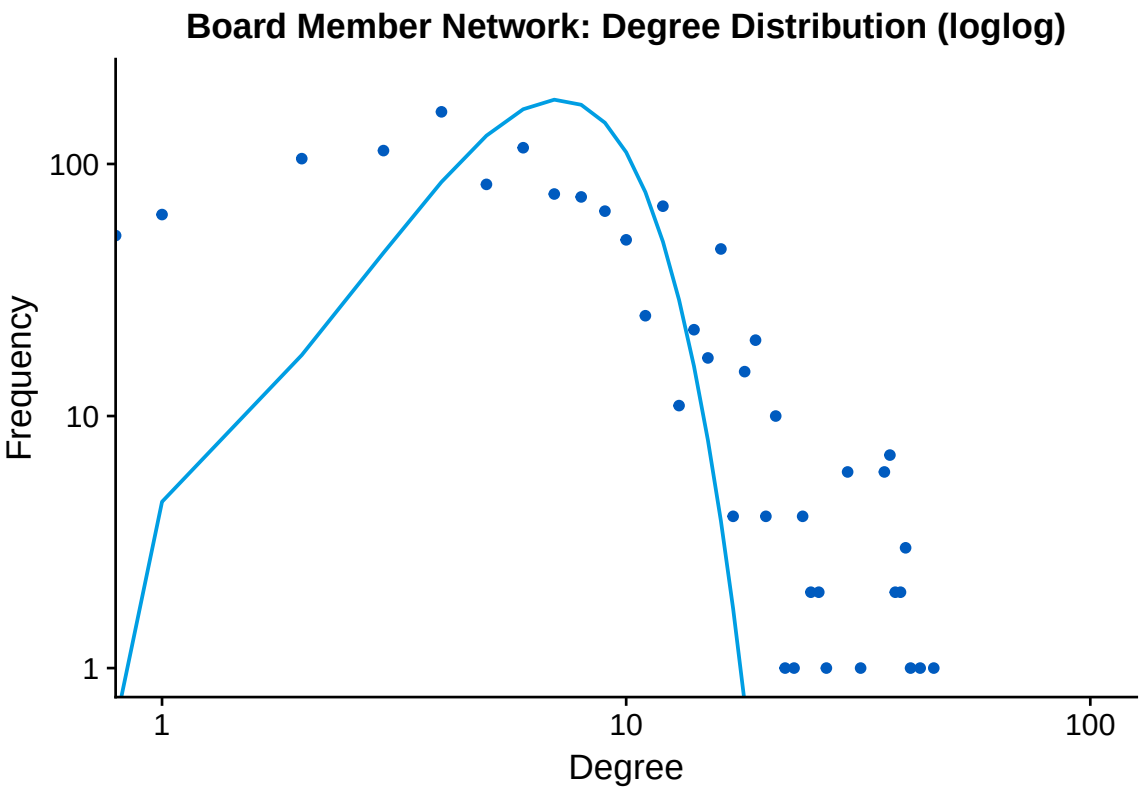


Figure 4.5: Individuals serving multiple boards: degree distribution (loglog)

Again, we do not observe a linear relationship on the log-log plot. Furthermore, the degree distribution does not seem to match the poisson distribution well. Therefore, we find that the results do not indicate that the discussed theories do not clearly describe the possible generation of the network of board members. Nevertheless, a remarkable appearance of the power-law distribution does appear in the board member network, namely, for the number of boards per individual. Note that this distribution is the same as the degree distribution of the individuals in the bipartite approach of the network, which was described in Section 2.2. To arrive at a distribution of the number of boards per individual, we take the initial data that simply lists the memberships of all boards and we count the number of times that each individual appears in the list. The resulting distribution is plotted on a log-log scale in Figure 4.6.

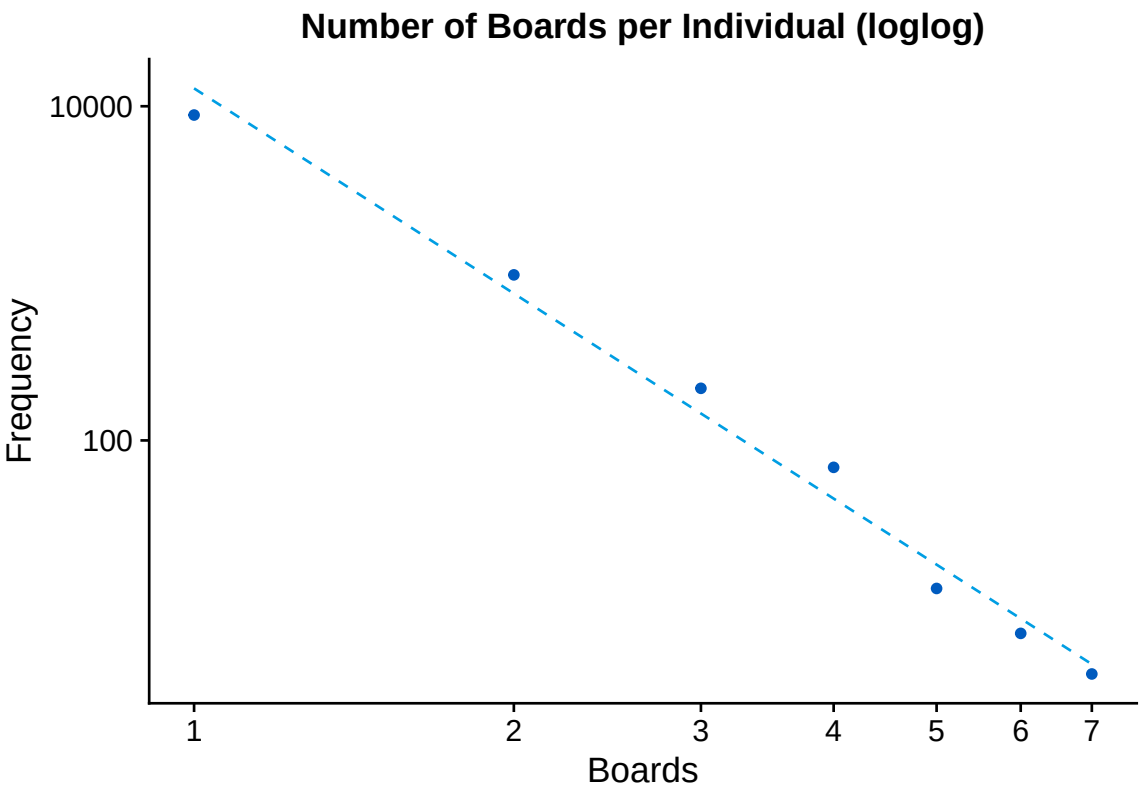


Figure 4.6: Number of boards per individual (loglog)

In Figure 4.6 we observe that the number of boards per individual is well described by a power-law distribution since it appears to follow a linear relationship on a log-log scale. The observed distribution of the number of boards per individual indicates that while many people are seated at just one or two boards of healthcare providers, a small number of people manage to be active on the boards of six or seven firms.

4.4 Networks of Firms

In Section 4.2, we constructed a social network of board members. However, as we have shown in Section 2.2, we can take a different approach of network construction by taking firms as nodes and shared membership of firms as the links between those nodes. In this way, we construct a network of firms, as has been done by Westra (2017). The extent to which firms have the potential to be heavily connected is dependent on the number of board members it has. The frequency distribution of combined board sizes, i.e. the total number of members seated in the two board tiers, is shown in Figure 4.7.

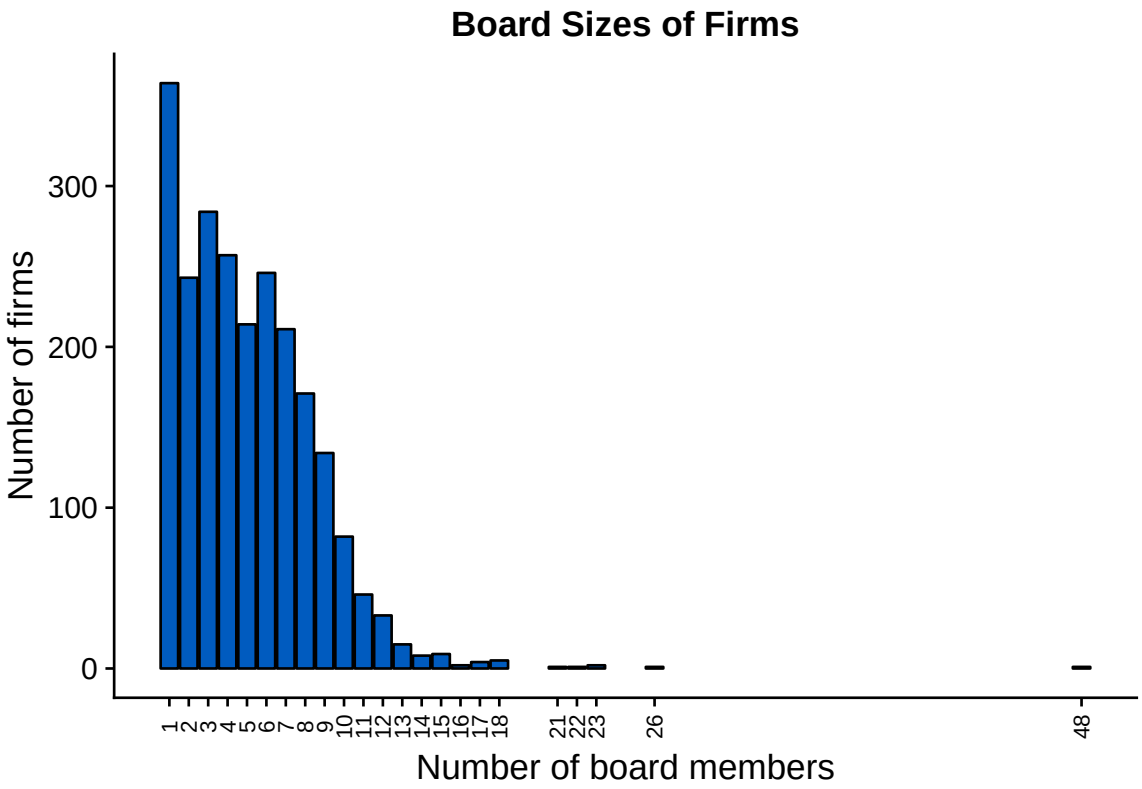


Figure 4.7: Number of board members per firm

In Figure 4.7 we observe that the firm with the greatest number of board members[1] has 48 board members and one-man boards without supervisory board members are most common. To further explore the occurrence of links between firms, we construct a network of firms. Upon construction of the firm-side network of interlocking directorates, we may add one of nine possible sectors to each firm as an attribute of the node. A sector is characterized by the type of care that is typically provided by the firms operating in it. The DigiMV dataset includes data concerning the revenue for each sector for individual firms. We label firms with a given sector name based upon the sector for which it has the greatest revenue. Firms for which no revenue was reported receive a label based upon which of the nine sectors it states to be a provider of. If multiple sector names are provided, we select a sector based on a ranking of sectors. Firms of which the sectors are not included in the nine main sectors will be labeled as being of the "other" sector, and firms which are have not provided a sector label at all will be labeled as "unknown". The ranking of sectors is loosely based on the typical firm size in each sector, i.e. a University Medical Center is typically larger than a private clinic in terms of revenue and number of patients. The sector ranking is as follows.

1. University Medical Center (UMC)
2. General hospital (AZKH)

[1]The firm with the greatest number of board members, Exodus Nederland, is a social care organization which operates five facilities in different regions as well as an umbrella organization. The board mainly consists of members who are dedicated to individual branches of the organization.

3. Categorical hospital (CatZKH)
4. Mental healthcare (GGZ)
5. Disability healthcare (GHZ)
6. Nursing homes, assisted living, and home care (VVT)
7. Rehabilitation care (REV)
8. Private clinics (ZBC)
9. Social shelter and women's shelter (MOVO)
10. Other

Upon labeling each node in the provider network with a specific sector name, we color each node based on their sector. The provider network plot is shown in Figure 4.8.

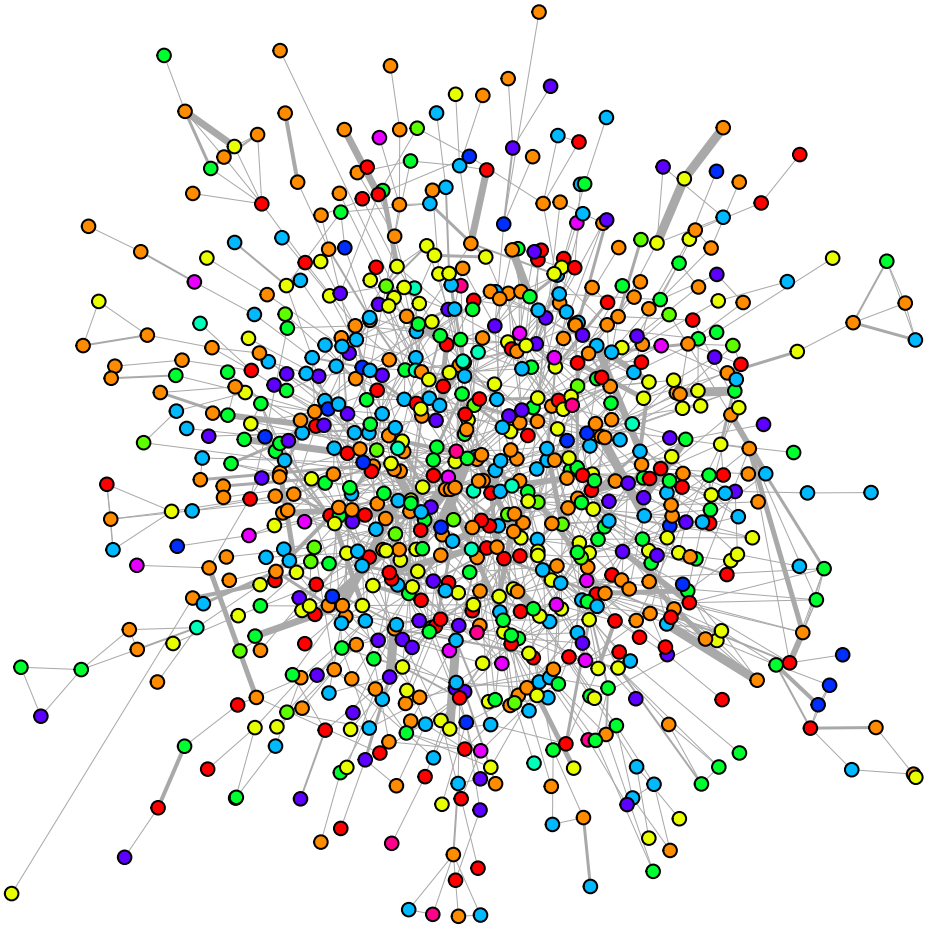


Figure 4.8: Sector-colored provider network

To further inspect the relations between sectors, we inspect the occurrence of links across sectors. Table 4.1 shows the number of links between different sectors.

Table 4.1: Between-sector links of firms

	# Firms	UMC	AZKH	CatZKH	GGZ	GHZ	VVT	REV	ZBC	MOVO	other	unknown
UMC	8	1	12	2	2	1	4	1	1	1	0	9
AZKH	69		30	11	50	31	76	10	29	11	6	59
CatZKH	17			0	5	3	12	1	5	0	0	5
GGZ	392				81	68	108	7	32	11	10	74
GHZ	296					42	134	13	20	8	8	53
VVT	718						169	16	46	22	19	99
REV	22							0	7	2	2	12
ZBC	315								92	1	8	42
MOVO	64									3	1	10
other	92										2	16
unknown	341											73

The data in Table 4.1 confirm the intuition that followed from Figure 4.8 of many links being present between different sectors. However, to test for preferences for links between certain sectors, we use a binomial logistic regression. Here, we leave out other and unknown sectors. Furthermore, we do not analyze the MOVO sector, as this sector is not regulated by the NZa. In order to conduct a binomial logistic regression analysis, we use the data shown in Table 4.1. Firstly, we compute the number of links that is theoretically possible between any two of the eight relevant sectors. If firms from different sectors are involved, this number is given by the product of the number of firms in each sector. If two firms of the same sectors are involved, the number of possible links is given by $\frac{n(n-1)}{2}$, where n is the number of firms in the sector in which both firms operate. As an illustration, consider the following examples. The number of firms in the UMC and REV sectors is respectively 8 and 22. Therefore, the possible number of links that is theoretically possible between the firms of the UMC and REV sectors is equal to $8 \cdot 22 = 176$ and the number of possible links between firms within the UMC sector is equal to $\frac{8(8-1)}{2} = 28$.

Using this data, the binomial logistic regression estimates the probability of a link being established between two firms, given the sectors in which the firms operate. Thus, for a given combination of two sectors (i), i.e. UMC-GGZ, the number of realized links for that combination (Y_i), and the total possible number of links for the combination (n_i), the binomial logistic regression fits the probability of a link being established (p_i) and the associated confidence interval. The underlying distribution for our model is thus characterized by equation 4.2.

$$Y_i \sim Binom(n_i, p_i) \tag{4.2}$$

The model computes a 95% confidence interval around the fitted probabilities. However, the regression fails to compute a confidence interval if the fitted probabilities are equal to zero. In such an instance, a confidence interval ranging from 0 to 1 is produced. Figure 4.9 shows the results of the binomial logistic regression.

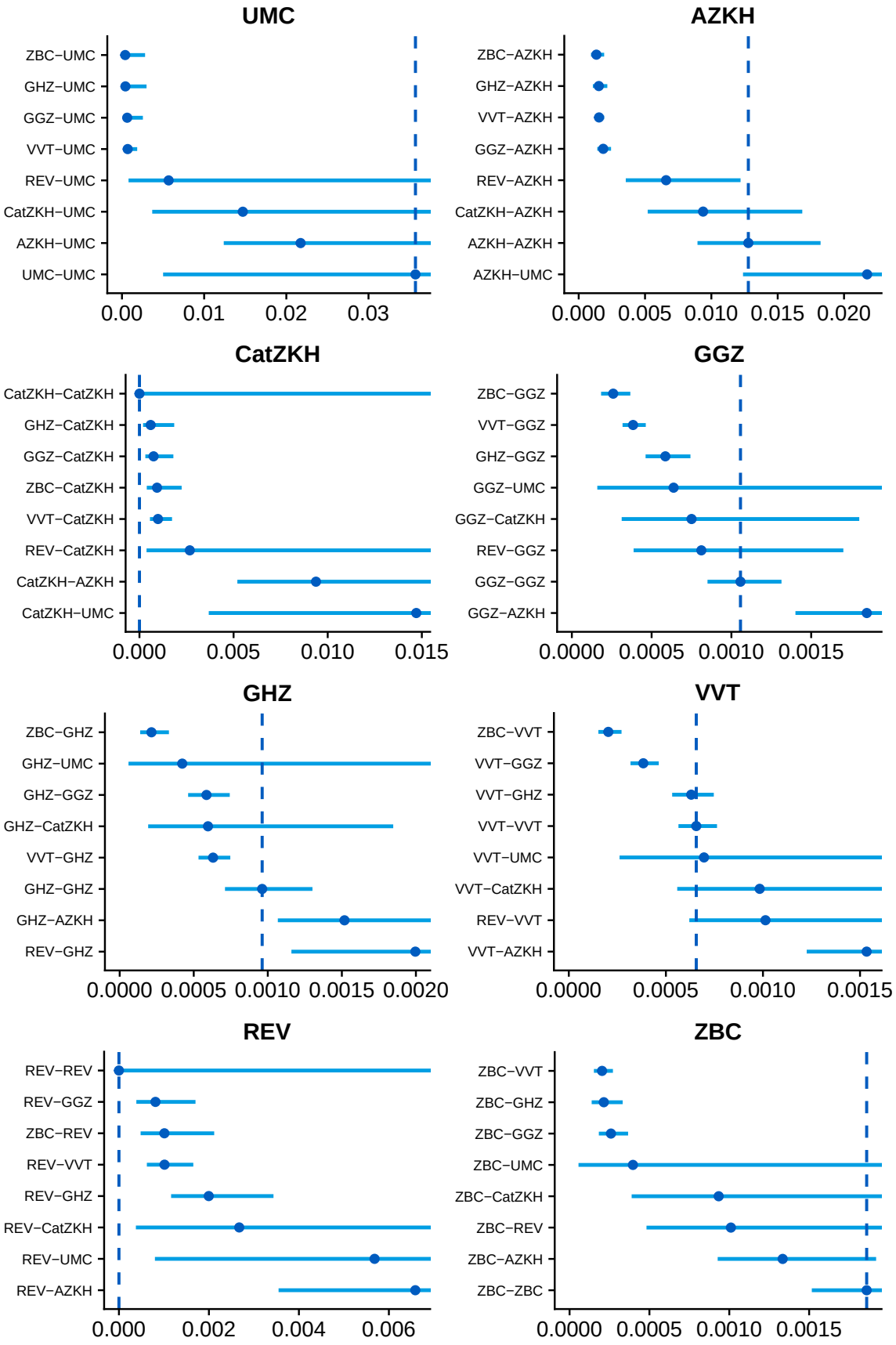


Figure 4.9: Between-sector link likelihood

We gain a number of insights from the results of the binomial regression analysis of links between healthcare providers across the most relevant sectors in Dutch healthcare. First of all, while it seems relatively likely for a link to be realized if the two firms are within the same sector for some sectors (ZBC, UMC, AZKH), the opposite seems to be the case for some other sectors (REV, CatZKH). Furthermore, 4.9 shows that some sectors exhibit greater overall probabilities of linking than other sectors. In particular, we find that the hospitals (UMC, AZKH, CatZKH) generally display greater probabilities, showing estimated probabilities above 1%. In general, the ZBC, GHZ, GGZ, and VVT sectors generally exhibit estimated probabilities ranging from 0 to 0.2%. The REV sector seems to deviate from all other sectors in terms of overall probabilities, as their estimated probabilities range from 0 to 0.6%. In order to get a more accessible overview of the estimated probabilities of links between sectors, we construct a network of the eight sectors that we have analyzed in our binomial logistic regression. Figure 4.10 shows the resulting network visualization. In this visualization, the width of a link and the distance between two nodes reflects the estimated probability with which firms of the two sectors link with each other. For clarity, we have not shown the probabilities of links within sectors, i.e. UMC-UMC. Furthermore, we have marked the sectors concerning hospitals using a blue polygon. Once again, we find that the hospitals appear to form a community in this small network.

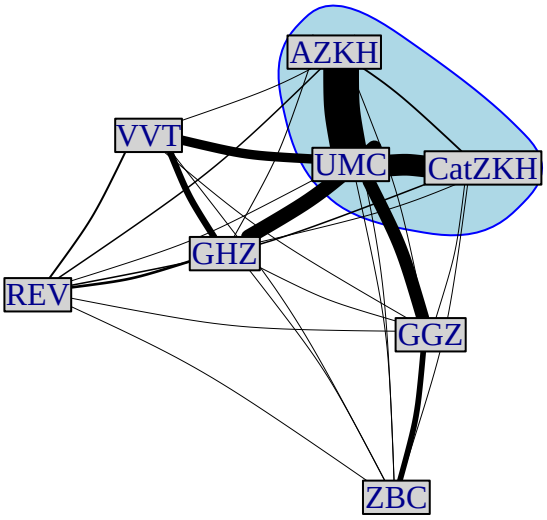


Figure 4.10: Network visualization of between-sector link probabilities

5 Influence Simulations

In this chapter, we compare different algorithms for choosing influential nodes in terms of their expected influence spread (performance), and computational costs. Many different algorithms are currently available in the literature. Here, we compare the degree heuristic, the degree discount heuristic, and the greedy algorithm. We analyze the performance of the algorithms using a benchmark scenario where we choose 10 seeds and have a 10% infection probability. Next, we test how our results change upon varying the number of chosen seeds and infection probability. As the heuristics have short running times, we are able to analyze their expected influence spread in greater detail. To test whether the variance of the performance distributions has any significance in the algorithm choice, we study the case in which the difference in variance appears greatest using expected utility theory, which is the reigning normative theory of decision making under risk.

5.1 Comparing All Three Seed-Picking Methods

5.1.1 Computation Costs

Here, we shortly elaborate on the computational costs of the three algorithms. We use the R implementation that has specifically been written as a part of this thesis. The R functions require an *igraph* network object with degree and name attributes for each node. Further inputs are the required number of seeds and the relevant infection probability. The functions are relatively easy to use, as earlier implementations are mainly written in arguably less accessible programming environments such as *C++*. The implemented algorithms differ greatly in their computation times. Approximations of the computation times of the seed-picking methods on a 1.4 GHz Intel Core i5 MacBook Air (4Gb RAM) are as follows. Choosing seeds costs the highest-degree heuristic 0.007 seconds regardless of the number of seeds. The degree discount heuristics takes 0.05 seconds to select 10 seeds. Furthermore, the computation time of the degree discount heuristic increases by 0.005 times for each node that is added. Unlike the degree heuristics, the greedy algorithm is computationally costly, taking 45 minutes per selected seed.

5.1.2 Expected Influence Spread

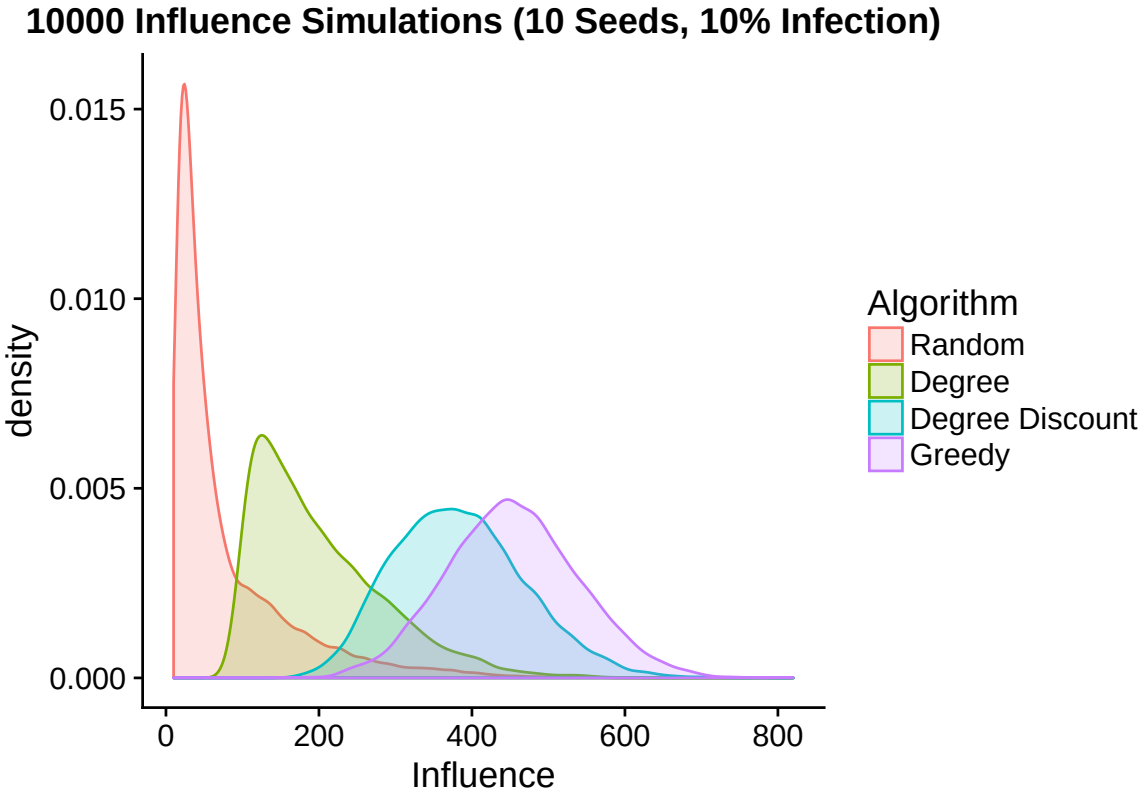


Figure 5.1: Algorithm performance distributions

To compare all algorithms in terms of their performance, we start our analysis using a seed selection of 10 seeds given a 10% infection probability. Figure 5.1 shows a density plot of the achieved performance scores over 10000 influence simulations for each of the compared algorithms. We find that of all tested algorithms, the greedy algorithm achieves the greatest expected influence spread, followed by the degree discount heuristic. The resulting performance distributions of the algorithms are as we would expect them according to Chen et al. (2009). While choosing seeds merely based on degree scores already provides a significant gain in influence compared to choosing randomly, more complex algorithms such as the degree discount heuristic and the greedy algorithm may provide more than two times greater influence spread. While the synthetic network that was analyzed in Chapter 3 was designed to show a great difference between both of the degree heuristics, Figure 5.1 suggests that this difference might not be very different from what we observe in the actual board member network. Due to the structure of our network, the highest-degree heuristic chooses a large number of people from the firm with the greatest number of board members. In fact, 7 out of the first 10 seeds and 11 out of the first 20 seeds chosen by the highest-degree heuristic is a board member of the firm with 48 board members.

We check the robustness of this result by varying parameter values. Here, we are interested in two

main effects. On the one hand, we seek to clarify how the difference between the algorithms differs for the number of chosen seeds, i.e. the total capacity for our hypothetical meeting. On the other hand, we explore the effect of having different infection probabilities (p), i.e. the extent to which a message will be passed on. Due to the long computation time of the greedy algorithm, a limited number of comparisons involving this algorithm is presented in this thesis. Using Figure 5.1 as a benchmark, we compare the results of the same process using multiple deviating parameter values. Figure 5.2 and 5.4 show the results of these analyses.

5.1.3 Varying Number of Seeds

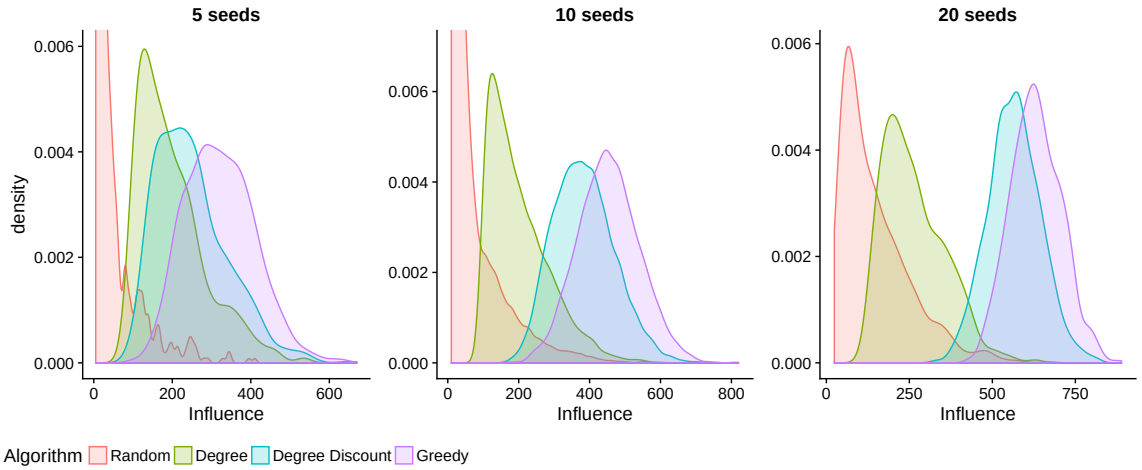


Figure 5.2: Algorithm performance distributions: varying number of seeds (10 percent infection)

Table 5.1: Performance difference for varying number of seeds (10 percent infection)

	Greedy/Degree Discount	Degree Discount/Degree
5 seeds	1.29	1.26
10 seeds	1.18	1.91
20 seeds	1.12	2.14

Using the results in Figure 5.2, we investigate how the number of seeds affects the difference in influence spread among the different algorithms. Increasing the number of seeds from 10 to 20 does not lead to greatly different outcomes. Still, the degree discount heuristic outperforms the highest-degree heuristic, now only by a slightly greater margin. Similar to our benchmark results, the greedy algorithm performs best, however, not substantially better than the degree discount heuristic. Furthermore, decreasing the number of seeds to 5 does lead to different outcomes. In this case, the influence spread differs less severely among the different algorithms. The greedy algorithm outperforms the highest-degree heuristic by about a factor of 1.6 on average whereas the degree discount heuristic outperforms the highest-degree heuristic by a factor of 1.3 on average. Table

5.1 shows the difference in performance for all investigated numbers of seeds. If we may conclude anything from these results then we would expect the difference between the degree heuristics to be positively related to the number of chosen seeds.

A possible explanation for such a finding might be that as the number of seeds increases, the seeds chosen by the different heuristics are overlapping to a lesser extent. We test this hypothesis by reporting the percentage overlap for the first 100 seeds and report the outcome in Figure 5.3, the points indicating a number of 5, 10 and 20 seeds are colored red. We find that for the three chosen seed set sizes, the percentage overlap is indeed decreasing as the number of seeds increases. Furthermore, the difference between the greedy algorithm and the degree discount heuristic seems to be decreasing as the number of seeds increases.

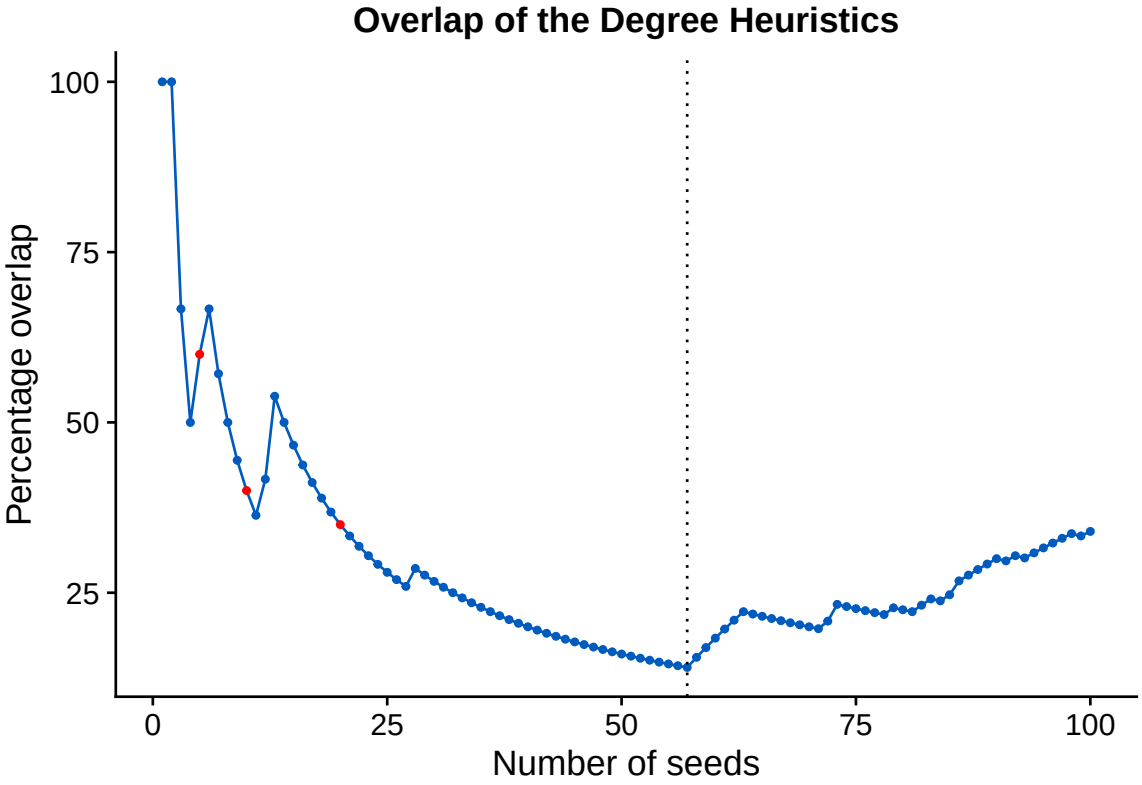


Figure 5.3: Overlap of the degree heuristics

5.1.4 Varying Infection Probability

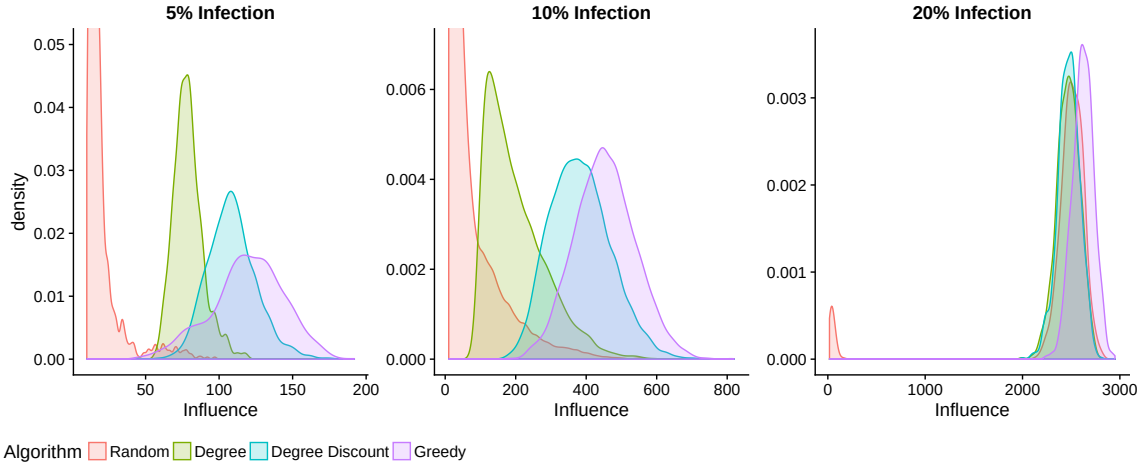


Figure 5.4: Algorithm performance distributions: varying infection probability (10 Seeds)

Table 5.2: Performance difference for varying infection probabilities (10 seeds)

	Greedy/Degree Discount	Degree Discount/Degree
5%	1.10	1.38
10%	1.18	1.91
20%	1.06	1.00

Secondly, we study the relation between the chosen infection probability p and the influence spread resulting from the different algorithms. Increasing the infection probability from 10% to 20% significantly affects the performance difference between the various algorithms. We find that if we increase the infection probability to 20% the two degree heuristics do not show to be vastly more effective than simply choosing seeds randomly. Nonetheless, if seeds are chosen randomly, a small chance exists that none of the ten seeds are part of the greatest component in the network, meaning that only a small part of the network can be infected. Furthermore, the greedy algorithm still is able to outperform the other methods of choosing seeds. This may be the result of the greedy algorithm being more effective at choosing seeds that are located further from each other in the network. Instead of increasing the infection probability, we lower it to 5% in our last simulations. We find that the effects of decreasing the infection probability are less pronounced. The differences among the algorithms again seem to be smaller than under our benchmark scenario of 10%, possibly indicating a hyperbolic-like relation between the infection probability and the influence difference between the algorithms. This supposition is later corroborated for the case of the degree heuristics in Figure 5.6. Table 5.2 shows the difference in performance for all investigated infection probabilities. The values in Table 5.2 suggest that the difference between the algorithm is (locally) maximized somewhere between 5% and 20%.

5.2 Comprehensive Analysis of the Degree Heuristics

The low computational cost of the degree heuristics allows us to further explore the relationship of the difference in influence spread to the parameters of interest. To better grasp the effect of the number of seeds on the difference in performance between the two degree heuristics, we take the average influence of both heuristics over 1000 simulations for a range of 1 to 100 seeds. The result of these simulations is shown in Figure 5.5.

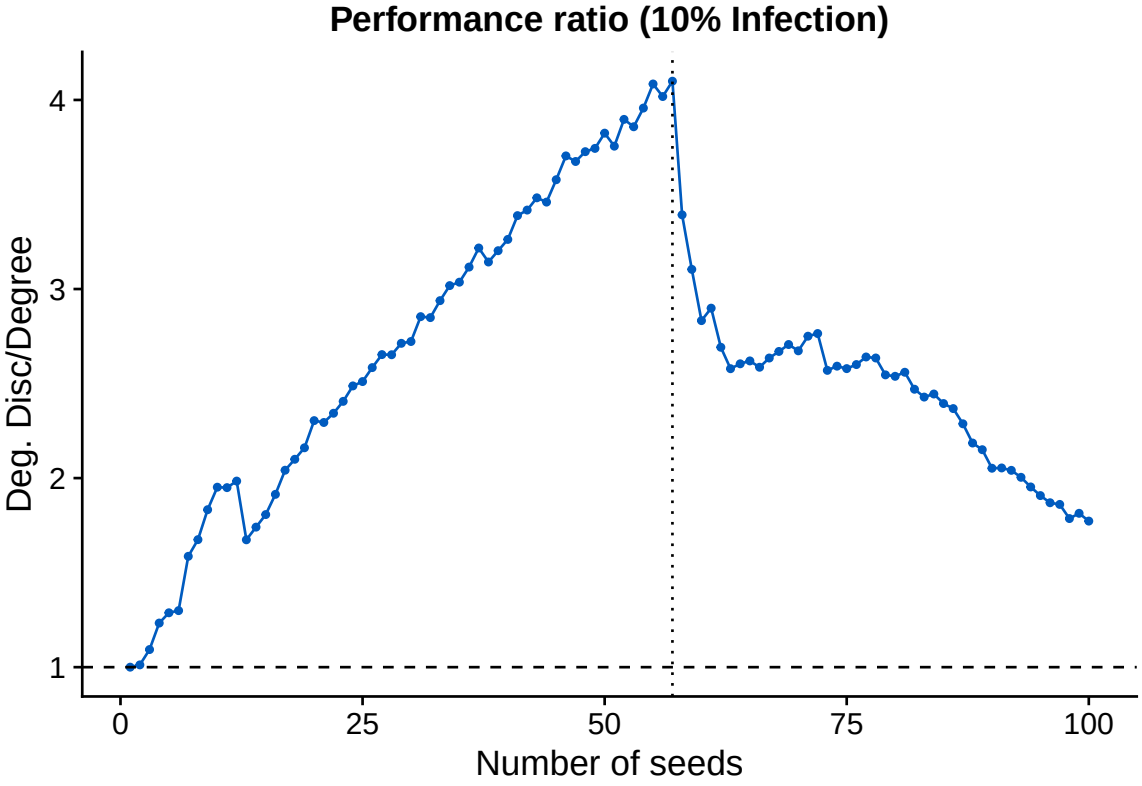


Figure 5.5: Heuristic performance ratio (Degree Discount/Degree): varying seed number

Contrary to what was suggested by the results in Table 5.1, Figure 5.5 indicates that there exists a non-monotonic relation between the chosen number of seeds and the difference between the two heuristics. The performance difference between the two degree heuristics shown in Figure 5.5 shows a remarkable analogy to Figure 5.3. The point at which the slopes of both relationships seem to switch sign occurs at exactly the same chosen number of seeds, namely between 57 and 58 seeds.

Similar to the analysis of the synthetic network in Section 3.2, we may explore the relationship between the infection probability and the difference in outcomes between the two heuristics. Figure 5.6 shows the resulting performance difference of 1000 simulations for a range of probabilities.

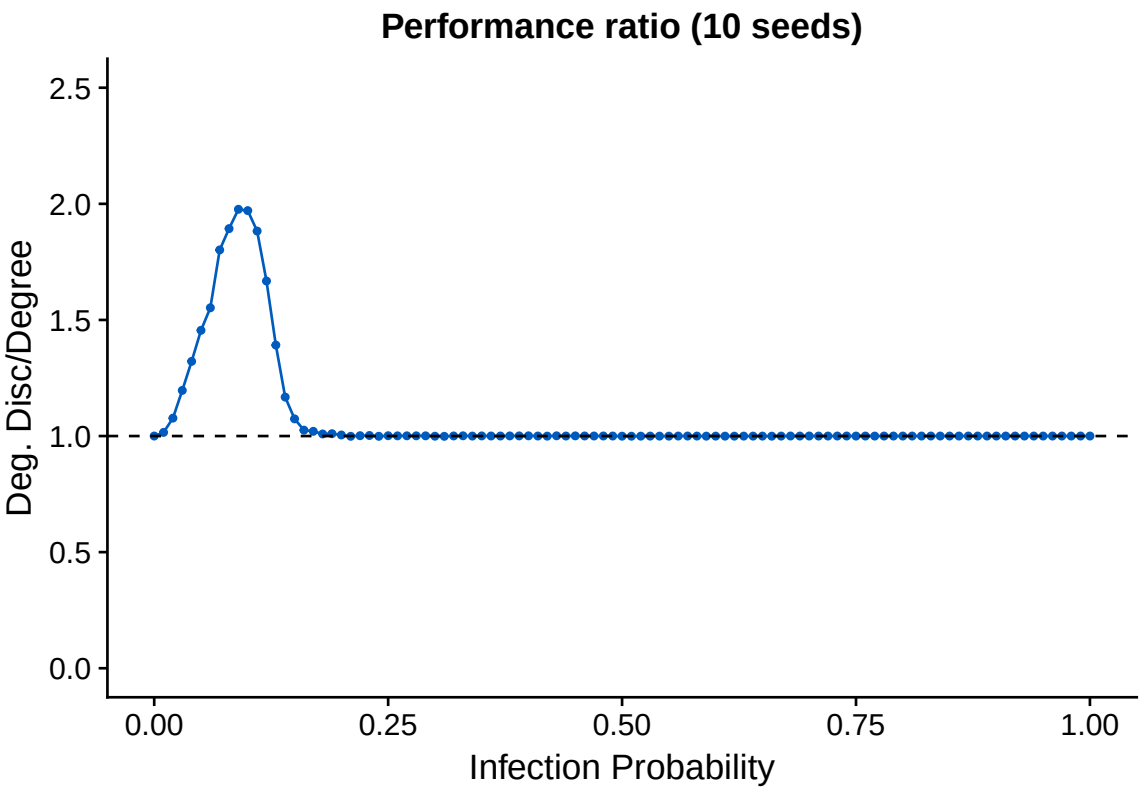


Figure 5.6: Heuristic performance ratio (Degree Discount/Degree): varying infection probability

While the relation between the infection probability and performance difference showed to be negative in our synthetic graph analysis (Figure 3.4), it seems to be more complex in the board member network (Figure 5.6). The infection probability for which the difference between the two heuristics appears to be maximized is somewhere in the region of a 10% infection probability. At this probability, the degree discount heuristic outperforms the highest-degree heuristic by a factor of two, given that we choose 10 seeds. This result is similar to the findings from Figure 5.1.

5.3 Decision Making Under Risk

Another factor in the tradeoff between the different algorithms is the amount of risk of that is associated with the algorithms. If we compare the standard deviations of the algorithm performances across the different infection probabilities using, we observe that the spread of the greedy algorithm performance is particularly high at 5% infection probability, at other infection probabilities (1%, 2%, 10%, 20%) the variance of the greedy algorithm performance does not greatly differ from that of the degree heuristics. To estimate whether the spread of the performance distributions has any significance in the optimal algorithm choice, we model the algorithm choice as a decision under risk. We use expected utility theory, as it is the reigning normative theory of decision making under risk. To compare the algorithms, we compute the relative risk aversion for which an individual is indifferent between choosing 10 seeds using either the greedy algorithm or the degree discount

heuristic at an infection probability of 5%. We assume a constant relative risk aversion (CRRA) utility function as described by equation 5.1.

$$U(x) = \frac{1}{1-r}x^{1-r} \tag{5.1}$$

Furthermore, we disregard the role of computational cost in the decision-making process and analyze the prospects of both algorithms by using the density functions shown in Figure 5.7 as probability distributions.

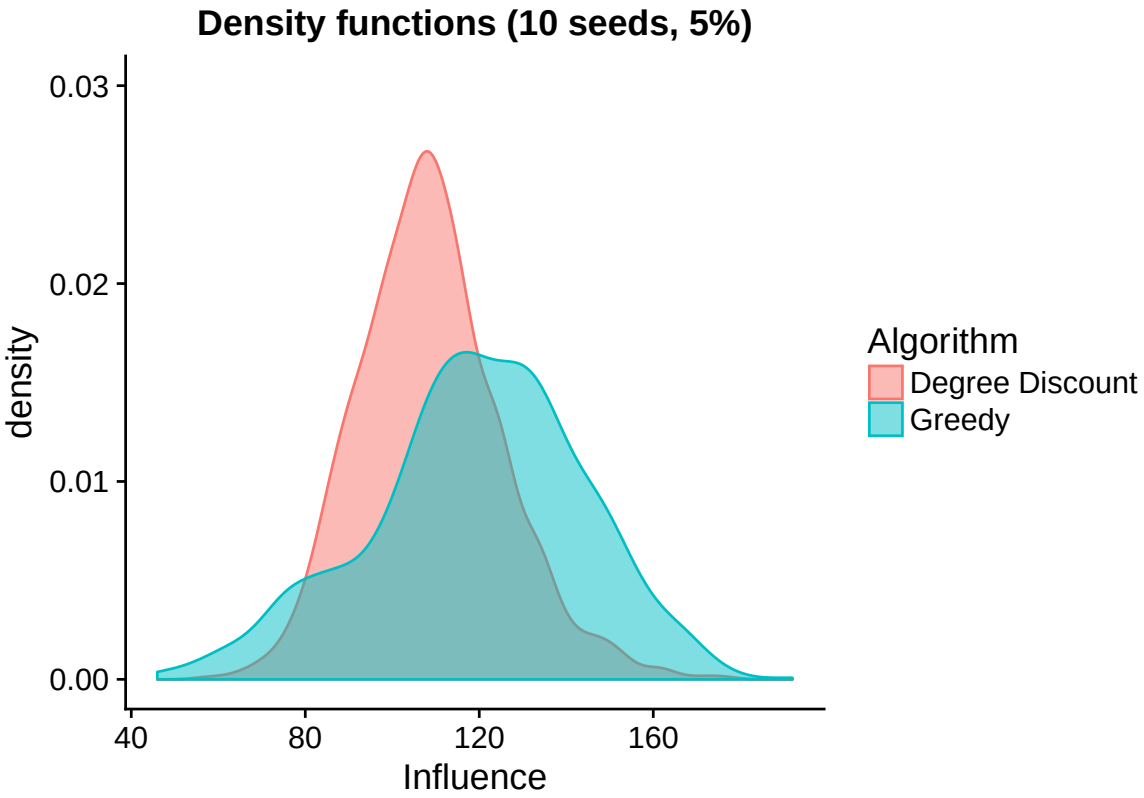


Figure 5.7: Density functions of Greedy and Degree Discount at 5 percent infection

Assuming CRRA utility we find that, for any value of relative risk aversion, expected utility theory states that the greedy algorithm is preferable to the degree discount heuristic in terms of influence spread. Figure 5.8 shows the absolute utility difference between the two algorithms following for values of relative risk aversion between zero and two.

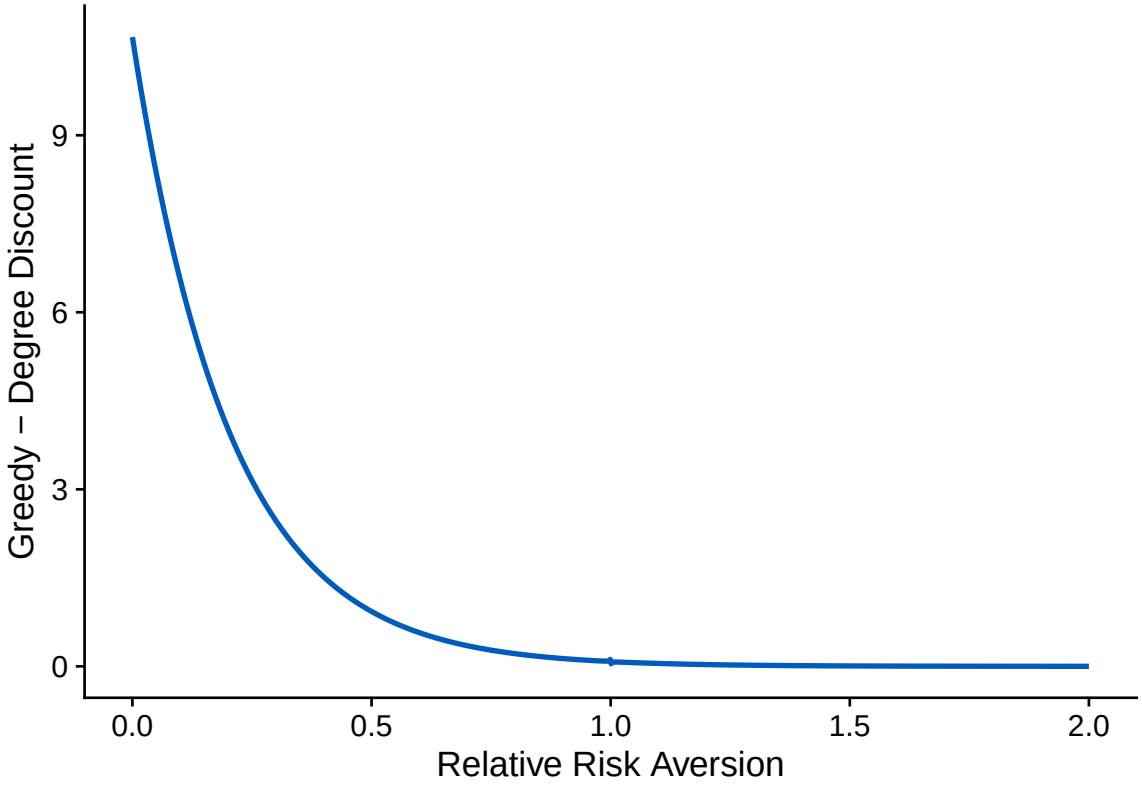


Figure 5.8: Expected utility difference between the greedy and degree discount methods

5.4 Conclusions

In this chapter, we compare R implementations of three methods that were reported in the literature for the influence maximization problem. We specifically address the influence maximization problem as finding influential individuals in the social network of interlocking directorates in Dutch healthcare markets. We find that there exists a tradeoff between computation costs and expected influence spread. Of the tested algorithms, the greedy algorithm established the greatest expected influence spread. However, as the greedy algorithm is computationally costly, one might prefer to use the degree discount heuristic, which achieves reasonable expected influence spread while having low running times. The difference in influence spread between the greedy algorithm and the degree discount heuristic seems to be decreasing as the number of seeds increases. Furthermore, compared to the benchmark infection probability of 10%, the difference between the greedy and degree discount methods seems to lower as infection probabilities get smaller, i.e. 5%. At 20% infection probability, simulations of the independent cascade model for influence predict there to be close to no difference between any of the tested algorithms.

In all cases, the simple highest-degree heuristic performs inferior to the other methods in terms of influence spread, therefore, we see no reason for the selection of influential individual has to be based merely on the number of links that individuals have. An in-depth analysis shows that the

difference between the highest-degree and degree discount heuristics appears to be increasing as the number of seeds increases up to a selection of 57 seeds. From 58 seeds onwards, the influence difference between the heuristics seems negatively related to the number of seeds. This relationship could be explained by overlapping seed selections of the two heuristics. Furthermore, we find that the expected difference between the degree heuristics is non-monotonic, concave, and peaks around 10% infection probability. There appears to be no difference between the degree algorithms if we observe infection probabilities greater than 20%.

Another difference between the degree discount heuristic and the greedy algorithm is the extent to which the resulting influence spread of the methods is volatile. We find that at 5% infection probability, the greedy algorithm exhibits a greater variance than the degree discount heuristic To test whether the spread of the influence spreads has any importance in the choice between the greedy algorithm and the degree discount heuristic, we have disregarded computational costs and used expected utility theory to test whether there exist values of risk aversion for which the degree discount heuristic is superior to the greedy algorithm. We find that there is no value of relative risk aversion for which the expected influence spread of the degree discount algorithm is preferable to the expected influence spread of the greedy algorithm.

6 Discussion

6.1 Applications

One key motivation for the topic of this thesis was the exploration of network analysis tools for the improvement of supervision and regulation in healthcare markets. This section sets out several applications that follow from the results and insights of the analysis in this thesis.

6.1.1 Distribution of Information

The first type of application is closely related to the setting that was discussed in Section 1.1. When regulators wish to obtain support among their stakeholders for a new policy, meetings may be organized to share information regarding the policy. Without the use of network analysis, it is possible that invitations to such meetings are distributed in a non-systematic and ad hoc fashion. The results in this thesis suggest that a list of people chosen by degree discount heuristic might serve as a computationally cheap and useful assistance to individuals who are responsible for the distribution of invitations. Further applications concerning the distribution of influence could include the spreading of information regarding the activities and findings of the authorities with respect to decent governance.

6.1.2 Detection of Information Flows

Another possible implementation of targeting influential nodes in a network relates to the detection of information flows. As it turns out, the same nodes should be targeted if instead of spreading influence, we aim to detect information flows (Leskovec et al., 2007). This phenomenon broadens the applicability of influence optimization using network analysis. Among the occupations of the supervisory department of the healthcare authority are affairs such as the detection of fraudulent behavior. An example could be the spread of information regarding loopholes in healthcare regulation that allow for undesirable behavior of directors. If regulators wish to be informed about this information, influence maximization algorithms could increase their chances of detecting such an information flow. Alternatively, regulators might use influence optimization techniques to identify problems with current policies, to get relevant feedback regarding their activities, and to learn about the healthcare sector in general. Other applications in this area could be unrelated to influence maximization but still very much related to social network analysis. If healthcare supervisors come to learn about mismanagement of a particular board or board member, social network analysis allows them to easily explore the position and neighboring nodes of this particular node in the network.

6.1.3 Network Visualization

In addition to static network visualization using the *igraph* package in R, tools such as the *Gephi* software package (Bastian, Heymann, & Jacomy, 2009) could be used for interactive visualization of the entire network or subnetworks. Figure 6.1 shows a basic Gephi visualization of a part of the bipartite network of boards. Gephi allows for intuitive network exploration by providing tools to filter and search for nodes, attributes, components, as well as many other network characteristics. Within Gephi, numerous layouts and aesthetic options can be specified.

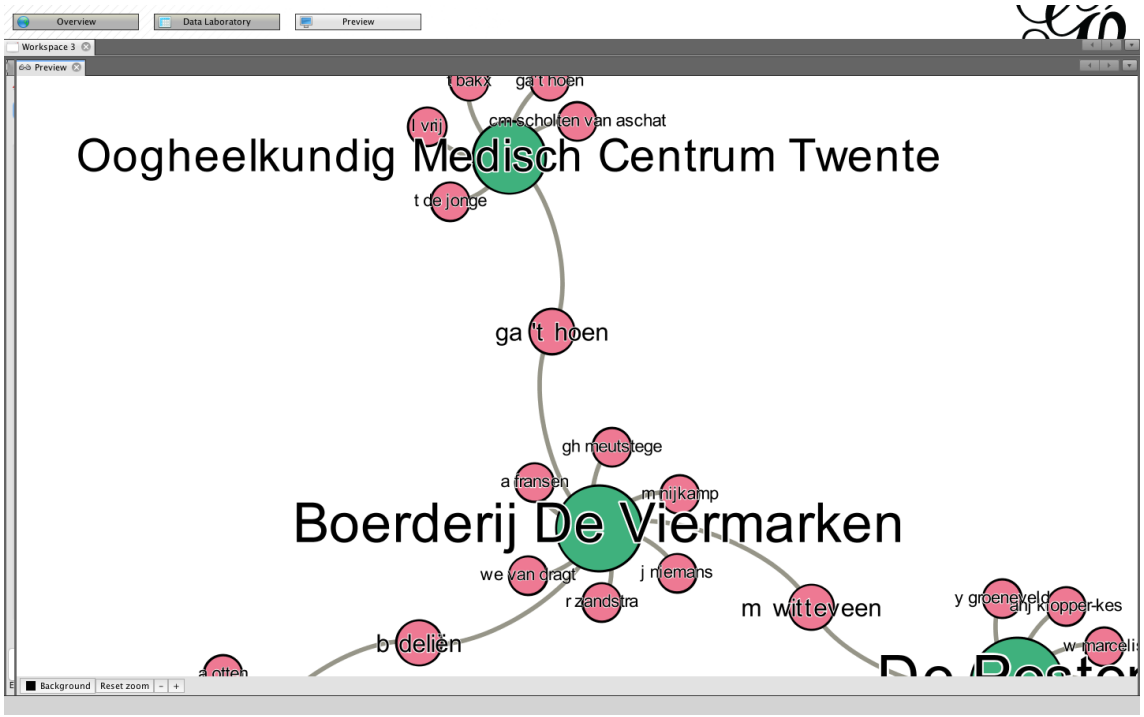


Figure 6.1: Gephi screenshot

6.2 Limitations and Further Research

For the results presented in this thesis to have any meaning in the actual practices of the regulators, the underlying model of information diffusion needs to be applicable to the real-life network of board members. Furthermore, the relations between the board members need to be captured to a sufficient extent by the links that are assumed in our constructed network.

The independent cascade model is a widely accepted model for information diffusion and is applauded for its simplicity. However, several more complex variations have been proposed to better model information spreading in social networks. The advantages of the independent cascade model include its ease of implementation, which in turn serves the computational cost. The main drawback of the independent cascade model is arguably the assumption of all links carrying the same infection probability. More realistic approaches for our board member network might account for the type of

link that is considered or the type of node that a link is connected to. For example, two executive board members may be more likely to share information than two supervisory board members due to a greater frequency of interaction. Similarly, information might be more effectively spread by chairmen of boards, rather than ordinary members. In order for future work to better capture actual relations between board members, one might wish to use arguably more realistic network data, from sources such as social media, i.e. LinkedIn.

As pointed out by Westra (2017), interlocks in Dutch healthcare governance have been increasing over time. However, we find no existing model for network generation that accurately accommodates the dynamics of networks such as the network of Dutch healthcare boards. Further research could help understand the network dynamics by developing a model which predicts degree distributions such as those in Figures 4.2 and 4.5 to arise.

Implementation of social network analysis in healthcare calls for accessible tools for network exploration and picking seeds. While *Gephi* provides a convenient tool for network visualization, no such software exists for the identification of influential nodes. Furthermore, current state-of-the-art approximation algorithms, i.e. IIM, TIM^+, $EaSyIM$, and SSA, could be implemented and compared in a similar manner as presented in this thesis. Alternatively, algorithms specifically designed to take into account the particular bipartite structure of board member networks could be designed.

6.3 Final Conclusion

In this thesis, we use network analysis techniques to disentangle the network of Dutch healthcare directorates. Specifically, we aim to locate the most influential individuals in this network. Existing literature regarding interlocking directorates in Dutch healthcare discusses and visualizes the network from the standpoint of organizations. Westra (2017) describes the network of interlocking directorates and raises competitive concerns regarding the occurrence and the development of interfirm relations. Influence maximization is an extensive topic in computer science literature, recent influence maximization algorithms and information diffusion models are ever more advanced as well as computationally efficient. This thesis aims to bridge the gap between the two well-established research topics of interlocking directorates and influence maximization.

We model the underlying social network of interlocking directorates by assuming individuals to be linked if they are seated on a board of the same company. We analyze multiple aspects of the constructed network. We find that the resulting degree distribution cannot be explained by either a random network generation model or a scale-free network model based on preferential attachment. Furthermore, we construct a model of interlocking directorates where firms are linked if there exists a shared member between them. Using a binomial logistic regression, we find that it is not generally

true that links appear most often between firms within the same sector. However, we do find that hospitals are more likely to interlock with other healthcare providers than other types of providers.

Using the independent cascade model of influence diffusion, we test several algorithms. We find that in any case, the improved greedy algorithm (Chen et al., 2009) achieves the greatest expected influence spread. Nonetheless, reasonable influence spread can be achieved by the degree discount heuristic. The degree discount heuristic might be the preferred method of selecting seeds as it is computationally cheap and reasonably effective. Computation of influential nodes using the greedy algorithm easily takes many hours and simple centrality heuristics are inferior as they exhibit a common flaw.

References

Arora, A., Galhotra, S., & Ranu, S. (2017). Debunking the myths of influence maximization: An in-depth benchmarking study. In *Proceedings of the 2017 acm international conference on management of data* (pp. 651–666).

Barabási, A.-L., & Pósfai, M. (2016). *Network science*. Cambridge university press.

Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi: An open source software for exploring and manipulating networks.* Retrieved from `http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154`

Brancheorganisaties Zorg (BoZ). (2017). Governancecode Zorg [Computer software manual]. Utrecht.

Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Chen, W., Wang, Y., & Yang, S. (2009). Efficient influence maximization in social networks. In *Proceedings of the 15th acm sigkdd international conference on knowledge discovery and data mining* (pp. 199–208).

Cheng, S., Shen, H., Huang, J., Zhang, G., & Cheng, X. (2013). Staticgreedy: solving the scalability-accuracy dilemma in influence maximization. In *Proceedings of the 22nd acm international conference on information & knowledge management* (pp. 509–518).

CIBG. (2018, Jun). *Jaarverantwoording zorg.* Ministerie van Volksgezondheid, Welzijn en Sport. Retrieved from `https://www.jaarverantwoordingzorg.nl/`

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems*, 1695. Retrieved from `http://igraph.org`

Domingos, P., & Richardson, M. (2001). Mining the network value of customers. In *Proceedings of the seventh acm sigkdd international conference on knowledge discovery and data mining* (pp. 57–66).

Galhotra, S., Arora, A., & Roy, S. (2016). Holistic influence maximization: Combining scalability and efficiency with opinion-aware models. In *Proceedings of the 2016 international conference on management of data* (pp. 743–758).

Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, *83*(6), 1420–1443.

Hang, Q., Zhu, J., Song, B., & Zhang, N. (2014). Game model of information transmission in social networks. *J. Chin. Comput. Syst*, *35*, 473–477.

Heemskerk, E., Hendriks, T., Wats, M., et al. (2010). Vormen de'old boys' een gevaar voor marktwerking in de zorg? *Goed Bestuur*, *5*.

Jeroen Bosch Ziekenhuis. (2015, September). *Profiel: Lid raad van bestuur.* Retrieved from `https://www.jeroenboschziekenhuis.nl/Website/Werken%20bij/Profiel_lid_RvB_2015.pdf`

Kempe, D., Kleinberg, J., & Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth acm sigkdd international conference on knowledge discovery and data mining* (pp. 137–146).

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., & Glance, N. (2007). Cost-effective outbreak detection in networks. In *Proceedings of the 13th acm sigkdd international conference on knowledge discovery and data mining* (pp. 420–429).

Li, M., Wang, X., Gao, K., & Zhang, S. (2017). A survey on information diffusion in online social networks: models and methods. *Information*, *8*(4), 118.

Nguyen, H. T., Thai, M. T., & Dinh, T. N. (2016). Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks. In *Proceedings of the 2016 international conference on management of data* (pp. 695–710).

Ohsaka, N., Akiba, T., Yoshida, Y., & Kawarabayashi, K.-i. (2014). Fast and accurate influence maximization on large networks with pruned monte-carlo simulations. In *Aaai* (pp. 138–144).

Papoulis, A., & Pillai, S. U. (2002). *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education.

Stokman, F. N., Van der Knoop, J., & Wasseur, F. W. (1988). Interlocks in the netherlands: stability and careers in the period 1960–1980. *Social Networks*, *10*(2), 183–208.

Tang, Y., Shi, Y., & Xiao, X. (2015). Influence maximization in near-linear time: A martingale approach. In *Proceedings of the 2015 acm sigmod international conference on management of data* (pp. 1539–1554).

Tang, Y., Xiao, X., & Shi, Y. (2014). Influence maximization: Near-optimal time complexity meets practical efficiency. In *Proceedings of the 2014 acm sigmod international conference on management of data* (pp. 75–86).

Westra, D. D. (2017). *Healthcare's competition conundrum: cooperative inter-organizational strategies in competitive healthcare markets* (Unpublished doctoral dissertation). Maastricht University.

Wikipedia contributors. (2018). *Np-hardness — Wikipedia, the free encyclopedia.* Retrieved from `https://en.wikipedia.org/w/index.php?title=NP-hardness&oldid=838074652` ([Online; accessed 1-August-2018])

Yule, G. U., et al. (1925). Ii.—a mathematical theory of evolution, based on the conclusions of dr. jc willis, fr s. *Phil. Trans. R. Soc. Lond. B*, *213*(402-410), 21–87.

Zhang, X., Zhu, J., Wang, Q., & Zhao, H. (2013). Identifying influential nodes in complex networks with community structure. *Knowledge-Based Systems*, *42*, 74–84.

Appendices

A Alternative Algorithms

In this appendix, we introduce several additional heuristics that may be employed to solve the influence maximization problem.

A.1 Betweenness Centrality

In addition to the highest-degree heuristic, we may base simple heuristics on many other network characteristics. One such characteristic is betweenness centrality. The betweenness centrality value of a node is defined as the number of shortest paths that pass through the node. Calculating betweenness centrality is computationally more costly than the calculation of degree centrality as a great number of paths has to be considered. Here, we analyze the *highest-betweenness heuristic*, which chooses the k nodes that have the greatest betweenness centrality value. We choose 10 nodes using the highest-betweenness heuristic in addition to the methods used in Section 5.1. Following, we compare the algorithms based on simulations of the independent cascade model with 10% infection probability. Figure A.1 summarizes the results of the simulations.

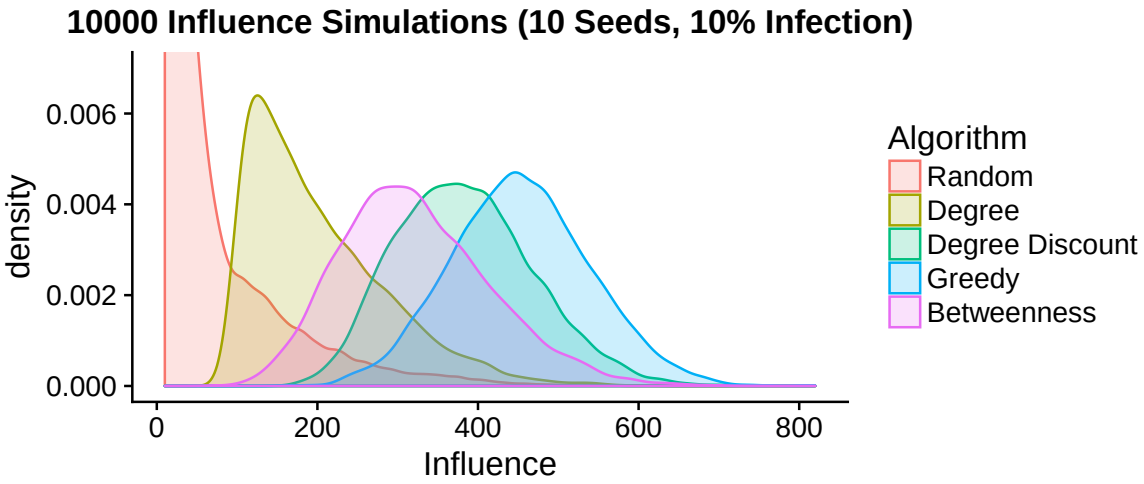


Figure A.1: Performance of the highest-betweenness heuristic

We conclude from Figure A.1 that the highest-betweenness heuristic might be promising as a simple centrality based heuristic. The highest-betweenness heuristic significantly outperforms the highest-degree heuristic in terms of influence spread. However, since betweenness calculations have some computational costs, the degree discount algorithm is likely superior to the highest-betweenness heuristic on all aspects. A possible explanation for the superiority of the degree discount heuristic is that the degree discount heuristic generally does not choosing neighboring nodes as seeds. We can implement a similar mechanism for the highest-betweenness heuristic. We develop the *discounted

betweenness heuristic*, which does not choose nodes with a high betweenness if neighboring nodes are already selected. Again, we simulate the independent cascade model and analyze the resulting performance in Figure A.2.

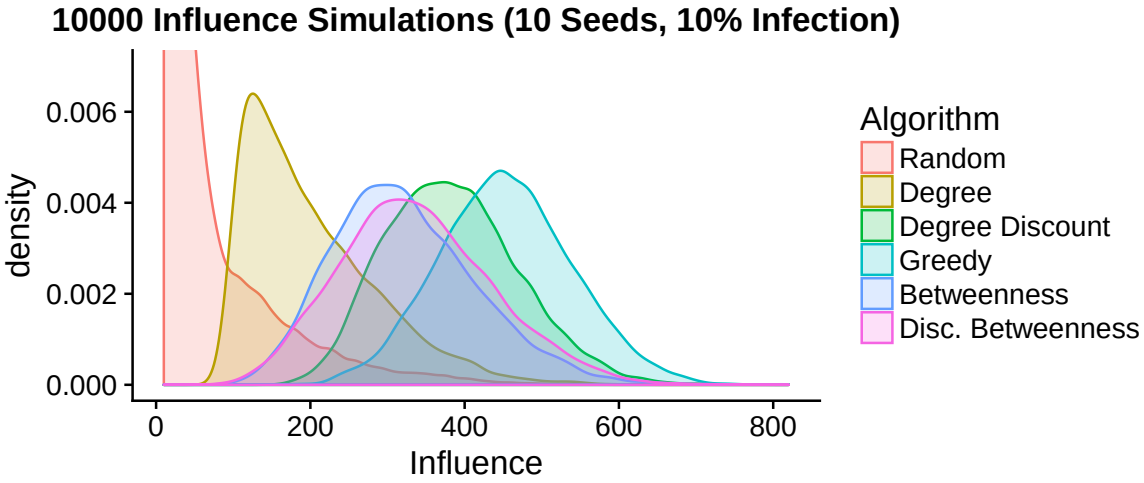


Figure A.2: Performance of the discounted betweenness heuristic

From Figure A.2 we conclude that the benefit of introducing a discounted betweenness heuristic is very small and arguably negligible. Therefore, the degree discount heuristic remains the preferred heuristic to select influential seeds at low computational cost. If computational costs are of no importance, the greedy algorithm is the preferred method.

A.2 Firm-Based Highest-Degree

In certain cases, it is apparent that simply choosing nodes which have the highest degree is unwise. For example, if there exists one firm in the network with an extraordinarily great number of board members, it is likely that the highest-degree heuristic will only choose seeds which are board members of that particular firm. To a certain extent, this flaw is similar to the flaw that the degree discount heuristic aims to correct, namely the flaw of choosing nodes that are close to each other in the network. Here, we attempt to exploit a distinctive feature of our board member network. As our board member network originated from the bipartite network of interlocking directorates, links between nodes in the network carry information regarding the firm of which the connected nodes are a board member. We develop a novel heuristic using this information in the following manner. The *firm-based degree heuristic* starts by choosing the node that has the greatest degree, as all degree heuristics do. Upon choosing the first seed, it creates a list containing all firms of which the chosen node is a board member. To choose a second seed, it attempts to choose the node which has the second-highest degree, however, this node is only selected if it is not a board member of any of the firms in the newly created list of firms. If the node is selected, the firms of which the chosen node is a board member are added to the list. This process continues until k nodes are selected.

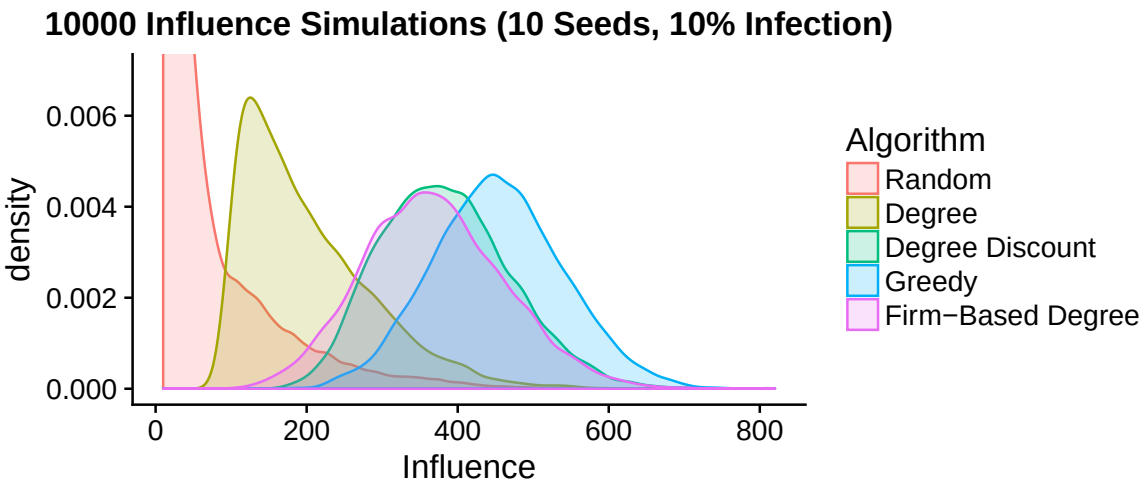


Figure A.3: Performance of the firm-based degree heuristic

The results of the influence diffusion simulations, as shown in Figure A.3, show that our firm-based degree heuristic achieves influence spreads similar to those of the degree discount heuristic. However, in its current form, running times of the firm-based degree heuristic are 20 times longer on average. We therefore conclude that, of all tested algorithms, the degree discount heuristic remains the best seed-selection method for the network of board members in Dutch healthcare.

B A Geographic Visualization of the Firm Network

This appendix shows the firm network plotted on a map of the Netherlands in Figure B.1. We conclude that such a representation on its own does not enable us to make inferences regarding the relationships between geographical distance or location and the likeliness of links to form. Therefore, we recommend for further research to study these relationships. Nevertheless, this visualization may provide powerful insights if explored in greater detail using interactive tools such as Gephi.

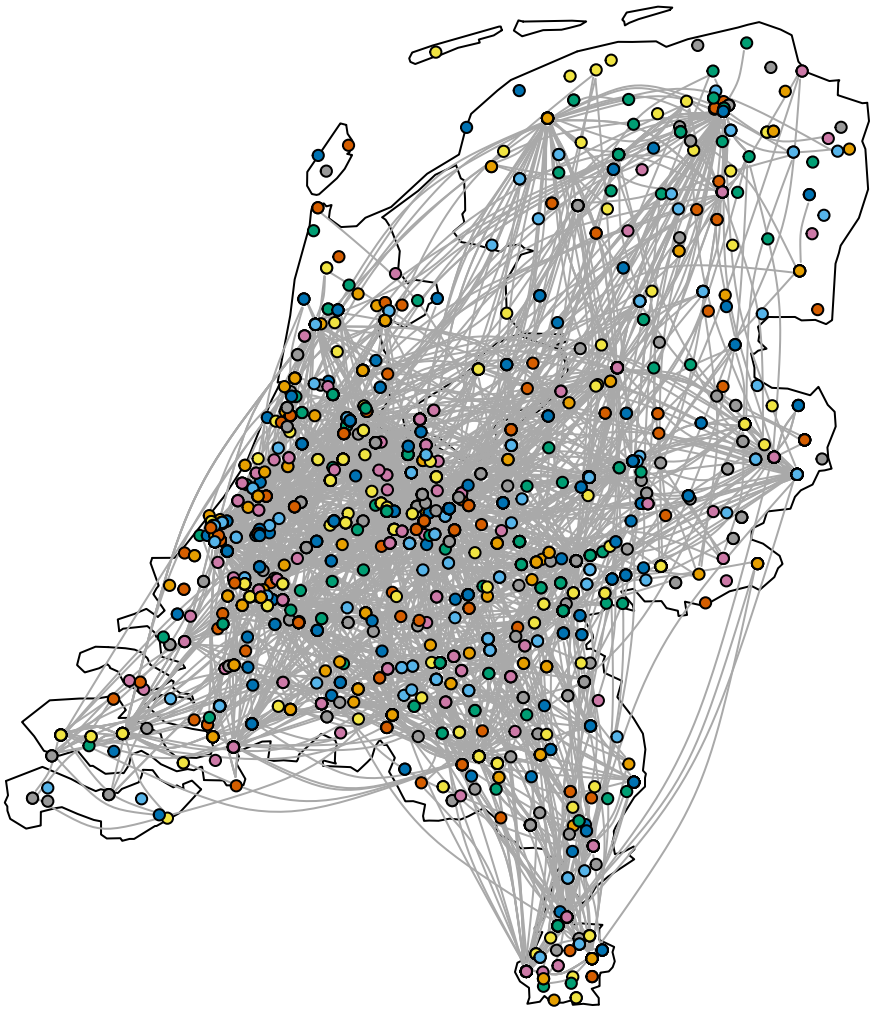


Figure B.1: Firm-network map of the Netherlands