University of Kaiserslautern

Master Thesis

Organizational Social Network Analysis

Author:
Wolfgang Schlauch

Supervisor:
Prof. Dr. Andreas Dengel
Darko Obradovic

May 7, 2012

Contents

List of Figures

List of Tables

List of Algorithms

Abstract

For this thesis the relations of employees of a company were examined. For this purpose, an online survey was created which is described in Chapter 3 . This is followed by a detailed analysis of the results. We used the Louvain algorithm to discover if people connected to their groups well or if there were inner group conflicts resulting in splits. Furthermore we used betweenness centrality to highlight the persons who can be asked if one needs guidance to the knowledge of other persons. In general, a positive picture of the company emerges because the groups are closely interconnected. Hazards are the few connections that are between the groups since shared knowledge might increase the results or the creativity.

Zusammenfassung

Im Verlaufe dieser Arbeit wurde eine Soziale Netzwerk Analyse innerhalb einer Forschungsabteilung durchgeführt. Hierfür wurde zuerst eine Umfrage mit Hilfe der Betreuer und der Firma erstellt, die sich auf die Mitarbeiter konzentriert und erfasst wer mit wem arbeitet. Zudem sollte genannt werden wie die Befragten ihre Beziehung zu ihren Kollegen selbst einschätzen. Aus diesen Daten wurde dann ein Netzwerk konstruiert, welches innerhalb der Forschungsgruppen der Abteilung einen guten Zusammenhalt zeigt während die Beziehungen zu den anderen Gruppen noch ausbaufähig sind. Auf dieses Netzwerk wurden dann Algorithmen angewandt um wichtige Mitarbeiter zu kennzeichnen die eine zentrale Position innerhalb der gesamten Abteilung innehaben sowie Aufspaltungen innerhalb einer Gruppe zu erkennen. Die Resultate dieser Bewertungen sind analog zu dem Gesamtnetzwerk gut, man erkennt einzelne Forschungsgruppen innerhalb der Forschungsabteilung ohne größere Abweichungen festzustellen.

1. Introduction

1.1. Objective

The objective of this thesis is to capture the knowledge and communicaton network of a scientific research company. We will gather information about the working relations. Afterwards we will analyze the relations to see if the connections are sufficient and if the company has obvious deficits. Since we will look at two different research groups we will also find out if they are connected well enough.

1.2. Introduction

Today online activities like watching youtube[1] videos, posting on twitter[2] or playing on facebook[3] are usual for the main street. In America there are at least 51% of the nation on facebook in the age between 12 and 50 according to a study by Edison Research (Arbitron/Edison Research, 2011) and according to a study conducted by the Bank of America 96% of the people who have participated in a survey were using facebook (Bank of America Merrill Lynch Global Research, 2011). This leads to increased interest in social network analysis (SNA) to find specific relations between the persons on those networks. Even companies have recognized that social networking is not only a byproduct of their work but that it can yield important information, better results and even new ideas if the team gets assembled based on their social network. Thus, it is not surprising that social media and social networking have become more important to everyone. One out of three employers rejected an applicant due to something found about them online and the number might increase since

[1] http://www.youtube.com
[2] http://www.twitter.com
[3] http://www.facebook.com

the percentage of HR professionals using social networking sites for sourcing potential candidates increases (CareerEnlightment, 2011).

This interest is not limited to the process of hiring. The process of performance reviews is still common but according to Culbert (2008) this process is obsolete and should be changed to performance previews, involving direct conversation between boss and employee. R. Cross and Parker (2004) provided even more examples. In their book the authors show several companies and the lack of social ties between subgroups of them. Without social network analysis they would not have seen the problem and thus could not have worked against that.

Therefore we gathered the social network of some of the groups in one branch office of a company. We did not stop there but gathered even more information than only the connections between the persons but also what they think about there co-workers on several levels such that we were able to call it an informal knowledge network, containing not only relations but also how well the employees know what their co-workers are capable of and their feeling about the ease of access to their knowledge. We analyzed the network to get key-persons in different views, expecting to see changes in the importance of some of the persons involved in the analysis due to their position in the company.

1.3. History of Social Network Analysis

Social Network Analysis as a term coined in the 1950s, even if the idea of a social network existed long before. J. A. Barnes started using the term “social network” to denote patterns in networks he analysed such as groups or clustering behaviour. That the analysis of groups and networks was done long before is well documented. Precursors of this theory were Émile Durkheim and Ferdinand Tönnies. Tönnis was more interested in the existence of social groups and the links between individuals whereas Durkheim was more interested in patterns in the society. He also made the difference between “mechanical solidarity” – in which individual differences are minimized – and “organic solidarity” – the development based on cooperation of different individuals. For a rather long time most studies seemed to be on small loose networks but in the 1930s J. L. Moreno started recording and analyzing the social interactions in small groups, while in Harvard W. Lloyd Warner and Elton Mayo started to explore interpersonal relations at work. Warner and Mayo were just observing people and discovered something they called "informal organization" of an organization, a social structure not related to the organigram but contributing much to the productivity of the employees (Scott, 2000; Wikipedia, 2011b; Chlldress, 2011). From there on many theories were developed like the "six degrees of separation", which states that in a sufficiently large network everyone knows everyone else by six degrees, or to re-formulate that in average everyone can be reached without going further than six degrees. There were experiments conducted by Stanley Milgram, who used chain letters to show that America is a "small world", showing with the experiment that he could send a letter to several random persons with the aim to get them to a colleague and they would arrive with approximately three friendships links on average. The experiment was performed by Duncan Watts in 2001 with e-mail chain letters who had a resulting average of six links. Nevertheless this experiment was not related to the increasing number of spam mails today. Also for instant messaging this was done in 2007 by Jure Leskovec and Eric Horvitz with a similar result of an average path length of 6.6 (Wikipedia, 2011a).

1.4. Related Work

Of course big networks like Facebook, Twitter and LinkedIn[4] are very interesting. But the networks of Warner and Mayo are more important in the context of this thesis so we will take a closer look at them. Usually every company has its own organization mapped as an organigram/organizational chart. They are expressive enough to convey the structure but they do not help in understanding the social network of a company. In a company a social network may hold cultural values or social values as it is capable of providing satisfaction by backslapping of one's peers or birthday cake. Those aspects are important for the structure and are basically the glue that holds a company together. We will investigate the knowledge network of a company. They differ from a social network as it is not based on the same things a social network is. A knowledge network tries to capture who works with whom for what reasons. Basically the knowledge of other people and their job description is an important part, since in a standard social network those facts are not necessarily known. We can call it a communication network since we only get to know the others abilities by

[4]http://www.linkedin.com

interacting or talking. For a good work relationship also trust is an important component such that some researchers tend to take this into account (Cross, Parker, Prusak, & Borgatti, 2001). Other researchers take more into account who talks to whom for what reasons. Cross, Borgatti and Parker asked to whom an employee would go for work related tasks and to whom for questions concerning not work related topics. The results were quite different (R. Cross, Borgatti, & Parker, 2002). In the same paper a manager was mentioned who recognized that something was wrong in his company and the analysts found with a basic question, who works with whom, the problem. There was no real interaction between the "soft skilled" and the "technical skilled" part of the team. With this very basic knowledge network the company then was able to restructure and improve the productivity and also the knowledge of the employees in both topics (see Figure 1.1 for the network and it evolved version).

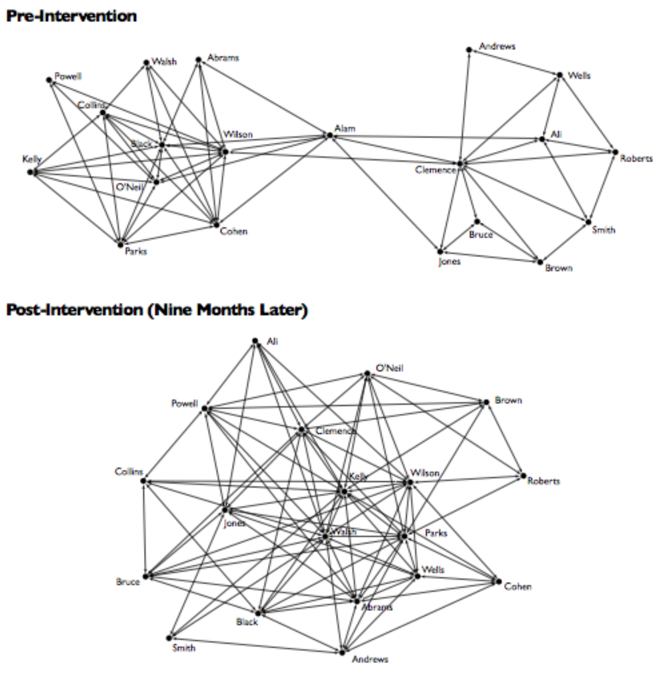


Figure 1.1.: The network before the intervention triggered by the investigation of Cross, Borgatti and Parker and afterwards

Another important difference between a social network and a knowledge network is the reflection where the company might lack skill. A very good example we already mentioned is given in the work of Cross, Borgatti and Parker (R. Cross et al., 2002). There it was a lack of communication but there can be also other important skills missing. Cross et al. (2001) showed different informal networks due to different levels of research. While the standard social network of the company researched was pretty dense, it got more sparse the more specific the questions were. As one might observe the ability to get in contact with the surroundings changed the dynamics of the network from dense to sparse and left several nodes with less or no connection to the group.

Research like this is used to increase the performance of a company or to discover where the problems are in working groups. In the context of this thesis we will retrieve the informal network of a company and analyse it for them by the means of graph theory.

In the next chapter we will explain the basic mathematical theory to understand what the rest of the thesis will be about. In the following chapter we will explain how we get the social network and how we did the questionnaire, resulting in the next part where we are going deeper into the observed knowledge network. We will close this thesis with a short overview about the things we did and give some ideas for further research on this topic.

2. Introduction to Graph Theory

Companies are usually based on a formal scheme called organizational chart. In some variants of those a clear hierarchic structure is visible in which the head of the organization is placed on top of all managers which are placed over their subsections (see Figure 2.1). The schemes not only show how the companies think they should be organized but also the way in which reports, news, updates and almost everything else should flow. And for some of these it might be okay to have this chart, for example for news from the head of the company who gives it to the managers who redistribute it among their groups. But if information or a question on a certain topic would always go from one subgroup to the managers level and from there down to the asked subgroup again and the complete way back the task would probably be outdated or invalid since employees can not waste their time with waiting for an answer but have to work on the problem. With that in mind one will acknowledge that there are other relations between employees than those defined in the organizational chart. And those are only the obvious relations a manager has to see, there are also many other relationships as social researchers remark.

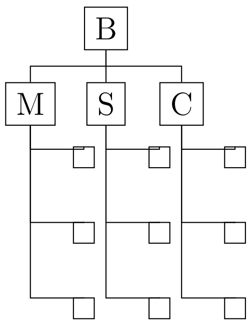


Figure 2.1.: Example of an organigram

2.1. Knowledge Networks

Social networks are today very common to have. If we take a look in the online community we see social networks everywhere, be it either connections by linkage (Obradovic, Rueger, & Dengel, 2011) or be it in a community like Facebook[1], the networks are everywhere. There are some things knowledge networks share with them. Social networks as well as knowledge networks are representable as graphs. But in a social network the main interest is how the people are connected, who knows whom, who is related to whom and so on whereas in a knowledge network different factors are more important. The factor of familiarity is still in the network, but other questions emerge concerning the working relationship like whom to ask for help, who is often available, who can explain a topic very well or help to get a different view on a problem. Every question asked in the research of a company may give the network a different shape and combined networks might look very different from the networks where only one question was asked (for comparison see Cross et al. (2001)). But how are networks displayed the best?

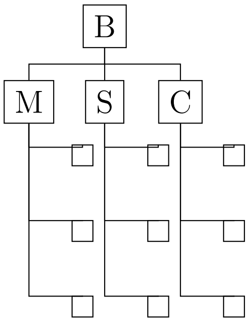
One of the easiest ways to recognize who knows whom and how they are connected is displaying the relationships and the persons graphically. Since it is practically impossible to observe this with pictures of persons and some complicated markers we will use graphs for that. People will be displayed as nodes and the relations between people will be the edges connecting two node. For an example we used the organigram from above and transformed it into a graph in Figure 2.2. To be able to use graphs we will introduce them formally in the following.

In the next section we will introduce some of the basic concepts of graph theory, followed by several algorithms. For the algorithms it is not necessary to understand the definitions from graph theory completely.

2.2. Basic Graph Theory

Definition (Undirected Graph) (Nebel, 2010) Let $G = (V, E)$ be a graph. We call a graph undirected if the edge connecting two vertices $v_i, v_j \in V$ have

[1]http://www.facebook.com



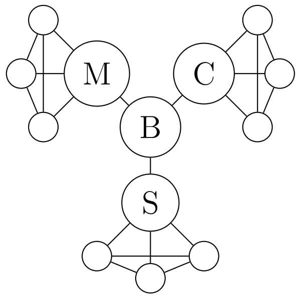


Figure 2.2.: (a) shows again the organigram, (b) shows the same structure as a graph

no orientation. This means that there is no difference between the edges (a, b) and (b, a), i.e., they are not ordered pairs but sets $\{v_i, v_j\}$.

With this definition it is possible that a node is connected to itself what will not harm and is for other examples not unusual. For the matter of a knowledge based social network it is not very useful to have those loops since they would not yield any information but the persons know them selves, consider themselves as trustworthy or else. So we introduce the notion of the simple graph as the simplest kind of network.

Definition (Simple Graph) (Nebel, 2010) Let $G = (V, E)$ be a graph. We call G a simple graph if it is undirected, contains no loops and if two vertices are connected only by one edge. This means that there are no edges $\{v_i, v_i\}$ and if there is one edge $\{v_i, v_j\}, i \neq j$ then removing this edge will disconnect the two vertices.

Up to this point we introduced the very basic concepts needed to create a network. As one can see it already is useful to map an organigram to a network (see Figure 2.2).

The next useful concept for the idea of a knowledge network is to introduce directions to the network. This is based on the fact that one might ask one person for their help but this person would never ask the other person, be it either for reasons of rank or lack of the needed knowledge or something else.

Definition (Directed Graph) (Nebel, 2010) Let $G = (V, E)$. We call G a directed graph if the edges connecting two node have an orientation. This is different from the undirected graphs since the edge $(v_i, v_j), i \neq j$ could be walked in both directions in them. In a directed graph we could on this certain edge only walk from v_i to v_j but not the other way around. For that we need another edge (v_j, v_i). This would then be a symmetric edge.

Remark Even if it is not defined as such we can speak of a simple directed graph, if the graph does not contain more than one edge between two nodes with the same direction.

Now we have the basic concepts to handle the basic case of a social network. The base case implies that it is only capable of handling one relation otherwise we would have several edges between two vertices. The general knowledge based on examples like Facebook implies that "add as friend" does not necessarily imply that one knows the befriended person on a personal basis or at any at all. This is a well recognized problem in studies concerning online social networks (Zain, 2011). For more complex research questions like the examples mentioned in Section 2.1 it is more complex to predict the result. If we want to have only one of the aspects in a network we can generate for each research task a different network. But if we want to get a glimpse on how the different research tasks interact then we have to put all the edges in one graph representation. For that there is the definition of the multigraph.

Definition (Directed Multigraph) Let $G = (V, E)$. We call G a multigraph if between two vertices $v_i, v_j, i \neq j$ there is more than one edge with the same orientation. The edges in a multigraph are sometimes also called links but we will not use this term in the context of this thesis.

Remark We will not use the possibility to make loops, indicating some self-references, since they would not give any useful information for a knowledge network.

We are now able to map a companies internal structure into a graphical representation. We mentioned before that this is useful for understanding concepts more easily. Another reason for doing so is the ability to change some of the edges in the graph to see if certain structures are common to a network like this or to discover dense connected subgroups. We now have introduced several different types of graphs. The most important to us will be the undirected and directed simple graph and the directed multi-graph.

2.3. Algorithms

After explaining the theoretical foundation of the concepts used in this work we will go ahead and name and explain the algorithms used. This will be followed by the real world part of the thesis where we dive into the analysis of the network itself.

For the explanation of the algorithms we will introduce some additional concepts such that it is more clear what the algorithm is supposed to return.

Before any algorithm will run we can start with looking at the data to get an initial understanding of our network. Nodes with a low connectivity and nodes with a high connectivity will be easily recognised. One of the assumptions in social network analysis is that people with a high connectivity are very important, so called connectors or brokers since they are able to direct a question to the correct person. People with low connections are assumed to be either new to the network such that they have few connections or the are self-sufficient (Ingelbrecht, Patrick, & Foong, 2010).

2.3.1. Floyd-Warshall Algorithm

Everyone who ever had something to do with any kind of network will know Dijkstra's algorithm by heart. Nevertheless we decided to go with the Floyd-Warshall algorithm since it gives us not only the best possible path but the distances from all nodes to each other. Furthermore, it works with negative edge weights what might be useful for the future (see Algorithm 1).

We create a matrix for the graph of size $N \times N$ where N is the number of nodes in the graph. The entries denote in the initial version the weights of the edges that exist and infinity if there is no edge between two nodes. Afterwards we go through the network and try to find shorter paths than the one given to connect the nodes. For example in 2.2 the nodes C and S would have no initial connection, thus an entry equal to infinity. There is a path between them via B such that the shortest path between them is the sum of the edge weights on the edges between M and B and B and S.

In this thesis we will handle only small networks with less than 100 nodes so we do not have to worry about the run time of the algorithm. However, if we should extend this research to a greater community the run time of the Floyd-Warshall algorithm has to be considered. This algorithm has a time complexity of $\Theta\left(N^3\right)$ which might be worse than the complexity of some other algorithms.

Algorithm 1 Floyd-Warshall Algorithm

Require: $G = (V, E)$
Require: N {number of nodes in G}
  for $i = 1 \to N$ **do**
    for $j = 1 \to N$ **do**
      if $\{i, j\} \in E$ **then**
        $d[0][i][j] :=$ weight of edge
      else
        $d[0][i][j] := \infty$
      end if
    end for
  end for
  for $k = 1 \to N$ **do**
    for $i = 1 \to N$ **do**
      for $j = 1 to N$ **do**
        $d[k][i][j] = \min\left(d[k-1][i][j], d[k-1][i][k] + d[k-1][k][j]\right)$
      end for
    end for
  end for
  return d;

Dijkstra's algorithm for example has time complexity $\mathcal{O}\left(\|E\| + \|V\|^2\right)$ which is most of the time reduced to $\mathcal{O}\left(\|V\|^2\right)$ due to the fact that the squared number of nodes should be much higher than the number of edges. Nevertheless, we decided against Dijkstra due to the fact that we want to get the distances from all nodes with possible negative edge weights. Dijkstra's algorithm is not capable to handle those correctly opposite to Floyd-Warshall's algorithm. If we assume we only have positive edge weights we derive a time complexity of $\mathcal{O}\left(\|V\|^3\right)$ using Dijkstra's algorithm if we want to get the shortest path from each node to every other node in the network, based on the assumption that the network is directed and not symmetric.

As stated above we will use this algorithm to get the shortest pathes from the nodes. We can use this knowledge to calculate the average path length by which we can judge if information flow is hindered or too slow in a certain path. We will also get the longest and the shortest path with this method such that we can see who appears to be the most central person in the network as also the most peripheral. With the distance matrix, the result of the Floyd-Warshall algorithm, we can also calculate the closeness centrality of a node, an indicator of how fast the person can reach each other person in the network (M. Newman, 2005). It is also possible to calculate the betweenness centrality that indicates how much the person is able to influence communication between other persons (Brandes, 2001; M. E. J. Newman, 2010). For the former it is enough to sum up all distances a node has to other nodes. The smaller the sum, the more central it is. For the latter it is counted on how many shortest paths the nodes in question is. This gets divided by the number of possible pairs of nodes in the network. The higher the resulting value the more between is the node. Those values are important for a network since they can also show who is too frequented and whose removal would damage the network (in the real world the company). For nodes that are too between it is useful to create new connections that can compensate that damage. As one can see in Figure 1.1 there was a node with high betweenness (Alam) in the state before the intervention. He was a bottleneck and thus a great danger for the company if he would have misguided any of the teams he was connected to.

From the centrality it is not very far to get to Google's PageRank algorithm since it is a variant of something called Eigenvector-centrality.

2.3.2. PageRank-Algorithm

Here we will describe how the PageRank algorithm works and we will also give one important variant of this algorithm.

The purpose of the PageRank algorithm was to estimate how popular a website is. This was done by finding out how many sites linked to it. The more sites linked to a specific one, the higher its influence on the whole network since it guides visitors to the sites linked on it. So a site linked on a highly linked site is assigned a higher probability to get visited. This is expressed with a simple equation (see equation 2.1). The damping factor d is set to a value between 0 and 1, usually $d = 0.85$. It refers to a random surfer, ending up on a page without links. He would have to jump to another random page to continue surfing. The value C_j refers to how many sites are linked, PR_j is the pagerank of the other site. PR_i would then be the probability that a random surfer will visit the page (Brin & Page, 1998).

$$PR_i = \frac{1-d}{N} + d \cdot \sum_{\forall j \in \{(j,i)\}} \frac{PR_j}{C_j} \tag{2.1}$$

The idea to use this in the context of a social network is valid. It is used for example to recognize leadership (see Pedroche (2010), Pedroche, Moreno, González, and Valencia (2011)) or simply key users (Heidemann, Klier, & Probst, 2010). We will consider that we have weighted edges and thus use not only PageRank, which uses unweighted edges, but also an enhanced version that uses this weights changing the equation (see equation 2.1 and equation 2.2 for comaprison).

$$PR_i = \frac{1-d}{N} + d \cdot \sum_{\forall j \in \{(j,i)\}} w_{ji} \frac{PR_j}{C_j} \tag{2.2}$$

In this equation C_j is not the number how many sites are linked but the sum of the weighted out-degree of the nodes linked to. w_{ij} is the weight of the edge between the two nodes considered (Wiggins, McQuaid, & Adamic, 2006). It is also stated that the PageRank-equation can be used do derive a similar formula that is able to calculate the SocialRank (Türling, 2007), a value that determines the influence of a person to a social network.

We will use both variants of this algorithm to discover not only who is regarded highly as a important person (first variant) but also in terms of the question who is good to be known. This can be found if the weights acknowledge the person an amount of knowledge or reach or power in the network.

2.3.3. Louvain Algorithm

Usually in a social network like Facebook groups are acting on common social interests or on social bindings like relationship easily discoverable. With organizations it is the same due to the projects persons are working on. We want to know if the informal network represents the paradigm required by the organization. We do this with the Louvain-Algorithm (compare Algorithm 2) developed by Blondel, Guillaume, Lambiotte, and Lefebvre (2008). They describe a method that should be capable to find communities in large networks by maximizing the modularity of the network. For that it assigns in the first iteration each node an own community it belongs to and then looks repeatedly if there are communities whose modularities would increase if a node would be added. If there is no such community it stays in the old. As soon as there is no community that can increase its modularity the process stops. We reimplemented the algorithm presented by Blondel et al. (2008) and compared the results to those of the original version on some artificial networks. The results are the same up to naming such that we can go on and argue about the problem with a community finding algorithm.

The problem with a community finding algorithm is that it is not guaranteed to succeed, in explicit it does not necessarily derive the best possible communities from a given structure. It is not claimed that the algorithm is correct but Blondel et al. (2008) wrote they compared their data with true networks with known communities and found good results. We took the examples provided by their working group and applied both versions on them.

Table 2.1.: Comparison Louvain method and reimplementation

Network Name	#nodes	$\#c_{\text{Louvain}}$	$\#c_{\text{our version}}$	matching
Zachary's Karate club	34	4	4	100%
arXiv citation	9377	59	59	61.03%

Algorithm 2 Short description of the Louvain algorithm

Require: $G = (V, E)$;
for all $i \in V$ **do**
  $C_i := i$;
end for
$improvment = True$
while improvment **do**
  improvment := False;
  for all $i \in V$ **do**
    $C_i := C_i \backslash \{i\}$; {remove node from its community}
    $best\_community := C_k$;
    $best\_gain := 0$;
    for all $j \in C$ **do**
      calculate modularity_gain
      if modularity_gain ¿ best_gain **then**
        $best\_community := C_j$;
        $best\_gain$:= modularity_gain;
      end if
      if $best\_community <> C_i$ **then**
        improvment := True;
      end if
      add node to best_community
    end for
  end for
end while

For the small real world network Zachary's karate club, a pretty popular network in social network analysis, the algorithm performs very well for our purpose. For a bigger network, the arXiv[2] citation network, the level dropped dramatically and had high changes. After looking into it and into the data we found that this depends heavily on the order in which the nodes are considered. This is supported by the authors who claimed exactly that. Taking this into account we have an algorithm that works good for our purpose.

We will use this algorithm to check if there are hidden subgroups in the company investigated or if there are only the communities given by the project structure. The existence of groups besides the groups supported by the company itself might be of good use in new projects.

[2] http://www.arxiv.org

3. Developing a Questionnaire

In this chapter we want to explain how the questionnaire was developed. We will give a short overview over the problems we had to consider given time and expandability of the survey.

3.1. Online versus Offline

There are several options one can use if he wants to investigate a community to get the structure. At first there is the option to conduct interviews. We dismissed this option due to several facts. The first one was that people tend either to not know people outside of their working group or to misjudge them based on the so called intergroup bias hypothesis that leads to a negative judgements of persons who are not in the subjects own group (Seger, Smith, Kinias, & Mackie, 2009). Secondly. it would have been very time-consuming to conduct a survey with each employee and the extendability would be very limited due to this fact. Additionally there was the potential to make the answers sound better due to the direct contact between the interviewer and the respondent.

Conducting the survey in an offline mode by handing out questionnaires to the employees was another option. Like the interviews the major problem was the time - sending out questionnaires and getting them back in a certain amount of time is time-consuming and can be an impossible task if a 100% return rate is needed. Luckily that was not wanted but a high participation was needed to get sufficient results. Other problems were printing costs and distribution costs and for future surveys there would be additional mailing costs which are eliminated by using online surveys (Kaplowitz, Hadlock, & Levine, 2004; Cobanoglu, Warde, & Moreo, 2000). Cobanoglu et al. (2000) even claimed that without the coding costs the online version would be not only cheaper but would also have a significantly higher response rate. That was true in their experiment and is supported by other research that states the response rates of web-based surveys are at least as high as in mail surveys (Kaplowitz et al., 2004; D. M. Bushnell & Parasuraman, 2003; Medlin & Whitten, 2001).

The fact that the average American person is in front of a screen approximately for 8.5 hours a day (Jameson, 2012) which should not be much different in Germany deemed us reasonable to conduct an online survey (Bitkom/gh, 2008). Expandability to other groups than the surveyed is a bonus for future research on this topic or for the company if it should be decided to investigate the overall communication network.

3.2. Questionnaire Development

After we decided to go with an online questionnaire we implemented the GUI in the web application framework Catalyst[1]. We will not provide details how it was implemented.

In several steps we developed the question set for the survey, following some guidelines provided by R. Cross and Parker (2004). The major problem is posing questions that do not influence the surveyed person nor should the question itself sound negative. It should also not carry any positive meaning otherwise the respondent might answer different than he would if the question is stated negatively or neutrally (Malhotra, 2006). The questions should also be closed and leave as little room as possible for interpretation such that they convey the same to every asked person. Thus we decided to give the questions a main text in which the actual question is contained and a additional information text which clarifies what is meant. We will now examine the questions without the answers and will discuss why each questions are important for the communication network.

Please identify persons who are important for your work in any form of information delivery, discussion, collaboration, delegation or coordination?

Starting with this question we get a nice overview of whom a person considers as important. We decided against providing a list of persons they could select with whom they work out of several reasons. First of all, the company investigated was a small to medium-sized business and

[1] `http://www.catalystframework.org/`

going through a list of all employees might be too time-consuming. It might also be confusing if there are names too similar or persons who are actually unimportant and get contacted rather infrequently might get choosen since they are listed. We hoped to get a description of the real work environment as precise as possible.

Please identify persons who have knowledge you might need in your daily work but you do not ask?

This question was supposed to get information about persons who are not asked either out of conflicts or time issues or or there are other sources or other reasons. The question is capable of revealing unused potentials since more contact to those persons might increase the productivity. In fact it is also able to show if someone is strained. If the contact has many listings in this category he may either be very distant to be an useful source of knowledge or he is too involved in other projects and does not have the time to get involved in other projects.

For this question it was problematic to convey what was meant without giving people misleading information.

From now on we always provided the list of persons the respondent gave in the first two questions such that we could assess the value of the connection between the persons.

How close to you do these persons work?

First of all this gives us some geographic information about the persons involved. If they work in the same room or in the same floor it is natural that they know and talk to each other even if it is just chatting in the corridor since it relaxes and increases creativity (*Why You Should Waste Time Chatting at Work*, 2011). If the connections are made to other floors or different departments in another city the connection seems to be more important since the maintenance of such relations is harder.

The locations were chosen according to how it was perceived the most reasonable. Some of the employees in the company investigated had a room with one of their coworkers, others were only on the same floor. Since the company is small to medium sized some branches have more than one floor so it is also reasonable to ask if it is only a different floor or a completely different city.

For how long do you know these persons?

This question gives not only information about how long the respondent works for the company but is also able to reveal together with other questions how good he is at social networking. If he has only contacts he taps since 10 years and no newer contacts this might be due to a lack of interest in new colleagues. The possible answers included up to one year if it was a new acquaintance, up to four years for some of the better established relations and up to ten years for long relations and ten years and longer for real long relations like it is usual for persons with administrative purposes in the company investigated.

How often do you typically interact with these persons in your work?

The information this gives might reveal some interesting information about the network itself. For example it might give some additional information about a connection to someone chosen as important co-worker. If one should be in this group but not often contacted the claim he would be important might give us a hint that the person in question is too involved in work as that he could be asked more often.

How much do you agree with the following statement for each person: "I know the capabilities, skills and knowledge the person has."?

In the additional text it was explicitly mentioned that this does not include capabilities to do anything related to them. Nevertheless it is important to know if a person has knowledge needed to solve a task and thus this question had to be included. It might have been purposeful to have information about the knowledge a person has himself to see if the connections are made by shared knowledge or if they are made to have access to different knowledge.

Do you agree with the following statement for each person: "I would be more efficient in my daily work if I were able to communicate more with this person."?

Together with the question regarding the interaction with other persons this might reveal structural issues or problems with co-workers who are rarely available due to meetings or business travel. As a stand-alone question it is also interesting to see with whom of the co-workers one remembered he would like to be more involved since this might be helpful for future projects if it is mutual.

Does the person engage in the question?

This question was at first intended to see how useful the answers a person provides are. Due to the formulation it sounded very negative and asking the other way around, who could give more useful answers, was considered as rude also. In the progress we decided to go with this question and added as additional information that one should consider how helpful the answer was. Please mind that we did not ask for useful but helpful since even chatting can be very helpful to come up with new ideas (*Why You Should Waste Time Chatting at Work*, 2011).

Did the intensity of communication with the person change over time?

This was asked to support some of the other questions. Mainly the question looking for co-workers not asked often is very interesting with this information since this could model a recent event between two persons. It is also worthwhile to see if new acquaintance are good to talk to or if they are regarded as too new to talk to.

The options of the last four questions were clustered in values since exact naming differs for everyone and people hesitate to give names to relations. Nevertheless, it is hard to decide if the value assigned is the same between two different respondents.

Now we have established a reasonable foundation for the questions asked and we will see in the next chapter some analysis for the answers we received.

4. Diving into the Real World

Up to now we discussed how Social Network Analysis emerged and got more and more important for work settings. We also discussed the graph theory that we will need for the analysis of the data which we discussed how to get in the last chapter. Therefore, we are now able to analyze the data we gathered and use the algorithms described in Section 2.3. Nevertheless, before we start analyzing we will present a different network for comparison.

4.1. Co-Authorship Network

This network captures the co-authorship of papers written by participants of the survey. It is created by the data gathered from the paper archive of the company. We extracted the authors and compared that list first of all to the actual name of the author. Then we reduced the amount of nodes by restricting it to the employees of the company and edited the weight of an edge to the number of papers the persons wrote together.

As one can see in Figure 4.1 the network looks nice. Besides that it is acknowledgeable that it does not tell very much without knowing what the purpose of the edges is. Edges indicate in this network if persons are co-authors for a paper. The thicker the edges are, the more papers name them as co-authors. The lowest number of co-authorship found was just one paper written together. The highest amount of papers written by an author together with one of the groups as co-authors was 20, the average was 4.39. The network has a diameter of 5 and a low density value of 0.125. Nevertheless it is just one component which means that there are no obvious working group separations. To find those we applied the Louvain algorithm (see Section 2.3.3 for further details) for finding communities. It resulted in a reduction to six communities. As one can see in Figure 4.2 there are two communities that are isolated from the overall network. That is reasoned by the fact that we did not take into account persons who did not respond to the survey. Thus some of the links

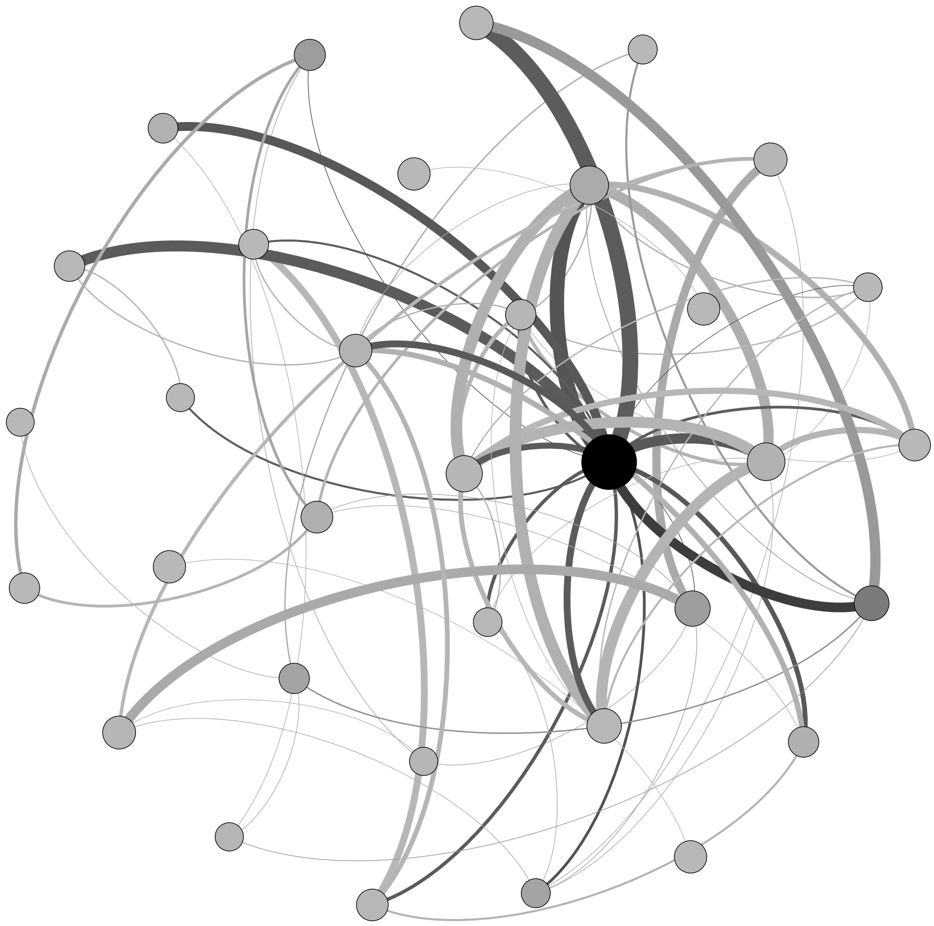


Figure 4.1.: The thickness of an edge indicates on how many papers the persons worked together

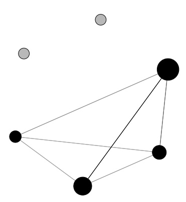


Figure 4.2.: Co-Authorship mapped to communities; the two small communities in the upper part are result of the omission of persons who did not respond to the survey

that exist in the real world do not exist in the shown network. For the other communities the result is reasonable. The working relations reflected in the co-authorship of the papers do exist to a great extent even if they are not an exact matching to the work groups. However, we found that there are certain problems with the co-authorship claims for research groups. Tarnow (1999) claimed that for physicists it would hold true that there were several names as co-authors or supervisors listed who did not write anything and sometimes not even read the articles. He wrote several follow-ups on this topic (see (Tarnow, 2002, 2008)) and came to the conclusion that there are usually more people named as co-authors than needed. It might be interesting to investigate this for the company.

We compared the communities yielded by the Louvain algorithm to the actual working groups and found that the partitioning done based on the co-authorship holds true forfive of the found communities. The algorithm was precise enough to distinguish a subgroup within one of the investigated groups. Three of the other communities that holds true after a closer look at the papers written since they are either written together or on the same topic.

For comparison we used an algorithm online designed by Rosvall and Bergstrom (2008) that analyses the flow of information to reveal a community structure. We took advantage of the already found co-authorship network and assumed that information flow is high between co-authors even if it is arguable according to Tarnow (1999, 2002, 2008). The algorithm found the same communities.

Despite the interesting research that this yields we will compare the communication network we developed with the questionnaire to the network presented here to see differences and commonalities the networks have.

4.2. Analysis of the Survey-Data

In this section we will investigate the data gathered by the survey and apply the former mentioned algorithms (compare Section 2.3). At first we will go to the basic network that shows who works with whom. Then we will analyse the different dimensions of the network (accessibility, knowledge of capabilities, engagement, communication changes) and combine them in the same order. We perform the research in this order since it is reasonable to say that if we know someones capabilities we will try to access him more often. Furthermore, if we know someones capabilities and have enough access it is interesting to know if the perceived engagement is sufficient. We apply in a last layer the changed communication behaviour since it modifies the overall evaluation to a more recent snapshot.

4.2.1. Data Gathering

Before we go into the analysis of the data we will shortly describe where we found problems.

First of all we want to mention the response rate since that was the most concerning problem. Without consistent reminding the people the response rate would probably be below 50% as in many other questionnaires. Even by reminding people it was hard to get to the response rate we got. We had the advantage that we could go to the persons directly and ask if they would not want to participate to reinforce them. Still, the response rate was below 80% (see Table 4.1).

The result of the questionnaire was quite satisfying. Looking at the plain data we found that there are many connections inside groups and only few boundary spanning connections. This might be a problem if one needs knowledge that is not contained in the group knowledge provided by his group and is thus dangerous. However, since it is intended that this separation in groups exists it is fair enough. We would have to remove only a small fraction of

Table 4.1.: Participation

Group \ Participation	no
Group 1	15.4%
Group 2	48.3 %
Overall	29.4%

the surveyed people from the company to completely disconnect both of the groups. The importance of the nodes creating the connections proved to be true by calculating their betweenness value which was the highest for those few nodes since they held the shortest connections between the groups. After we looked into the group with more answers we found two persons were not mentioned at all but they also did not respond to the survey. Overall, there were 7 persons in total who did neither respond nor were they mentioned as important.

In comparison to the co-authorship network (see Figure 4.1) there were much more nodes due to communication with organizational units while we had a concise description for the groups for the co-authorship network such that we were able to include only the persons who actually worked there. We decided to exclude persons who were only mentioned and did not work in the surveyed groups else we would have too different results and a much lower network density. Also, if we consider the PageRank value, there might be some misconceptions caused by the random surfer model. For the nodes that were only mentioned we would have an edge added to each other node with the same probability to use this edge and this would lead to wrong values if we include every mentioned person. Still, we will include members of the surveyed groups who did not respond as long as they were mentioned by someone. The missing response of those persons can be a result of vacations, or privacy concerns or anything else.

4.3. Basic Network

Moving on to the next investigative step we looked at the distribution of in- and out-degree in the network. In the real-world example investigated this is who is considered as important for work (in degree) and whom does one consider as important (out-degree). We hoped to see something like a power law distribution in the connectivity. For the in-degree this did not hold but for the out-degrees the assumption that a social network has a power law distribution did hold. Nonetheless, we had more nodes with low than high in-degrees. Some of the persons did not reply and thus had an out-degree of 0 while their in-degrees varied between 1 and 17.

Besides that there is an imbalance between the in-degrees and out-degrees. Most of the surveyed persons had more outgoing than incoming edges. Still, there were several nodes clearly visible as important since they had a high in-degree. The network itself was not very dense with 10.7% of the possible connections but still had a low average path-length of 3.04 (see Table 4.2 for comparison). This reveals a structure that is capable of high performance if persons are willing to ask freely for advice if necessary.

It is also important to note that the perception of this basic question is different for each person. More than half of the edges, 59.06%, were perceived mutual, not including edges to non-responding persons.

Table 4.2.: Analysis of the basic-network

#Persons	#edges	density	average degree	average path length
60	387	0.107	6.3	3.04

Let us now analyse the basic network based on the other algorithms we wanted to use. We included the network (see Figure 4.3) with some information encoded in the network[1]. The darker the node, the higher its betweenness centrality. The bigger a node, the higher its PageRank. In opposite to the co-authorship graph the edges are all of the same size since there we encoded in the thickness how many papers the persons produced together whereas we here are only showing who knows whom.

By visualizing the network in this way it becomes apparent that removing two nodes, namely the black ones, would divide the network into two strongly clustered subnets. It is also interesting to see that betweenness centrality and PageRank are not directly correlated. The black nodes are very small, thus have a small PageRank value. Other nodes that are completely white, a small

[1] All networks from this section on will also be in a better resolution in the appendix. We will provide an appropriate scheme for referencing the correct depiction

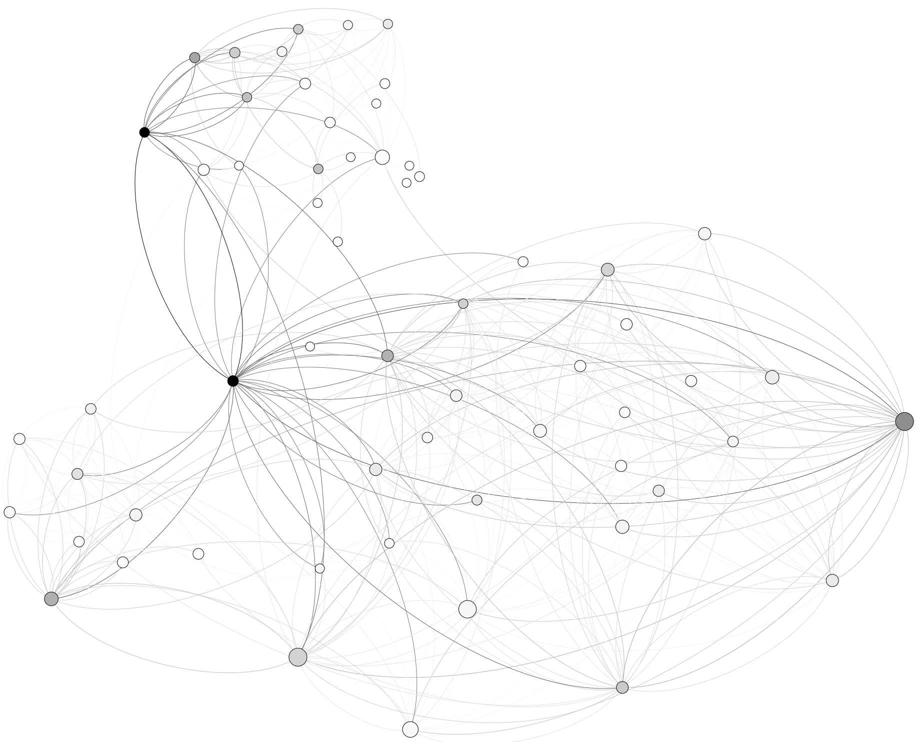


Figure 4.3.: The network show who works with whom and includes the betweenness centrality (color coded) and the PageRank (size coded)[A.1]

betweenness centrality, are very big, having a high PageRank value. This reveals that those persons are highly regarded as contact person. It is not obvious if this circumstance is caused by their seniority or their knowledge or something different as important but they have many ingoing connections.

The Louvain community detection algorithm was very precise and found four communities. Two of those communities are exactly one (sub-)group. For the remaining two it is an interesting fact to acknowledge that one of the communities is actually one person who did not participate. Apparently he was also not very often named and would thus not have increased the modularity of any of the existing communities. This result is very good and gives us more reason to trust the Louvain algorithm.

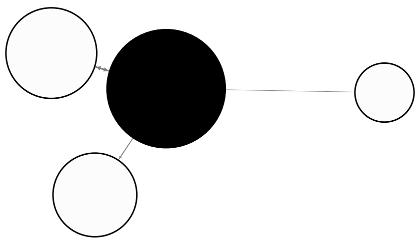


Figure 4.4.: The basic network mapped to communities[A.2]

The Louvain algorithm reveals at this point two important things. This first network was only based on the fact who works with whom the most. Keeping this in mind we see that the main groups as well as the subgroup are well presented in this network as they are prescribed by the organigram. More important, the connections are tied to the organigram, too. The subgroup communicates with its main group and the two main groups with each other but one main group does not communicate with the subgroup. The extra community for the single person has no tie to the groups it does not belong to, also.

Up to know we looked only at the basic network including the work relationship without further adjustments. In the next section we will look at different variations of this network according to the dimensions presented with the survey.

4.4. Evolved networks

At first we analysed the network along the basic dimensions, namely who has access to whom or where is a lack thereof, do the participants know the capabilities of each other, do the persons known engage in questions they are asked and did the communication with someone change recently.

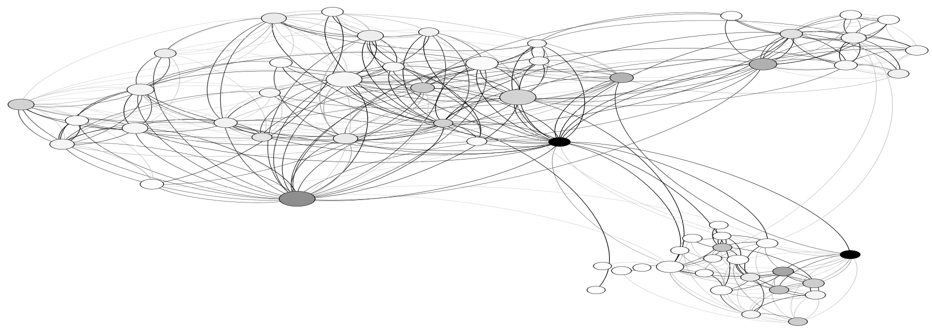
As before we did run the analysis for each of the networks. The density and the average degree did not change since we wanted to get the information for all persons that were named in the survey. Also, the betweenness centrality and the closeness centrality stayed the same since the calculation is done without weights on an edge.

4.4.1. Access-network

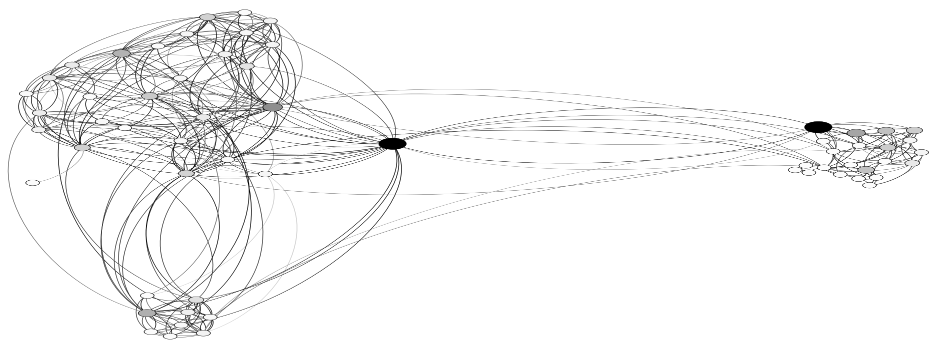
As one might observe the structure of the network, shown in Figure 4.5(a) is the same as we had before. There are three subgroups clearly visible and only two nodes are really dark, indicating a high betweenness value. The PageRank value did change. The most changes were small and only one place up or down. In the middle range the values stayed even constant due to having not as many connections as other nodes. It was not very surprising to see that persons who were known by most of the people also had a higher PageRank value. According to Pedroche (2010) using PageRank to model social competences is possible. So we do have a slightly different ranking in this network.

With the Louvain algorithm we did discover interesting changes on the edge weights as also in the grouping behaviour (see Figure 4.7(a) for comparison). The intergroup values stayed high as one would expect. The group spanning connections were more interesting. The edge from the subgroup to its main group increased in the same manner as the reverse edge did. The other main group split into three groups. One consists like before out of one person, the other one of only three. The weights on the edges from these subgroup to their main group and vice versa are actually pretty low, indicating that the persons inside the groups either did not participate in the survey or that they are a self-sufficient subgroup or they are just new to the company. Moreover, they are not accessible for anyone in the group but the persons in this community who participated in the survey. Still, to those subgroups there are no edges from the other groups.

As we can see inside the groups there is a pretty good communication and

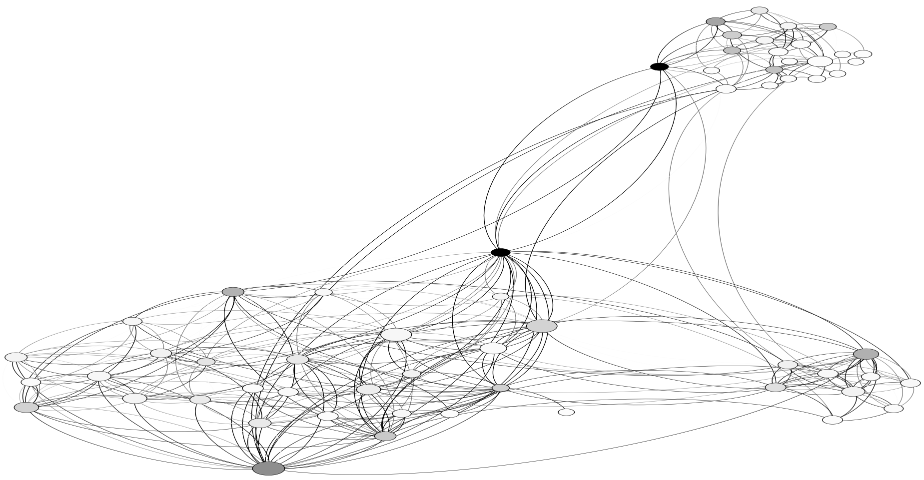


(a) Access-Network[A.3]

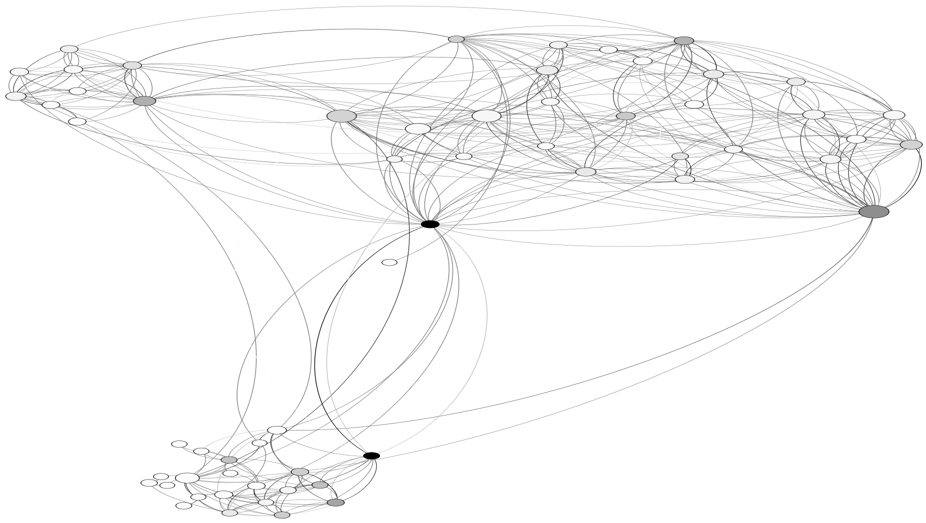


(b) Capabilities-Network[A.4]

Figure 4.5.: (a) shows the feeled accessibility; edges are darker if the people felt they have good access to each other. (b) shows the capabilities network where edges are darker the more people thought they know the capabilities of the other



(a) Engagement-Network[A.5]



(b) Communication-Network[A.6]

Figure 4.6.: (a) shows the perception of engagement; edges are darker if the engagement perceived is higher. (b) shows how the communication changed recently. The darker the edges are the more the communication increased and vice versa.

the persons do not need more access to each other. Especially the lower group, the subgroup, is very well connected and has good access to each other. The most interesting part of the analysis was to see that people who were thought after the most by some persons were mentioned as very good accessible by other persons.

Based on the interaction between the persons we were able to see some quite different interaction patterns. While most of the mutual edges seemed to be well established there were some edges where a clear lack of mutuality was obvious. Some of the persons needed more access to the knowledge of the counterpart of the interaction while they were satisfied with the existing interaction.

4.4.2. Capabilities

The second dimension we wanted to look at was the knowledge about other persons, i.e., how well do employees know what their co-workers are capable of. For the weighted PageRank this did not have any major influence, the order is the same up to some nodes in the range where the values were the same such that a new order there can be ignored. The top nodes and the low nodes did not change which shows that persons who are well known or easily accessible are considered as well known or their experience is well known. It is a natural behaviour for humans that they tend to think to know their surrounding very well. Actually only two persons admitted not to know very well what the other person knows or does. This is a pretty good result. It shows that employees of the company known their peers well enough to make a statement about their knowledge and they know whom to ask if a problem occurs. If this knowledge would not exist it would hinder the work flow since they might ask co-workers who do not have the correct knowledge. This would not only be confusing for the other part but would also take some time to figure out and thus delay the solution of the problem.

The frequency of interaction was better than before. According to our survey most of the mutual edges indicated that people who talk to each other more frequently also know each others capabilities very well. Only a few edges with to little interaction had a high amount of knowledge of the other persons capabilities.

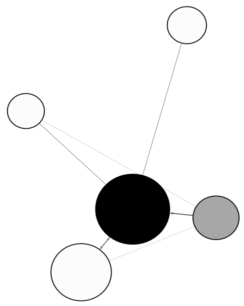
The Louvain algorithm gave us this time a very loose network (for comparison see Figure 4.7(b). There is no possibility to come from one group to every other. This is similar to the community structure from the basic-network (see Figure 4.4). In comparison to the access-community structure it seems like the persons would know the capabilities of the three person group even if they are not accessible.

4.4.3. Engagement

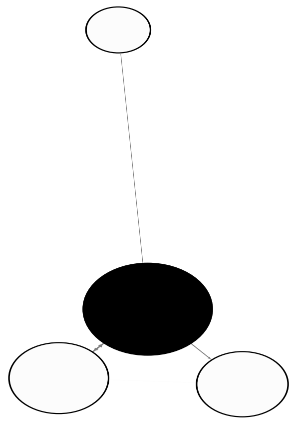
This time we start with the result of the Louvain algorithm. We found four communities at all which is a result of the very strong structure of the network shown in Figure 4.6(a). In the network a strong colored edge means that the persons feel that the other engages very much in the question and tries to be of assistance in answering the question. As we can see there are a lot of strongly colored edges. 11% of the edges are actually indicating too little engagement. Nevertheless, over 50% of the edges indicate that the engagement is either not known because the persons are not asked enough or the answers provided are satisfying but not significantly better than other answers. Due to so many good answers we have quite a different structure and the nodes are evenly distributed in the communities. This circumstance explains that the betweenness centrality of all communities is the same. Due to the shifted community affiliation the network is completely connected and every node is on all shortest path or all paths are shortest paths.

4.4.4. Communication

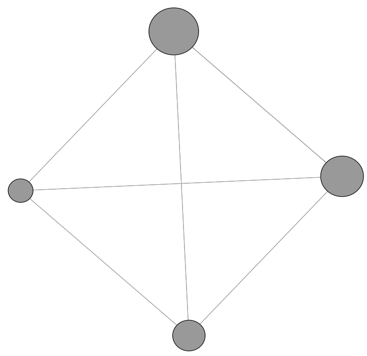
Let us now have a look at the communication network. Here we wanted to get a look at how the communication changed recently in the groups asking if the level of communication increased or decreased or stayed the same. The visualization is here the most helpful part (see Figure 4.6(b)). As we can see there are again the three clusterings, this time indicating that the communication inside the clusters is very good. Like usual the color of the edges indicates how good the connections are, the darker an edge the more the persons communicated recently. As expected only a small percentage decreased their communication (14.8% of all edges have weights indicating a decrease) whereas there are more increased communication channels (27.7% of the edges show weights indicating an increase in communication). The rest of the communication stayed the same over time or did not change in the recent past. The most interesting in this network is the black node that is in the middle of the complete network.



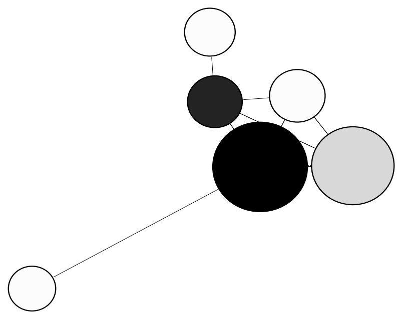
(a) Access-Community[A.7]



(b) Capabilities-Community[A.8]



(c) Engagement-Commity[A.9]



(d) Communication-Community[A.10]

Figure 4.7.: The community networks of the corresponding network. Color indicates the betweenness centrality. For (c) there is no different color since the path is fully connected.

Due to the layout algorithm used a node tends to be near nodes where the edge weight is high. In the previous networks the node was always on the edge to its man group but still far away from the other groups since it has many connections not only in its group but also to the other. Since the weight of the edges influences the placement of a node and the groups are not as far from each other as in the other networks it is fair to say that the person who is behind this node is to the other groups as communicative than he is to the other groups.

The Louvain algorithm yielded a network with 5 communities. This network is very different from the ones we saw before. The community in the lower right corner consists out of one person belonging to the group right next to it. The community in the upper left corner consists also of one single person. Those persons did not seem to be as important to the judgement of the algorithm. It may very well be that another order of the nodes might pull them to their according group. Nevertheless, we can also observe a split of the group that is in all other community-networks in Figure 4.7 the black or the big one. This sounds reasonable if we take into account that most of the edges in this network had a weight that indicated neither more nor less communication.

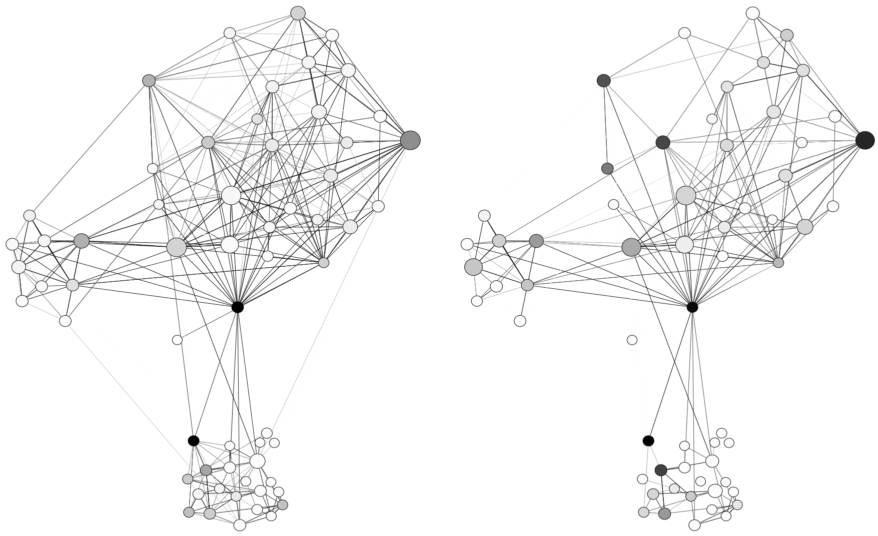
After analysing the four basic dimensions we investigated we will have a look at the combinations. This idea was shown to work well in (R. Cross & Parker, 2004). There the researches had a different type of questionnaire and were thus able to cut edges if they had a weight below a certain threshold or if the person was not mentioned at all. We added the weight of the edges up to get a simple, directed network.

The statistics for the next networks stay the same like provided in Table 4.2.

4.4.5. Capabilities-Access

As we mentioned above we had always the complete data set given. This enabled us to do a qualitative research of the network. For two dimensions we were able to determine a certain threshold where the persons had enough access to each other as also enough knowledge about the capabilities. Thus we are able to compare the actual network resulting from adding the edge values and a comparison network we analyzed.

As one can see in Figure 4.8(b) the overall combination of accessibility and knowledge of capabilities is pretty good. To be exact there are still 51.85% of all edges that were in the network in Figure 4.8. That is a really good value

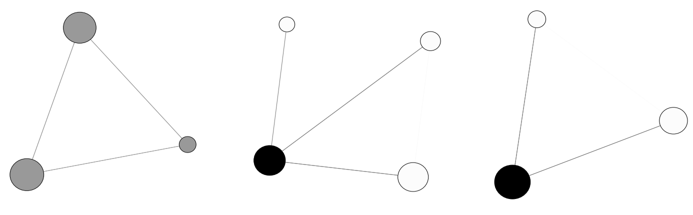


(a) Capabilities-Access[A.11] (b) restricted Capabilities-Access[A.12]

Figure 4.8.: (a) shows the combined network for the dimensions of knowledge of capabilities and accessibility; (b) shows only edges with a sufficient weight such that those edges do not lack in any of the specified dimensions.

since we left out all edges that showed a lack of accessibility or showed somehow a lack of knowledge about the capabilities. It increased also the betweenness centrality of some of the nodes. Of course this is caused by the vanished edges such that there had to be new shortest ways calculated. More importantly, it shows that there are several nodes inside the groups which can be easily accessed and they can be used as so called connectors. They might be able to retrieve needed knowledge or handle questions to the correct persons. It is also important to note that some of the former more between nodes, possible connectors so to say, are now lower while others have arisen due to a good combination of accessibility and their known capabilities. Specifically we want to note the three big nodes in the upper right cluster which increased their betweenness. They all have a high number of in-degrees but also a reasonable to high number of out-degrees. This provides even more reason to believe in those persons as connectors.

The Louvain algorithm gave us a new layout (see Figure 4.9(a)). All of the connections are bidirectional and we ended up with only three nodes. We have to take into account that this is a result of the order the communities got calculated but it is a convincing result. Considering the distribution of the knowledge of the capabilities and the accessibility of the co-workers the algorithm ended up with only three communities with is congruent with the number of actual work groups.



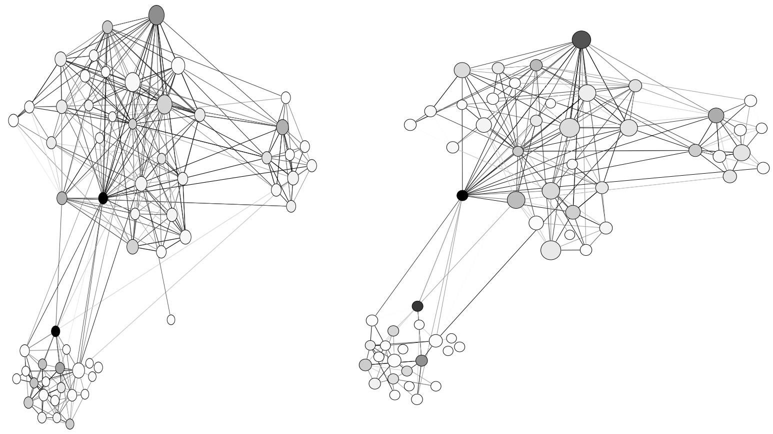
(a) Capabilities-Access-Communities[A.13]

(b) Capabilities-Access-Engagement-Communities[A.14]

(c) Capabilities-Access-Engagement-Communication-Communities[A.15]

Figure 4.9.: Community structures for the combined networks

4.4.6. Capabilites-Access-Engagement



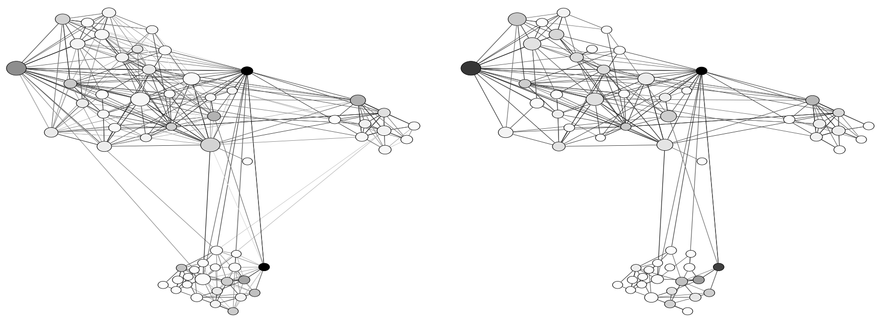
(a) Capabilities-Access-Engagement[A.16] (b) restricted Capabilities-Access-Engagement[A.17]

Figure 4.10.: (i) shows the combined network for the dimensions of knowledge of capabilities and accessibility and engagement; (j) shows only edges with a sufficient weight such that those edges do not lack in any of the specified dimensions.

As before we restricted the edges in Figure 4.10(b) such that the worst options dropped off the list. The remaining edges capture mostly edges with at least average valuation or if one of the options is high and the others are in the mediocre range. In contrast with the network shown in Subsection 4.4.5 the nodes changed their betweenness centrality differently. The change in the betweenness is based probably on the fact who engages more. We can not conclude from this fact that persons who are not as engaging as others are not willing to since it may be that they have only limited time. As a significant change one might observe the lower subgroup. In Figure 4.8(b) there were two nodes with high betweenness centrality values whereas in Figure 4.10(b) is actually only one node with a high and one with a fairly high value. The one with the fairly high value is the former high valued node.

The Louvain algorithm showed us four different communities (see Figure 4.9(b)). The lower left three communities are completely connected even if the connection from the lowest community to the most left community is relatively weak. The community in the upper right consists only of one person. This could be again a result of the order in which the algorithm works his way through the nodes but it may also be a result of a low connection. The person is still connected to the network in Figure 4.10(b) but the connection is overall the only this person has.

4.4.7. Capabilities-Access-Engagement-Communication



(a) Capabilities-Access-Engagement-Communication$^{A.18}$ (b) restricted Capabilities-Access-Engagement-Communication$^{A.19}$

Figure 4.11.: (a) shows the combined network for the dimensions of knowledge of the four dimensions; (b) shows only edges with a sufficient weight such that those edges do not lack in any of the specified dimensions.

Finally, we combined all four dimensions into one evaluation. The edges are all not very strong but we are still able to make some assumptions about them. Overall there are 35.71% of the edges with a very low weight indicating that they not only lack in the first three dimensions but that also the communication between the persons decreased. The rest of the edges, 64.29%, are in a range where it is reasonable to say that they have the needed access to persons with knowledge and either a normal or increased communication with them. We can see the restricted network (Figure 4.11(b)) with only the reasonable edges in comparison to the complete network (Figure 4.11(a)). It is simple to recognize why some nodes in the right figure have a higher betweenness centrality. Taking the vanished edges into account we see that several of the old shortest paths have vanished leaving fewer of them. In comparison with Figure 4.10(b) we have more nodes with a higher betweenness centrality. Since this is the most complicated network we can assume that those persons are important connectors in each of the groups. If these nodes would vanish it will have an impact to the complete network.

The Louvain algorithm gave a network consisting out of three communities. It is fair to assume that this network is a good representation of the working relations we were able to discover with the survey. Of course this depends on the willingness to answer which was not 100% (see Table 4.1). Important for this community network is to recognize the missing link between the two white nodes, implying that the communities are not exactly a mapping to the working groups.

4.4.8. Difference-Network

After we have analyzed the four basic dimensions we will take a short break from going deeper into this and we will analyze the difference network. We were able to produce a difference network based on the first and second question, showing people who were not chosen as important co-workers. Now we are able to see where the problem is.

Although we call this section difference-network we will not show the actual network since it would increase the possibility to conclude which company was investigated and even the persons.

Overall, there were only 14 connections between very few persons. Accessibility was only mentioned two times as a major problem. In both cases the communication did also not increase, so we have to assume that these connections will eventually fade completely from the network. The most prominent problem seemed to be the willingness to engage in a question followed by the communication. While the former was always voted really low the latter was almost always showing either no recent changes. That may include no communication at all. Thus we needed to analyze this with the three additional dimensions, namely time, interaction and physical proximity. With the interaction as additional value alone it was obvious why there was no communication change and why there was also little to none engagement. They just did interact with each other too irregularly to get to a better value. This does not imply that this is a crux. It can be that those sporadic connections are only for administrative issues and thus they should not be forced to change their pattern of behaviour. It would require a personal interview with the persons to investigate these issues.

We did not include the three dimensions before for a simple reason. Neither did it change the overall appeareance of the network – it was still clustered in three regions – nor did it do anything for the analysis. Most of the employees do work in the same building and a working group is most of the time settled on one floor. So the factor of the closeness does not reveal very much only for a scant few connections that spanned not only floors but cities. These connections proved to be stronger and changed the PageRank value accordingly. Still, the changes were not significant since there were only few connections to other cities. On the other hand, the number of connections to another floor was higher but it still did not change much since the people tended to connect more to their floor than to the other floor. Overall, there were only 11.9% connections made not on the same floor. The interaction between persons did also not influence the network very much. With only 10.5% of all edges classified as unregular (or not at all, 0.7%) contacted persons the company can be seen as pretty communicative. The additional 13.75% classified as fewer than several times a month might be more influencing but there is still to consider that the questions were not answered by the same person. Therefore we can not be entirely sure if some answers do not fall actually in the category “several times a month” and the other way around.

Up to now we omitted the question about the time two persons know each other. And this is the first time the results changed. Figuring that newer working relations are not as well established as old relations we set the weight for newer connections higher. As it is visible in Figure 4.12(a) the clustering from the other networks seems to dissolve but in the lower left corner. This group appears to be fairly new assembled or it contains a lot of new hires such that there the edges had more weight. This proved to be true after a lookup in the extracted data. The upper half contains both, persons who are long in the company and are responsible for the far spread as also some relatively new guys as also new hires. For example, in the upper right corner we can see a small clustering consisting out of several persons who know each other for roughly the same amount of time.

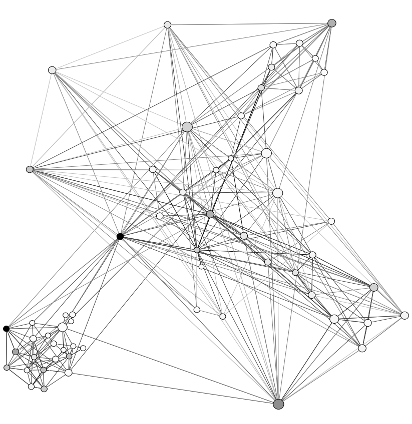
In Figure 4.12(b) is this group even tighter clustered as in Figure 4.12(a) and we can distinguish it as the former mentioned subgroup again. In this visualization seems to develop a new cluster in the upper right part of the network. This proved to be true in the community analysis with the Louvain algorithm. As it is shown in Figure 4.12(c) we now have five communities, one consists almost out of the node in the mentioned new cluster in Figure 4.12(b). The other clusters are actually the communities we were able to find in all the other networks.

Since the other networks did not yield any new information about the structure of the groups we will now consider again the co-authorship network from Section 4.1.

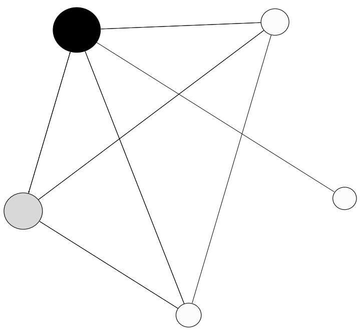
4.4.9. Comparsion between different networks

First of all we will compare the connectivity. Since in the co-authorship network are less nodes than in the networks retrieved by the survey the values of the PageRank-vectors were different. But there were other changes as well. As the most important, the order of the persons sorted according to their PageRank value changed very much. This is important to consider for the relations. If we are in the co-authorship network there are different nodes important. That might lead a new hire to judge that those persons are either competent or helpful in writing a paper. Taking some different network like the basic network modified with the interaction evaluation a new hire would see more important persons to contact if he wants to get to know people who can help him in his research.

To make this more explicit look at Figure 4.13(a). A new hire might try to communicate with the dark node at first and discover that he could lead him where he needs to be. Since this node has a high betweenness value it is fair to assume his time is already scarce and he might not have the needed time to supply the new employee with the needed information. In the Figure 4.13(b) he not only has more choices but he also has the ability

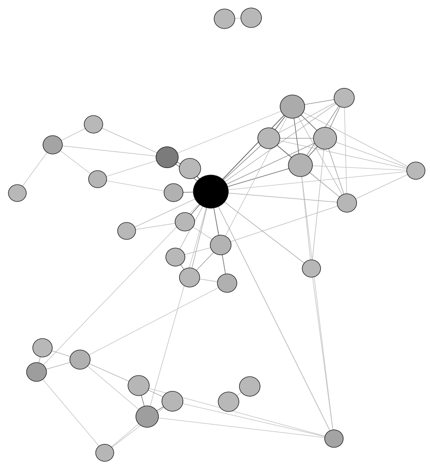


(a) Time modified basic network[A.20] (b) Time modified access network[A.21]

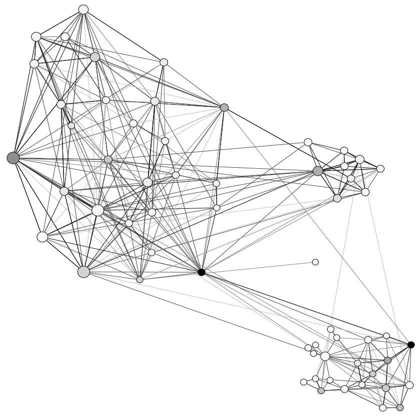


(c) Time modified access network communities[A.22]

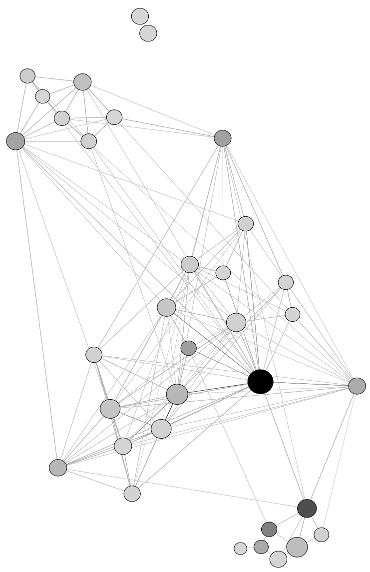
Figure 4.12.: (a), (b): Two examples of how the relations are changed if time is considered as an influencing factor; (c) shows a different community structure than observed before due to the time-factor



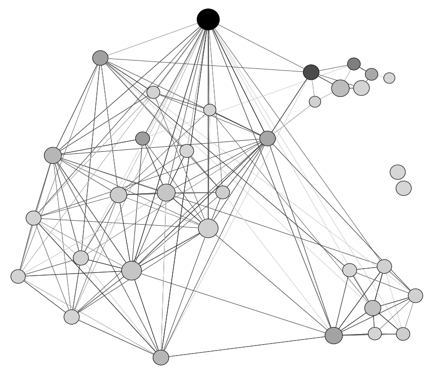
(a) Co-Authorship[A.23]



(b) Basic-Network, modified with interaciton[A.24]



(c) Both networks mixed[A.25]



(d) Four dimension-network, mixed with co-authorship[A.26]

Figure 4.13.: (a) shows the co-authorship network from Section 4.1 again with a different view; (b) shows the basic network modified for the interaction; it is clearly visible that not only the PageRank values have a different distribution but also the betweenness values have changed; (c) shows the mixture of both networks.

to discover to which working group the person belongs to. He can still ask the persons with the highest betweenness values if they have time to show him to the needed contacts. Now he has also the information needed to tell groups apart and thus he can interact with the correct persons immediately. Both networks together might tell him even more about the working relations inside the company and thus lead him more directly to the needed contacts (Figure 4.13(c)). For the combination of the two networks we restricted our selves to the persons who participated in the survey. Or the combination of the network with the four dimensions from Figure 4.11(a) together with the co-authorship network as a map of whom to talk to in this matter might be an even better idea to guide new employees. This network has like the co-authorship network only one very between node but as before we can see the groups more clearly and we have a slightly better adjusted estimation of the PageRank.

5. Conclusion

We will give a short summary at first about what we have seen. After this we will give some options on how to improve the ideas used and how the developed questionnaire might be improved for further uses. Also we will revisit the analysis to give examples about possible variations of our algorithms or other algorithms that might be used not only for companies but also for research interests. In the end the own opinion of the author will be stated with some thoughts about how to increase the willingness of employees to participate in the survey and why not everyone participated.

5.1. Summary

After the introduction to the topic where we introduced several other researchers and their findings we introduced the reader to graph theory. This was a basis for our research and thus needed. With the concepts introduced we were able to explain the algorithms used. For some of the algorithms, for example the Louvain-algorithm, it would have been possible to explain this one without the graph theoretical background and explain it by talking about social communities and how to set them the best but this would need much more experience on social behaviour on our behalf and is thus left. Nevertheless, this chapter was not necessary to be understand completely to get the idea of the next chapter where we dived into the explanation of the questionnaire. Afterwards we went on and examined the results of the four weeks survey with some interesting results. As we saw most of the communities we found were either close or exactly the existing working groups which shows that the ties in the company are well made. Also, it shows that the existing relations should either be deepened or at least maintained. Some of the community structures showed that very clearly, see Figure 4.12(c) for comparison, where we had one new community based on the ages of their relationship. If they should drift of from the network there is the possibility for a loss of knowledge for the company. Overall, the results showed a positive result indicating no further problem inside the investigated research groups. In addition to the existing communication network we compared it with the co-authorship network and discovered that the combined network might yield more information if we want to find the correct research group. Nevertheless it is not as helpful to use the co-authorship network since there are first of all not alls employees mentioned whereas in the survey everyone is mentioned or able to answer.

5.2. Future research

The main idea of doing the survey online was saving money and saving time. This is still valid but it is complicated to motivate everyone to participate in an online survey. Reminding them via mail might lead to the feeling of an obnoxious researcher who is only interested in his success. Thus it might be more intelligent to go with the idea of D. M. Bushnell and Parasuraman (2003) who send their survey either with mail or a postcard in advance and had relative good response rates. Admittedly this results in higher prices to conduct the complete survey with prices between 1.10 € for a letter and answer letter up to more than 11 € for a certified mail package. It might also be a good idea to get the employees more motivated by informing the head of the corresponding group of the lack of motivation.

Nevertheless, there are also algorithmic ideas that can be introduced into the field of research to improve the analysis. We know there is research in mail correspondence as well as in email correspondence (Adamic & Adar, 2008; Culotta, Bekkerman, & Mccallum, 2004; Viégas, Golder, & Donath, 2006) and this might be useful to get an additional social network with references to the abilities the persons involved have. To the best of our knowledge no use of both, the social network as perceived by the persons as well as their email network, has been reported. Thus it might be interesting to see if there are additional persons in one of both networks yielding information about why, how or where the person had been contacted. As an additional advantage of this research the company would also get extensive information about the knowledge the scientist have since it is possible to extrapolate the topics emails are about (Culotta et al., 2004).

While this might yield additional information about the persons involved in the daily work it also might be interesting to introduce a flow analysis based on the email correspondence to discover communities as it is done by Rosvall and Bergstrom (2008) (Rosvall & Bergstrom, 2007). Furthermore, it might be interesting to predict changes and thus it might be interesting to use the Alluvial-algorithm developed by Rosvall and Bergstrom (2010) that implements bootstrap resampling of an existing network to get the structual change in the network. Doing so not with bootstraped networks but real networks might reveal important changes over time as for example behavioural changes depending on projects or arguments.

Up to now we presented only additional ideas how to get information. We know, there are problems with some of the ideas due to the data protection act. One would have to guarantee that all emails either are only work related before evaluating them or evaluating all without looking into them and discard all mislead topics. Both are not necessarily suitable solutions but to our best knowledge there is no email system that separates emails directly in only-work-related and miscellaneous. Thus we will now also go shortly over some algorithms that we were not able to test due to time restrictions.

There are algorithms that are supposed to analyze opinion propagation dynamics. They can be influenced by vanity. Usually this is modeled with static values for influence of the propagation factor and the vanity factor (Deffuant, Carletti, & Huet, 2012) but one can easily think about inverse PageRank values as the propagation factor for each person. Without loss of generality we expect to see from these algorithms some interesting predictions for the future trend of communication and opinion making. Considering the wide range of behaviours defined by the social sciences one can model even more diverse behaviours. For example, Rosvall and Sneppen (2006) developed a different model in which chatting, cheating and lying was considered and edges got created by random and they talked always about one other person in the network where in the Leviathan model (Deffuant et al., 2012) the talk was about the minimum of people known and a hard defined minimum higher than 1. Also the models differ in the link-behaviour, the Leviathan model keeps information and therefore links while the model presented by Rosvall and Sneppen loses links by rewiring. It is possible to define a combined model or to define something completely different, an advantage of the social sciences. We can also think about some non-social uses of the algorithm. Considering the scale-free nature of social networks one might use this network and apply a Monte-Carlo-Markov-Chain-algorithm to get different configurations of the network. In these the researcher then might try to find interesting patterns that are not special on a first peek.

The questionnaire itself can be improved by asking the participants how the questions can be made shorter, more self-explaining and more clear. While we have there some easy options for the simple improvement of the language and maybe the appearance we still do not have the wanted response rate of 100%. To the best of our belief it is not possible to get this response rate in a real life survey. Nevertheless we have some ideas how to improve the number of responses:

- visit the department investigated and explain yourself
- get the head of the department to support your interest
- set an incentive

Although we followed all of our suggestions we did not get close to the rate we were hoping for. After re-consideration we came to the conclusion that we should have been more persistent and spoken more directly to the employees of the departments.

5.3. Final words

To come to an end of this thesis I will present my own opinion about the problems I had to face.

One of the biggest problems was of course the response rate like written above. Even with the support of the heads of the departments it was hard to get the people to collaborate with me. It depended strongly on the relation with the person and where the person worked. For one group it was very easy since I knew most of the people and they knew me such that answering the questionnaire was not a problem. Still, that defied the fact that an interviewer should be neutral against the interviewed person (Rubin & Babbie, 2005). The other group was neutral and the number of responses dropped with the familiarity. I guess it was good that I was only to investigate the two groups. If I would have done all of them it would have been more than complicated to get he people to participate if they do not know me. Even in the departments investigated there were enough problems, even after setting an incentive and getting the heads of the departments to have my back and support me.

It was also an interesting task to develop a questionnaire. I never thought that there are so many different interpretations of so few words like I used in the questions posed. I want to thank specially my advisor for his help in the design of the questions and also the persons who helped me improving the questions.

References

Adamic, L., & Adar, E. (2008, July 3). *How to search a social network* (Vol. 27) (No. 3). Available from http://arxiv.org/abs/cond-mat/0310120

Arbitron/Edison Research. (2011). *The infinite dial 2011.* http://www.edisonresearch.com/home/archives/2011/04/the_infinite_dial_2011.php.

Bank of America Merrill Lynch Global Research. (2011, May). *Report: Do almost all americans use facebook yet?* Retrieved 02.11.2011, from http://www.allfacebook.com/report-do-almost-all-americans-use-facebook-yet-2011-05

Bitkom/gh. (2008). *Umfrage: Deutsche sitzen lange vor dem computer.* Retrieved 27.03.2012, from http://www.computerbild.de/artikel/cb-News-PC-Hardware-Bitkom-Umfrage-Deutsche-sitzen-lange-vor-dem-Computer-3276484.html

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. (2008) P10008*.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, *25*(2), 163-177.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual search engine. In *Proceedings of the ACM Conference on the World Wide Web.*

CareerEnlightment. (2011, August). *How businesses use social media for recruiting.* Retrieved 19.04.2012, from http://mashable.com/2011/08/28/social-media-recruiting-infographic/

Chlldress, B. (2011). *Hawthorne experiments.* Retrieved 02.11.2011, from http://www.referenceforbusiness.com/encyclopedia/Gov-Inc/Hawthorne-Experiments.html

Cobanoglu, C., Warde, B., & Moreo, P. J. (2000). *A comparison of mail, fax, ans web-based survey methods.*

Cross, Parker, A., Prusak, L., & Borgatti, S. (2001, November). Knowing what we know: Supporting knowledge creation and sharing in social networks. *Organizational Dynamics*, *30*(2), 100.

Cross, R., Borgatti, S. P., & Parker, A. (2002). Making invisible work visible: Using social network analysis to support strategic collaboration. *California Management Review*, *44*(2), 25–46.

Cross, R., & Parker, A. (2004). *The hidden power of social networks:*

understanding how work really gets done in organizations. Boston: Harvard Business School Press. Hardcover. Available from http://books.google.de/books?id=vQ3mM4Vpix8C

Culbert, S. A. (2008, October). *Get rid of the performance review!* Retrieved 19.04.2012, from http://online.wsj.com/article/SB122426318874844933.html

Culotta, A., Bekkerman, R., & Mccallum, A. (2004). Extracting social networks and contact information from Email and the Web. In *Ceas-1.* Available from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.59.7117

Deffuant, G., Carletti, T., & Huet, S. (2012, March). *The leviathan model: Absolute dominance, generalised distrust and other patterns emerging from combining vanity with opinion propagation.*

D. M. Bushnell, M. L. M., & Parasuraman, B. (2003). Electronic versus paper questionnaires: Comparison in persons with asthma. *Journal of Asthma*, *40*(7), 1-11.

Heidemann, J., Klier, M., & Probst, F. (2010). Identifying key users in online social networks: A pagerank based approach. In *Icis 2010 proceedings.*

Ingelbrecht, N., Patrick, C., & Foong, K.-Y. (2010). *User survey analysis: Consumer marketing using social network analysis, worldwide* (Tech. Rep.). Gartner. Available from http://www.gartner.com/it/page.jsp?id=1409213

Jameson, M. (2012, January 2). *From more time in front of screens and less in bed, the american lifestyle promotes added pounds.* Retrieved 23.01.2012, from http://www.sun-sentinel.com/health/os-why-americans-are-fat-lifestyle-20120102-20,0,5960309.story

Kaplowitz, M. D., Hadlock, T. D., & Levine, R. (2004). A comparison of web and mail survey response rates. *Public Opinion Quarterly*, *68*(1), 94-101. Available from http://poq.oxfordjournals.org/content/68/1/94.short

Malhotra, N. (2006). *The handbook of market research: Uses, misuses, and future advances* (R. Grover & M. Vriens, Eds.). Rajiv Grover and Marco Vriens.

Medlin, B., & Whitten, D. (2001, Juli). *A comparions of internet and mail survey methodologies.* Retrieved 23.01.2012, from http://www.quirks.com/articles/a2001/20010703.aspx?searchID=331297459&sort=9

Nebel, M. (2010). Skript zur Vorlesung Entwurf und Analyse von Algorithmen.

Newman, M. (2005). A measure of betweenness centrality based on random walks. *Social networks*, *27*(1), 39–54.

Newman, M. E. J. (2010). *Networks: An introduction.* Oxford University Press.

Obradovic, D., Rueger, C., & Dengel, A. (2011, 10). Core/periphery structure versus clustering in international weblogs. In A. Abraham, E. Corchado, R. Alhajj, & V. Snasel (Eds.), *Proceedings of the 2011 international conference on computational aspects of social networks (cason 2011). international conference on computational aspects of social networks (cason-11), october 19-21, salamanca, spain* (p. 1-6). IEEE. Available from http://dblp.uni-trier.de/db/conf/cason/cason2011.html#ObradovicRD11

Pedroche, F. (2010). Ranking nodes in social network sites using biased pagerank. In *2. encuentro de álgebra lineal análisis matricial y aplicaciones.*

Pedroche, F., Moreno, F., González, A., & Valencia, A. (2011). Leadership groups on social network sites based on personalized pagerank. In *Modelling for engineering and human behaviour.*

Rosvall, M., & Bergstrom, C. T. (2007). An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, *104*(18), 7327-7331. Available from http://www.pnas.org/content/104/18/7327.abstract

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118-1123. Available from http://www.pnas.org/content/105/4/1118.abstract

Rosvall, M., & Bergstrom, C. T. (2010, 01). Mapping change in large networks. *PLoS ONE*, *5*(1), e8694. Available from http://dx.doi.org/10.1371%2Fjournal.pone.0008694

Rosvall, M., & Sneppen, K. (2006). Modeling self-organization of communication and topology in social networks. *Phys. Rev. E*, *74*(1), 016108.

Rubin, A., & Babbie, E. (2005). *Research methods for social work.* Thomson/Brooks/Cole. Available from http://books.google.de/books?id=eAdbEn-yZbcC

Scott, J. (2000). *Social network analysis: a handbook.* SAGE Publications. Available from http://books.google.com/books?id=okSev70ua3gC

Seger, C. R., Smith, E. R., Kinias, Z., & Mackie, D. M. (2009). Knowing how they feel: Perceiving emotions felt by outgroups. *Journal of Experimental Social Psychology*, *45*(1), 80 - 89.

Tarnow, E. (1999, Juni). *An offending survey.* Retrieved 02.04.2012, from http://www.salon.com/1999/06/14/scientific_authorship/

Tarnow, E. (2002). Coauthorship in physics. *Science and Engineering Ethics*.

Tarnow, E. (2008). Subconscious ratings of inappropriate coauthorship in physics. *The Open Ethics Journal*.

Türling, F. (2007, Juni). *Business intelligence und web 2.0: Modellierung von social networks.* http://www.prudsys.de/Service/Downloads/files/DMC2007_tuerling_shsviveon.pdf.

Viégas, F. B., Golder, S., & Donath, J. (2006). Visualizing email content: portraying relationships from conversational histories. In *Chi '06: Proceedings of the sigchi conference on human factors in computing systems* (pp. 979–988). New York, NY, USA: ACM Press. Available from http://dx.doi.org/10.1145/1124772.1124919

Why you should waste time chatting at work. (2011, September). Retrieved 26.03.2012, from http://bitesizebio.com/articles/why-you-should-waste-time-chatting-at-work/

Wiggins, A., McQuaid, M., & Adamic, L. (2006). The small world of academic hiring networks. In *Applications of social network analysis: Proceedings of the 3rd conference on applications of social network analysis.*

Wikipedia. (2011a, October). *Six degrees of separation.* Retrieved 02.11.2011, from http://en.wikipedia.org/wiki/Six_degrees_of_separation

Wikipedia. (2011b, October). *Social network: Social network analysis.* Retrieved 02.11.2011, from http://en.wikipedia.org/wiki/Social_network#Social_network_analysis

Zain, B. (2011). *How many real offline friends do you really have?* Retrieved 20.01.2012, from http://thereasoner.com/articles/how-many-real-offline-friends-do-you-really-have

A. Network depictions

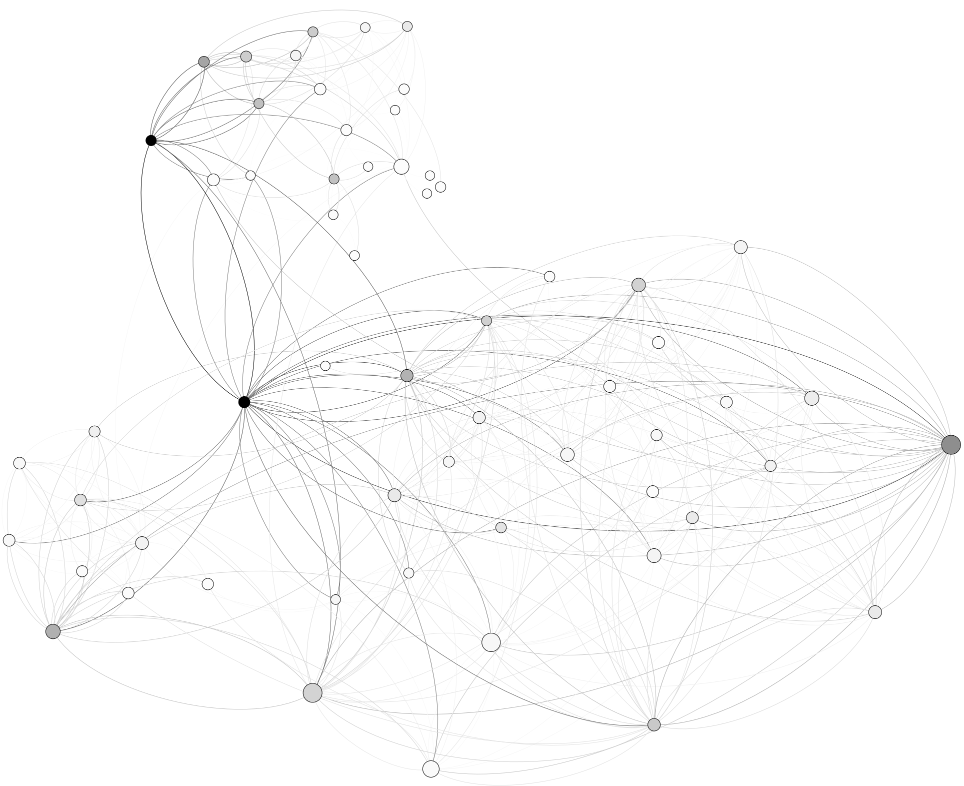


Figure A.1.: Basic-Network: Who works with whom[4.3]

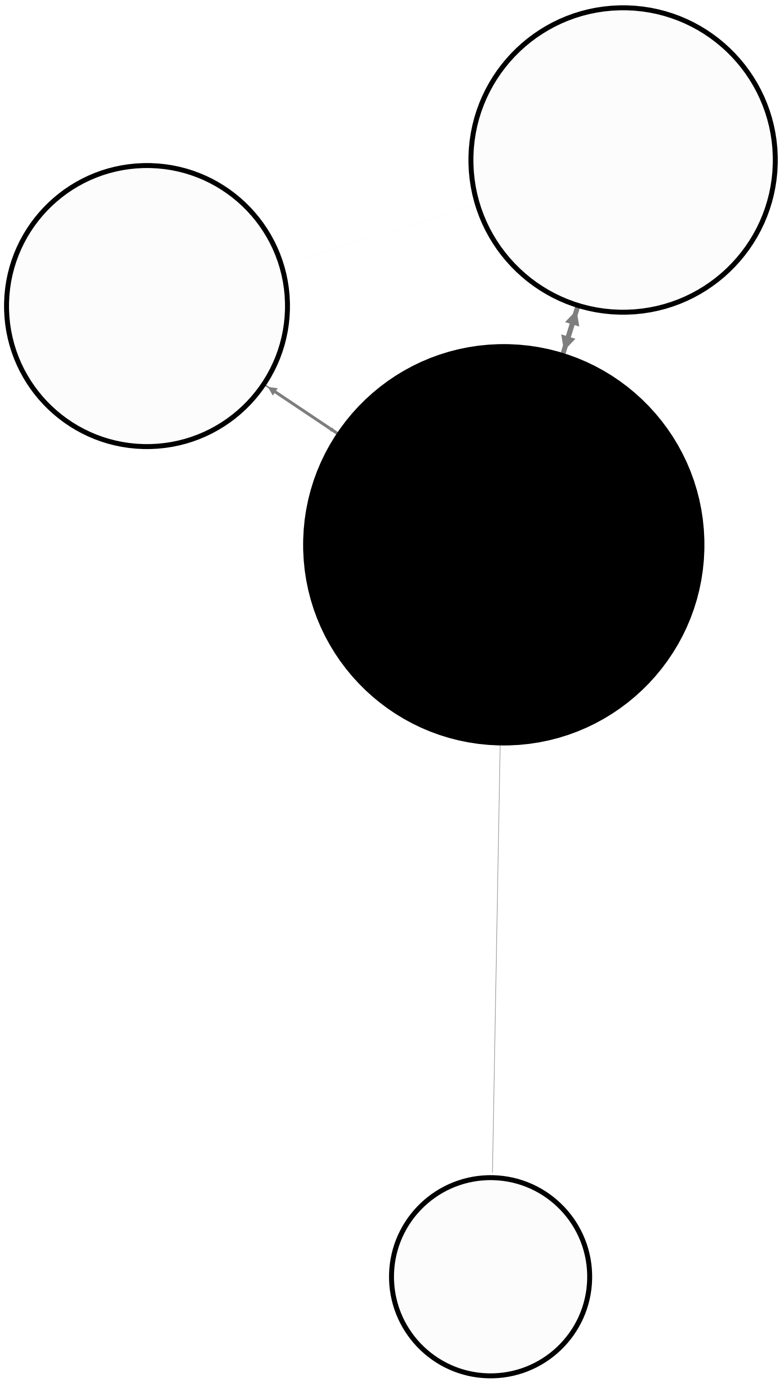


Figure A.2.: Basic-Network mapped to communities[4.4]

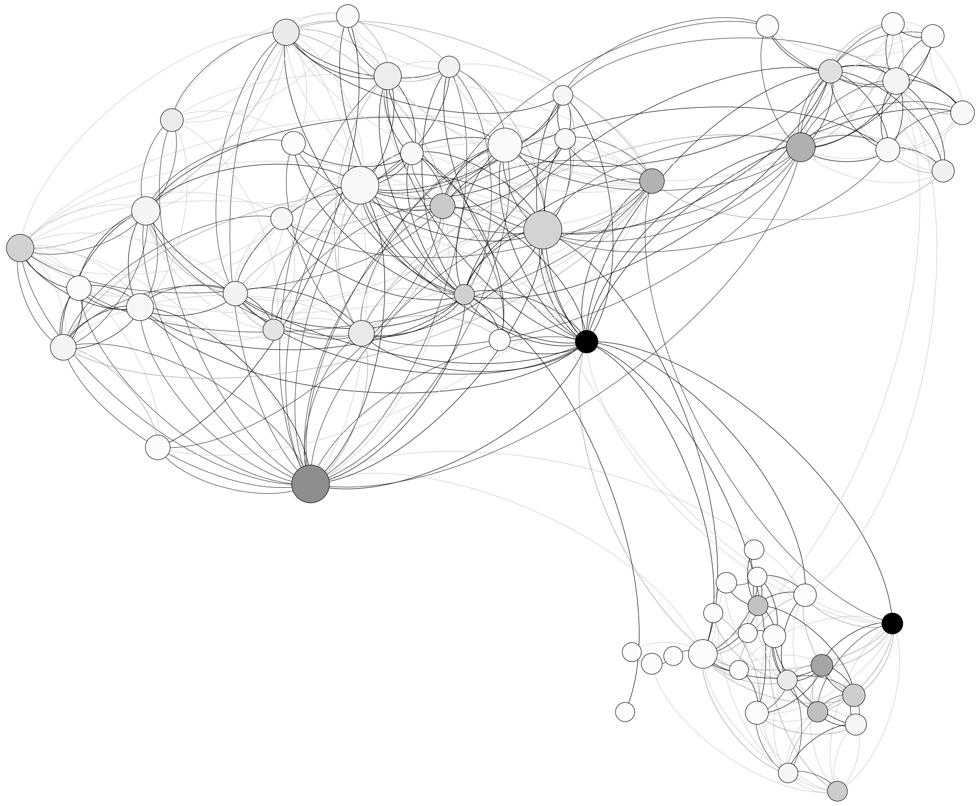


Figure A.3.: Access-Network[4.5(a)]

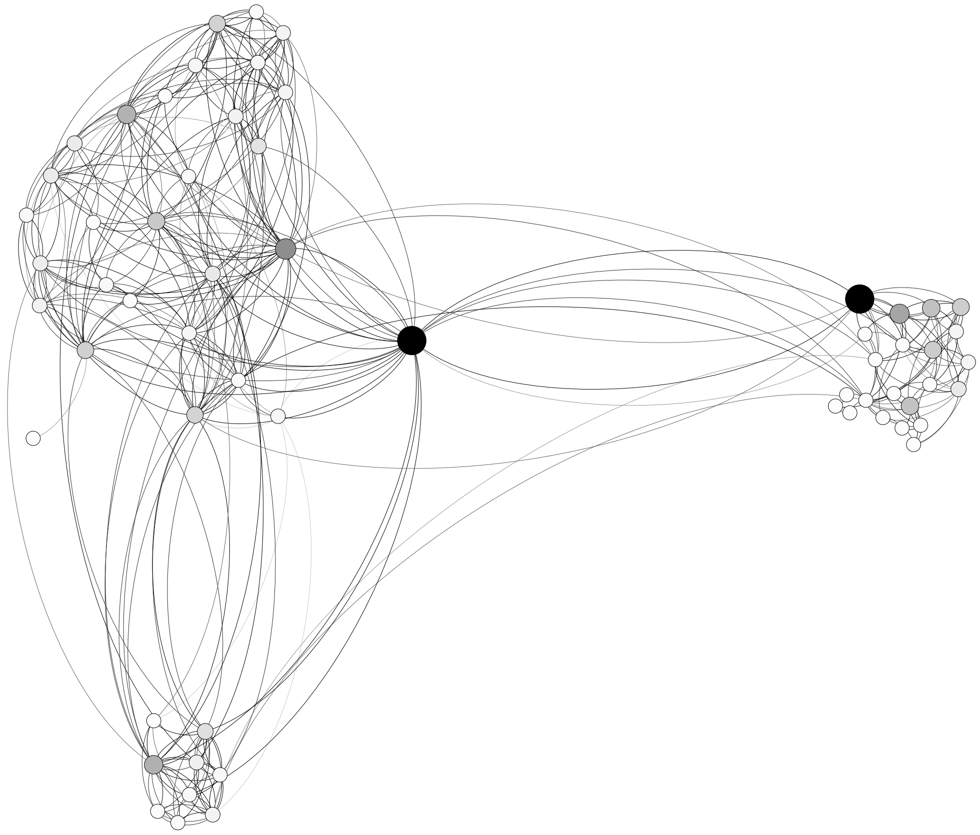


Figure A.4.: Capabilities-Network[4.5(b)]

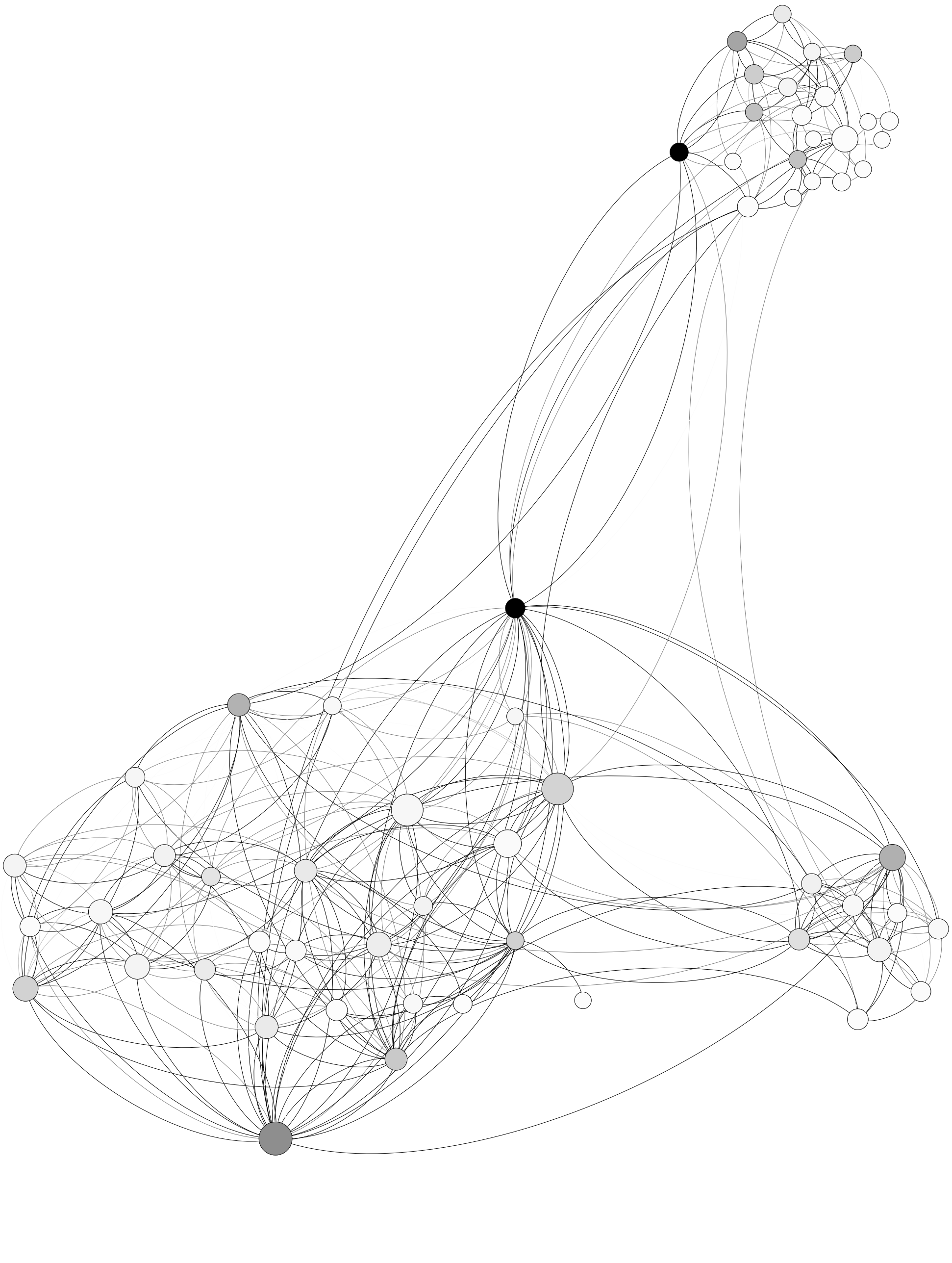


Figure A.5.: Engagement-Network[4.6(a)]

Figure A.6.: Communication-Network[4.6(b)]

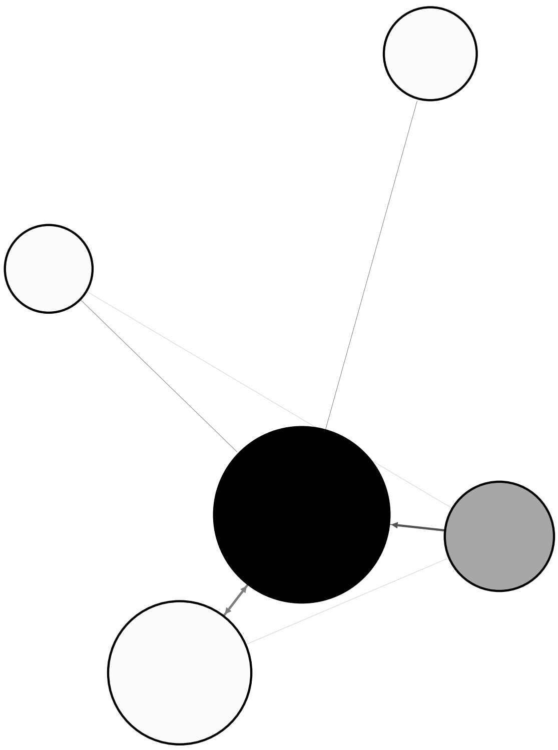


Figure A.7.: Access-Network mapped to communities[4.7(a)]

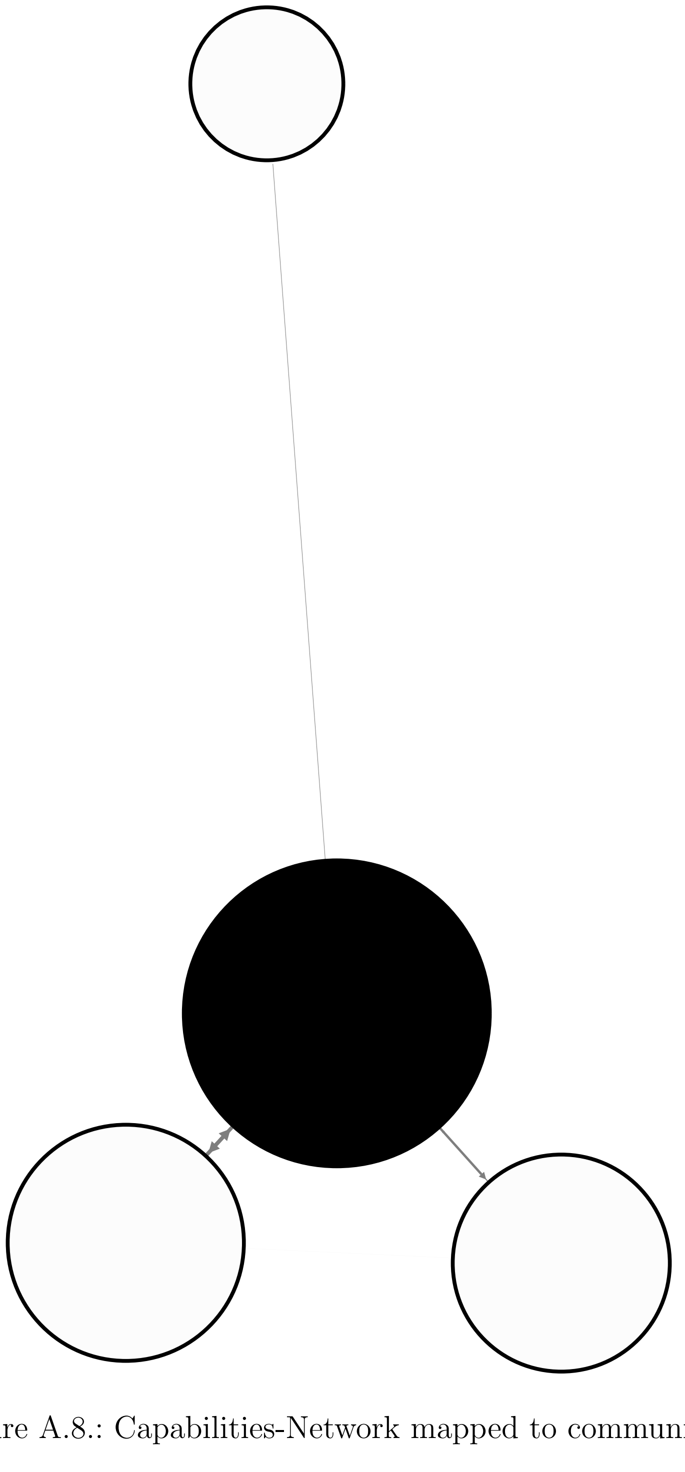


Figure A.8.: Capabilities-Network mapped to communities[4.7(b)]

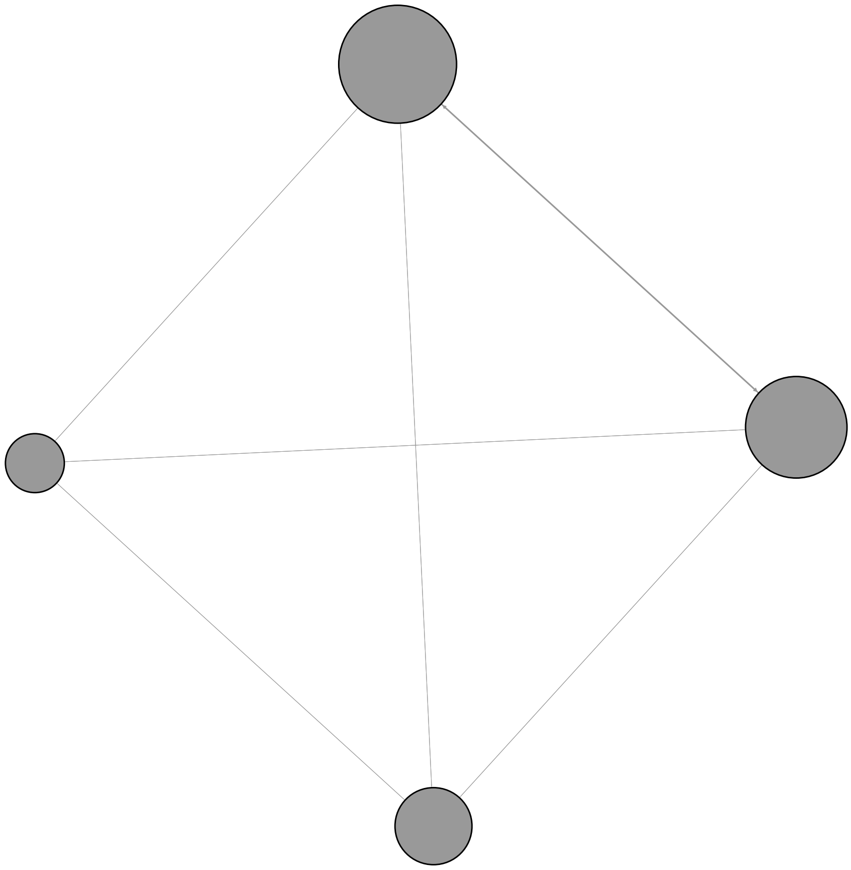


Figure A.9.: Engagement-Network mapped to communities[4.7(c)]

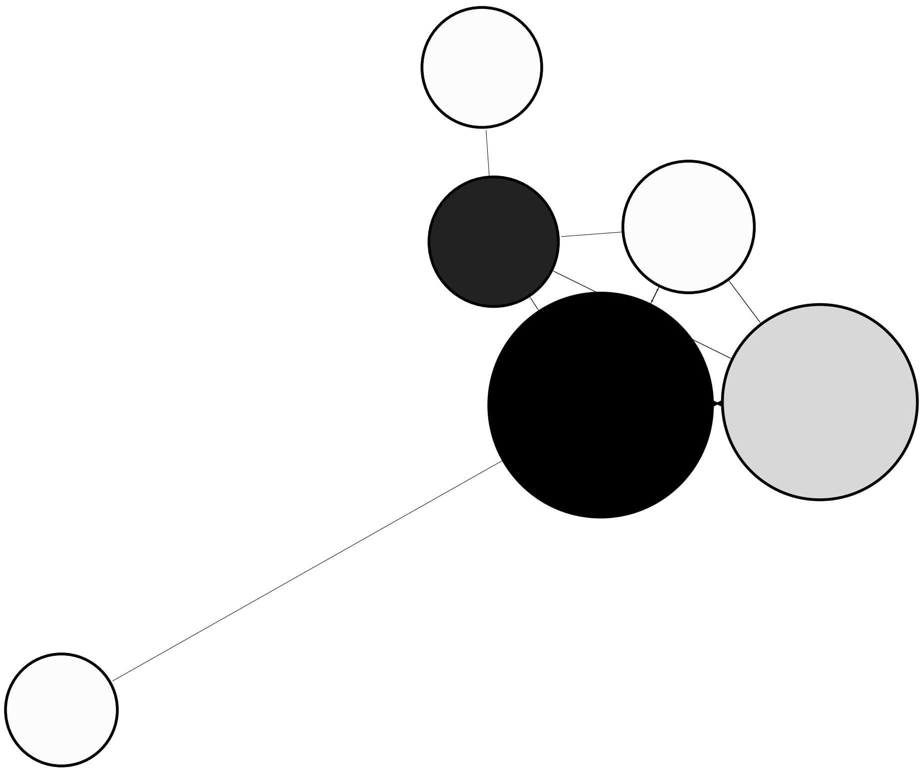


Figure A.10.: Communication-Network mapped to communities[4.7(d)]

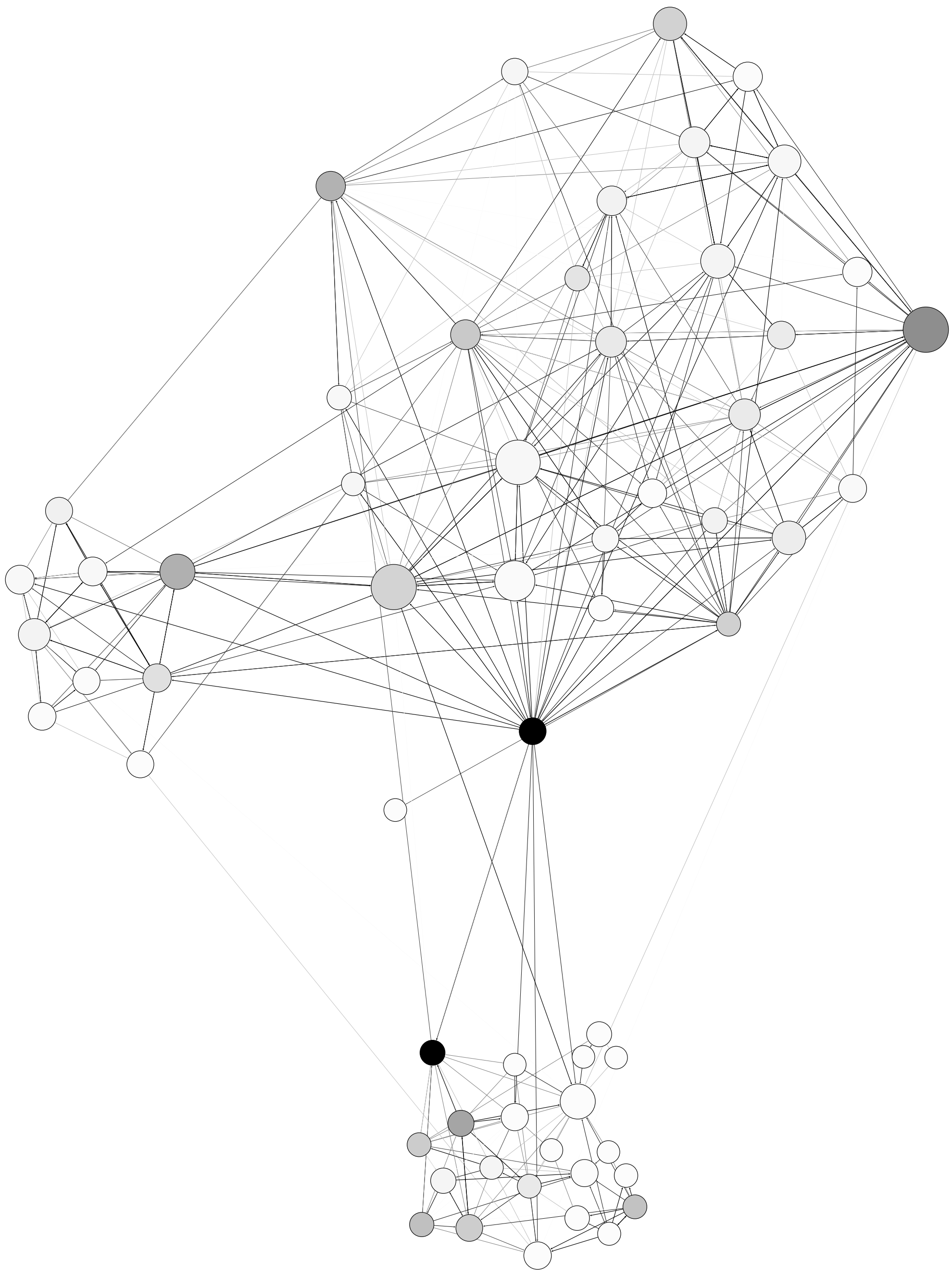


Figure A.11.: Combination of capabilities and accessibility[4.8(a)]

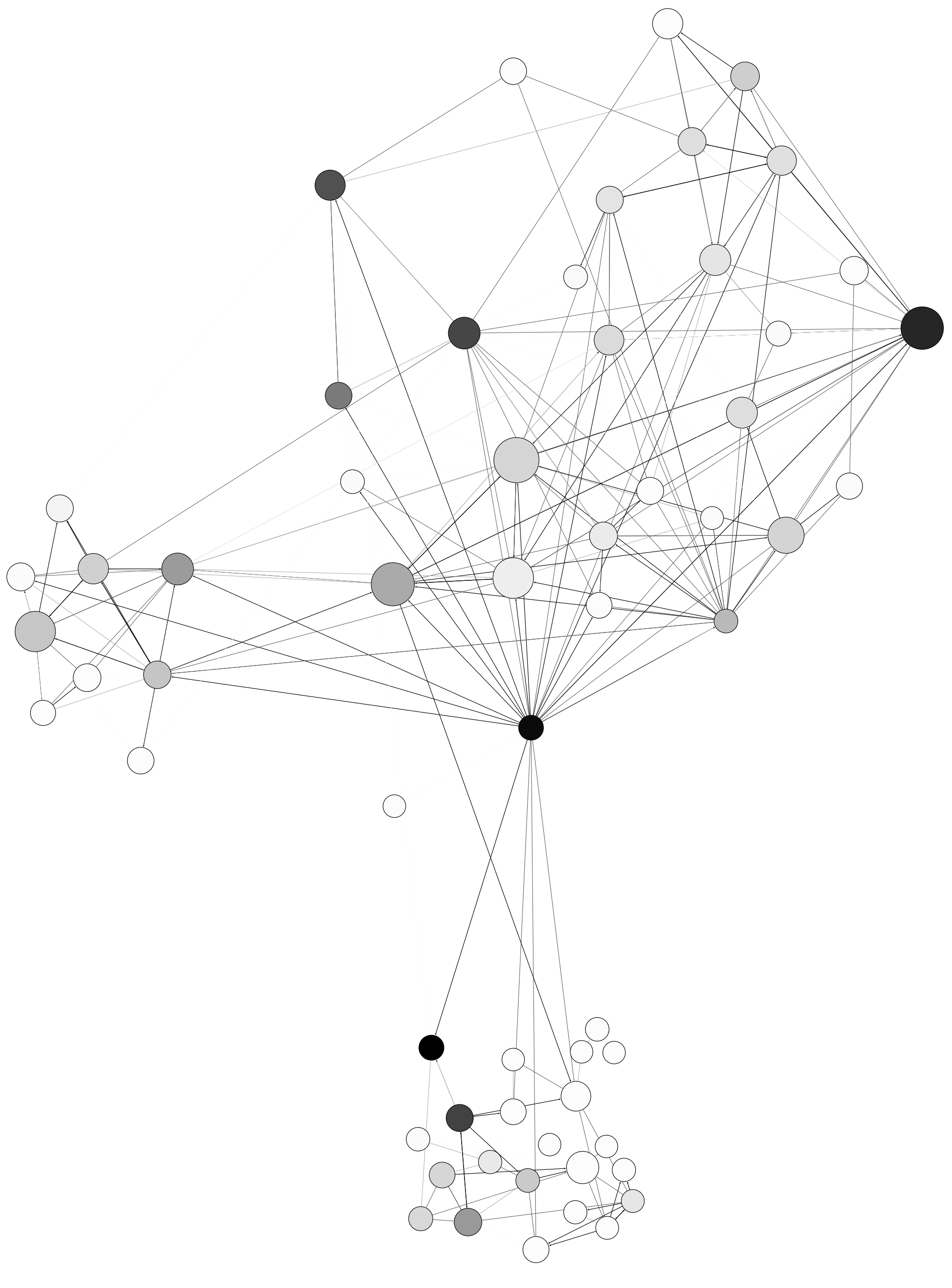


Figure A.12.: Capabilities and accessibility network restricted to most promising connections[4.8(b)]

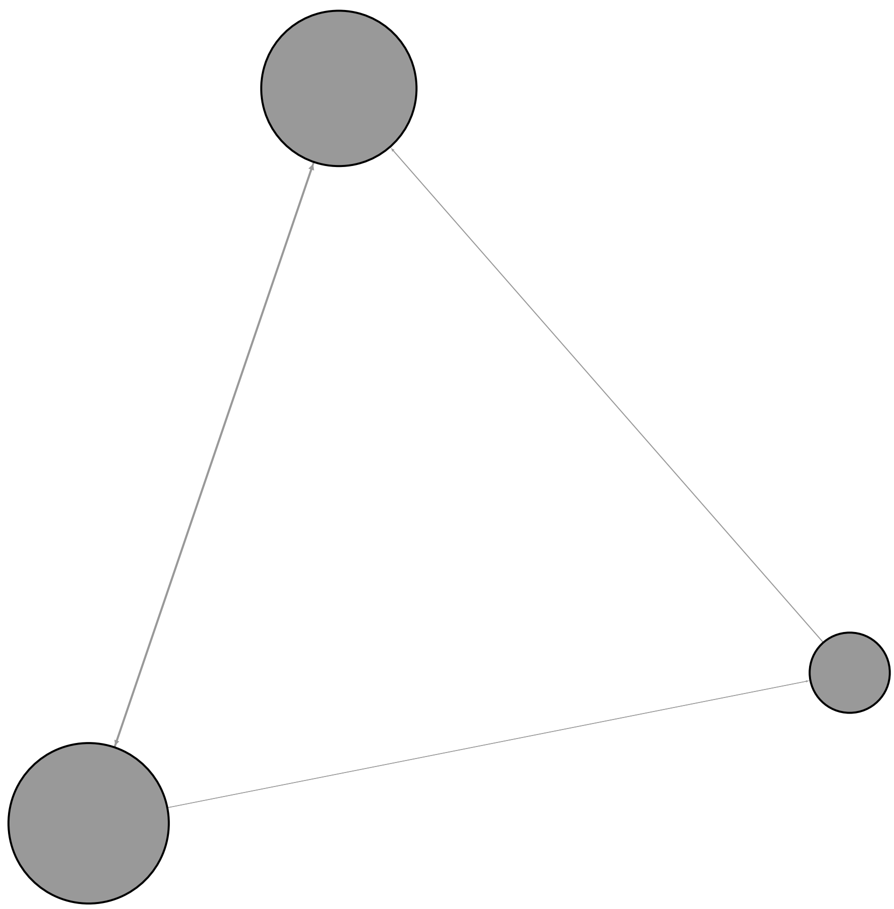


Figure A.13.: Capabilities and accessibility network mapped to communities[4.9(a)]

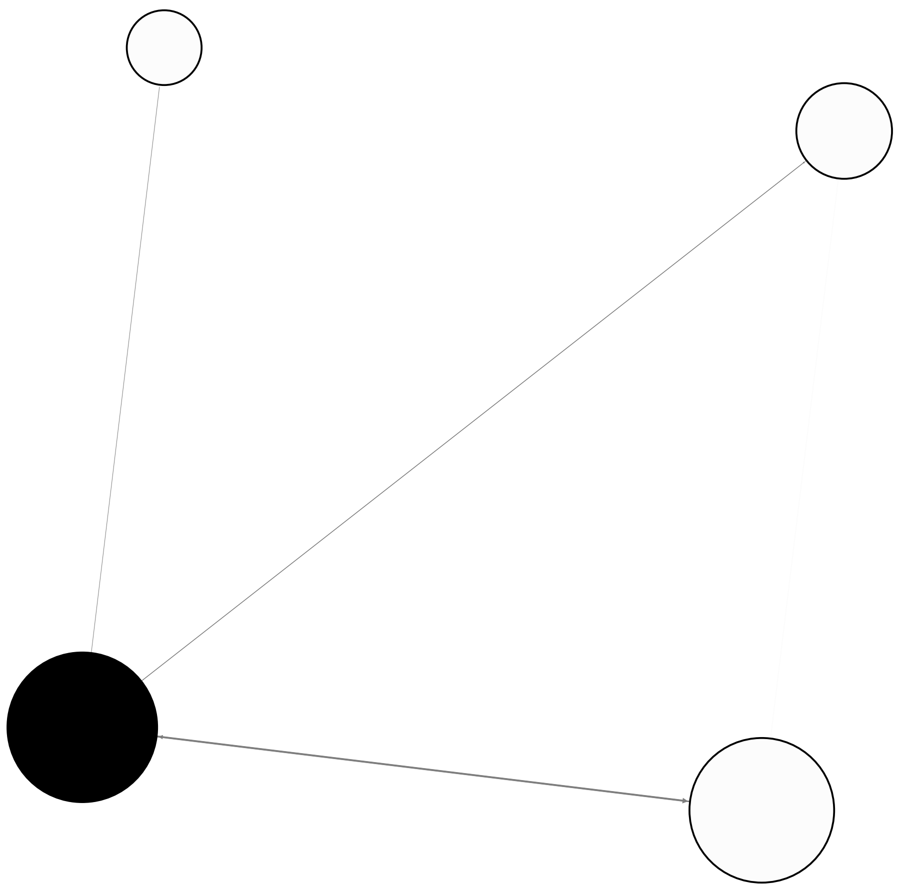


Figure A.14.: Capabilities and accessibility and engagement network mapped to communities[4.9(b)]

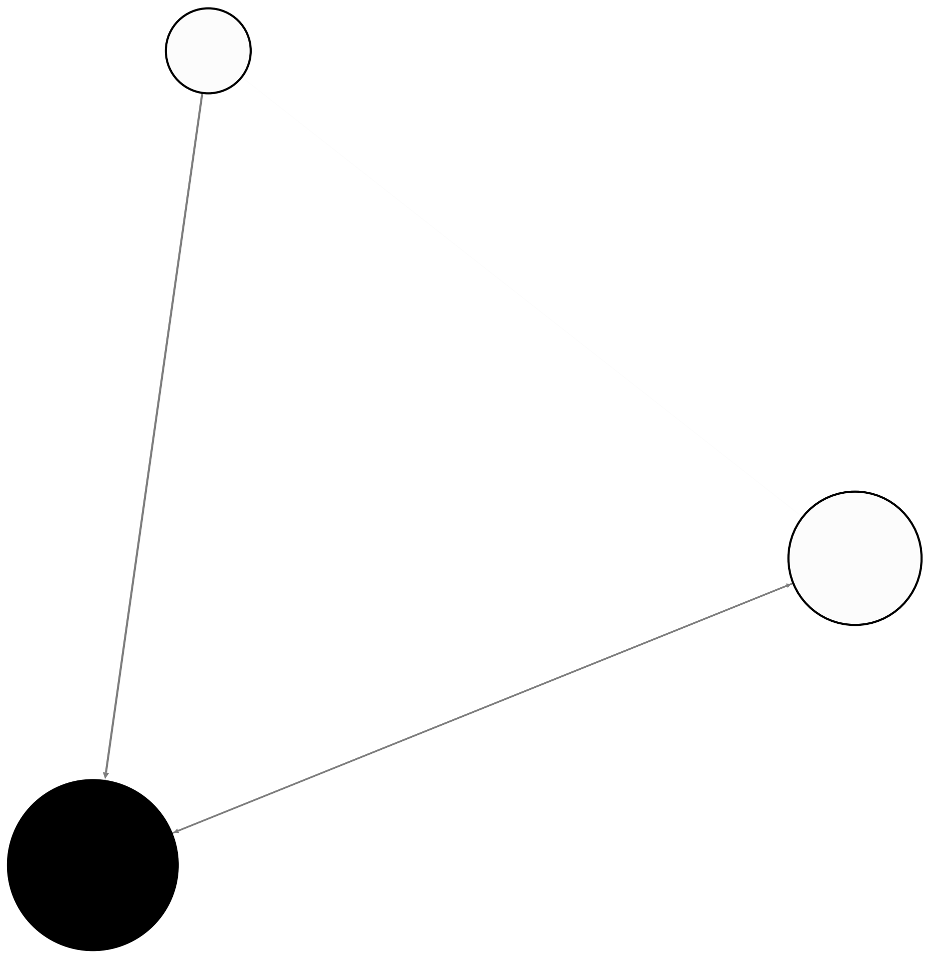


Figure A.15.: Capabilities and accessibility and engagement and communication network mapped to communities[4.9(c)]

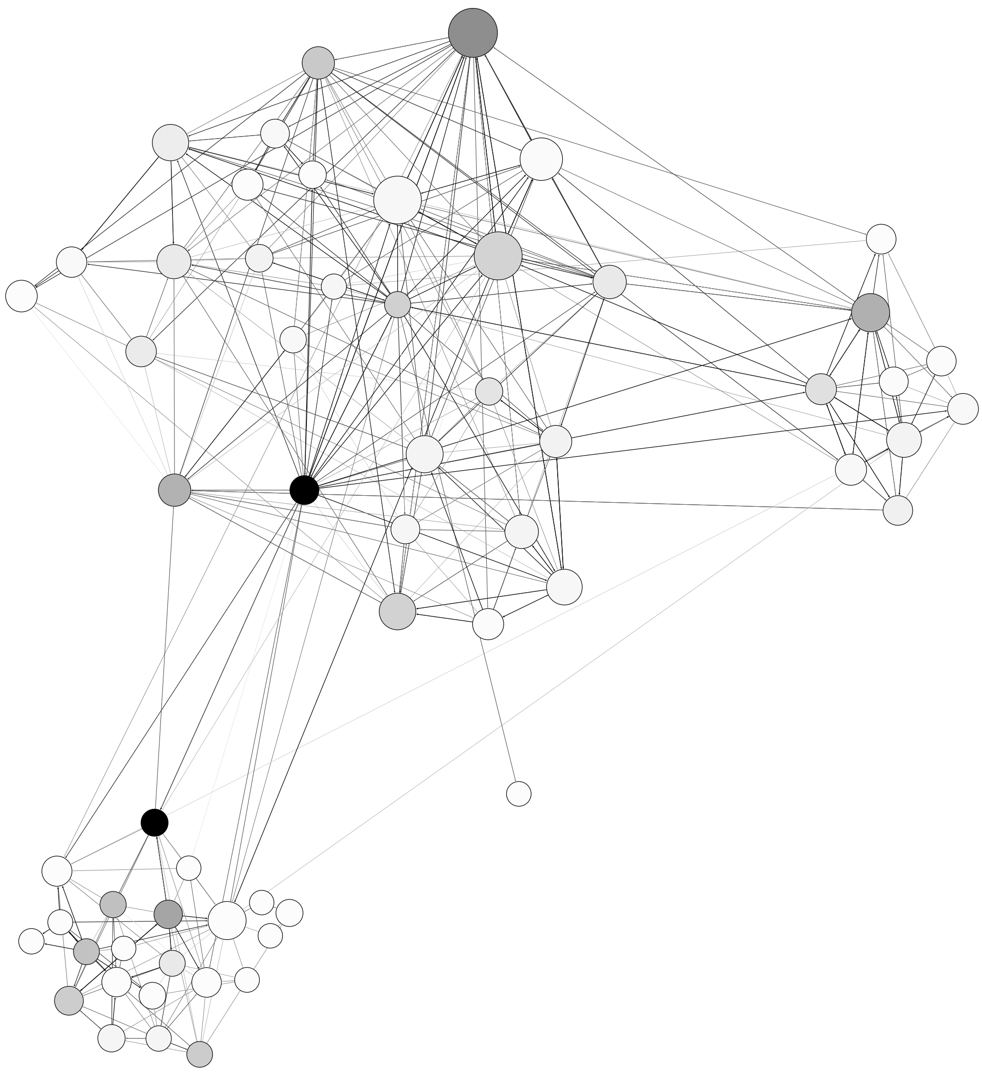


Figure A.16.: Capabilities and accessibility and engagement network[4.10(a)]

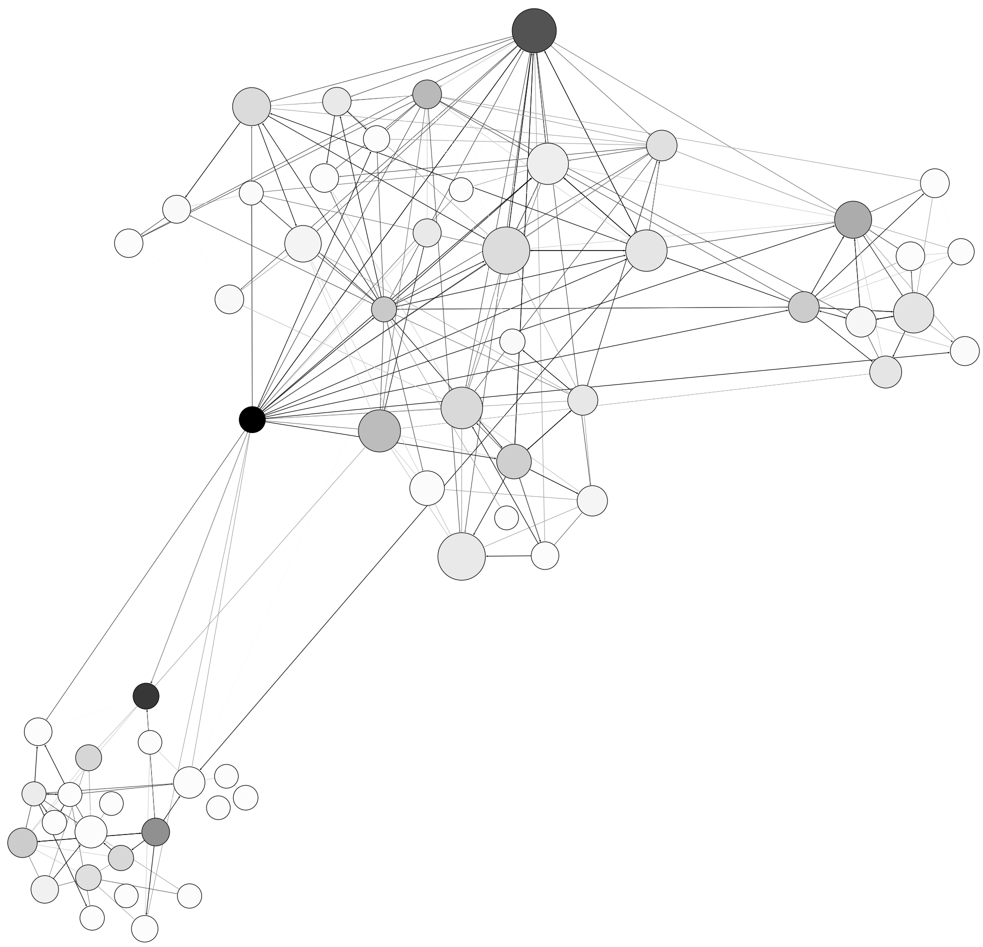


Figure A.17.: Capabilities and accessibility and engagement network restricted to most promising connections[4.10(b)]

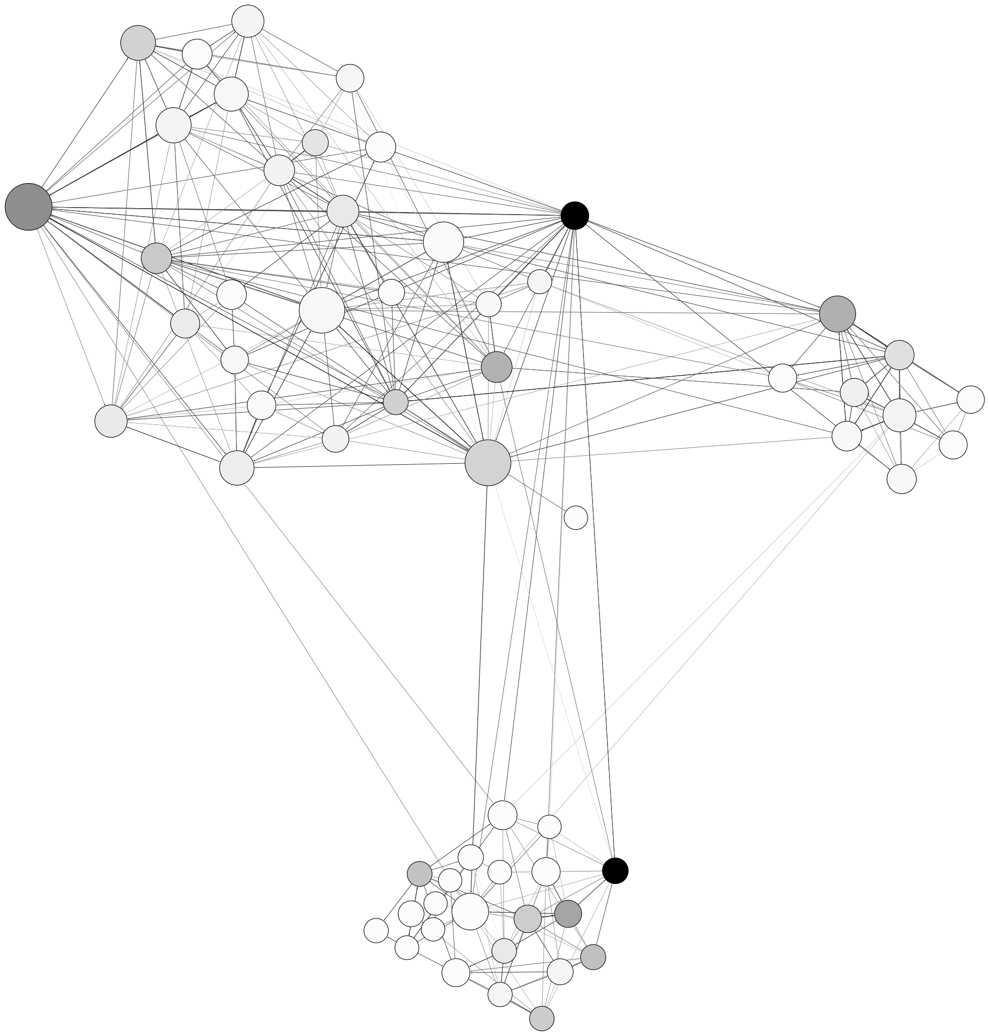


Figure A.18.: Capabilities and accessibility and engagement and communication network[4.11(a)]

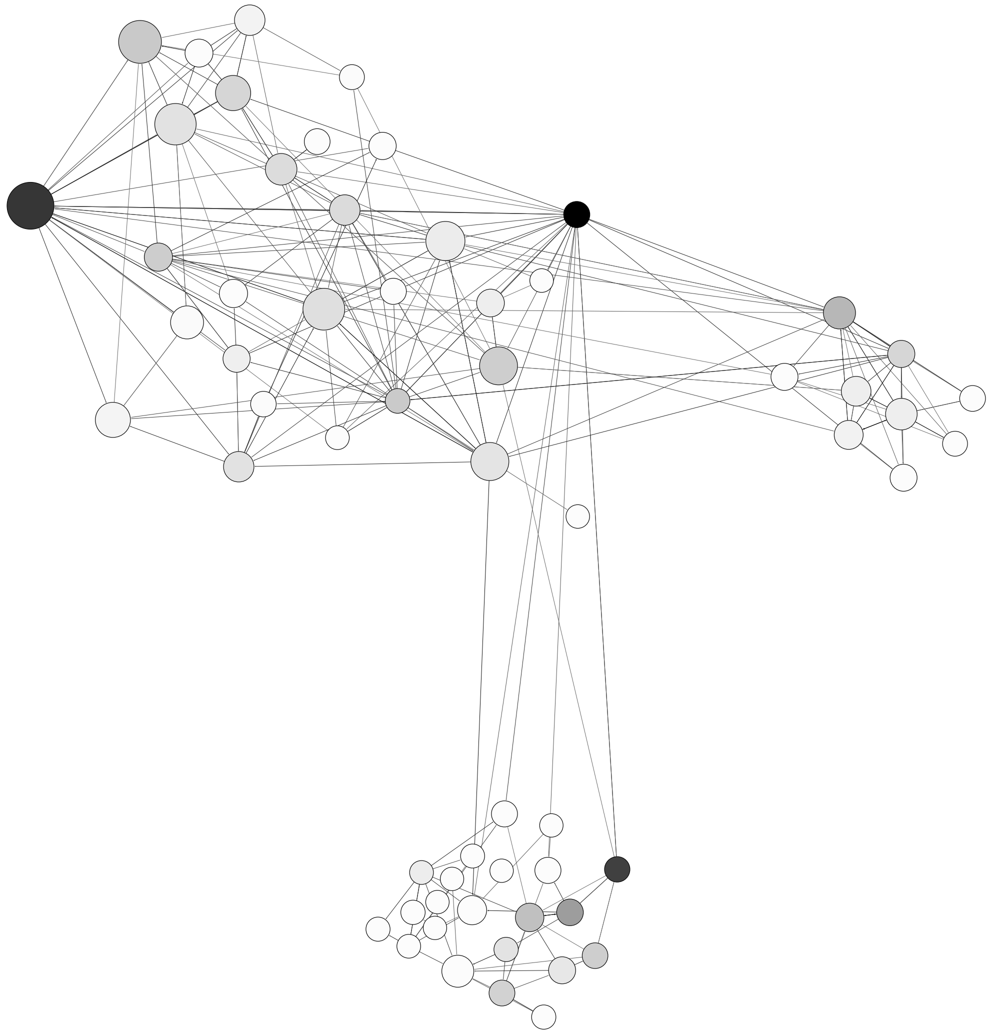


Figure A.19.: Capabilities and accessibility and engagement and communication network restricted to most promising connections[4.11(b)]

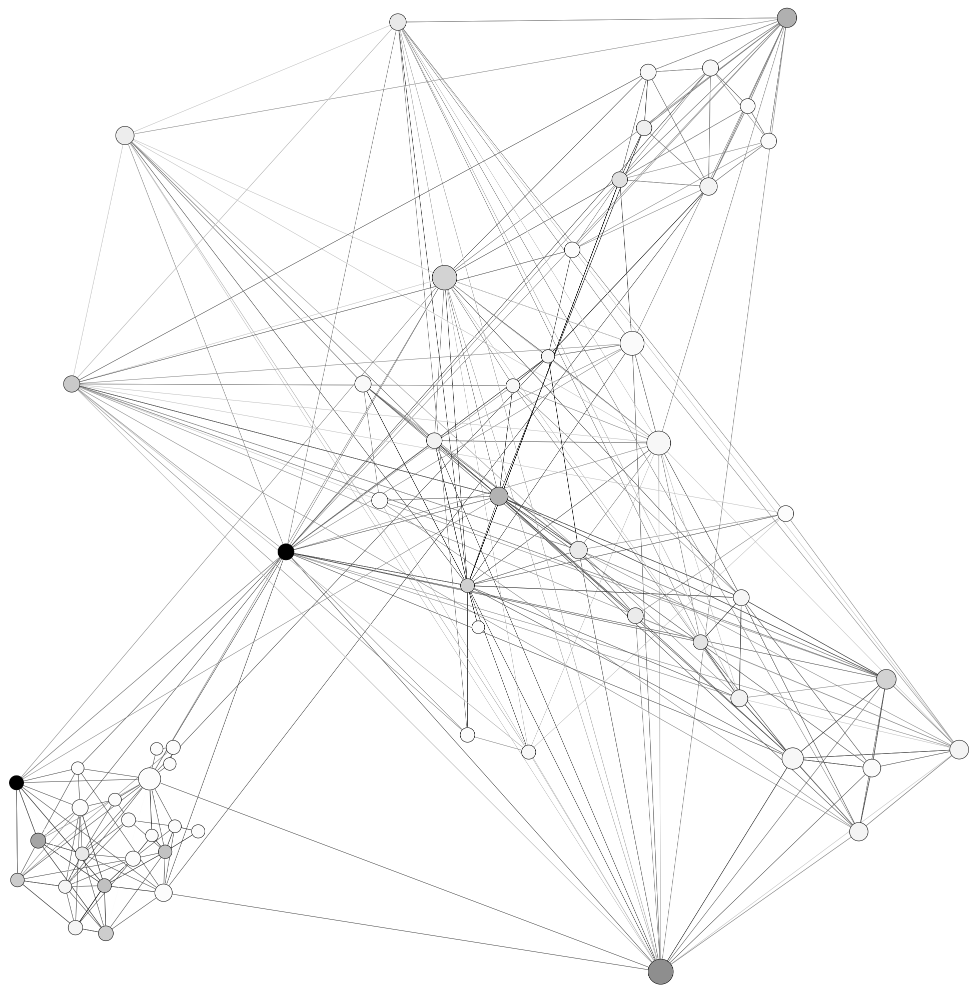


Figure A.20.: Basic-Network: who works with whom modified by time known[4.12(a)]

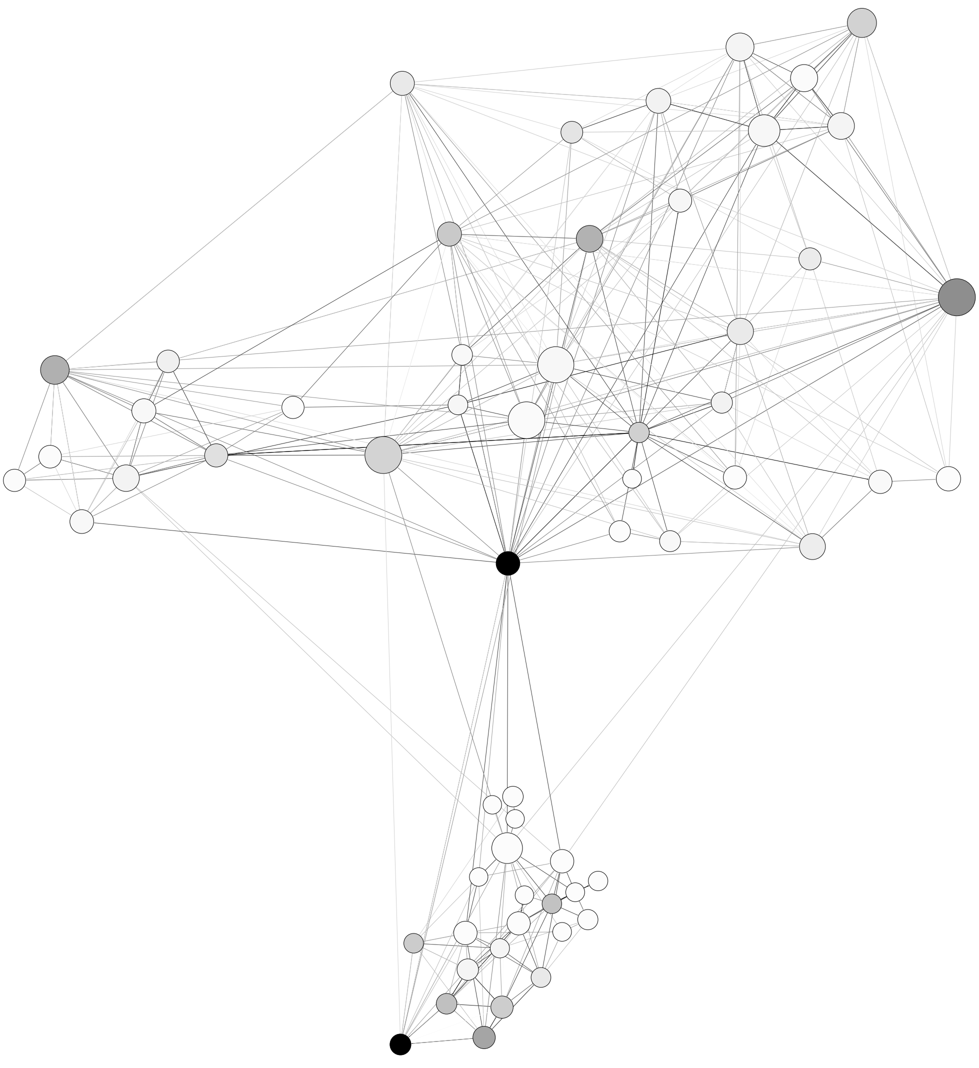


Figure A.21.: Access network modified by time known[4.12(b)]

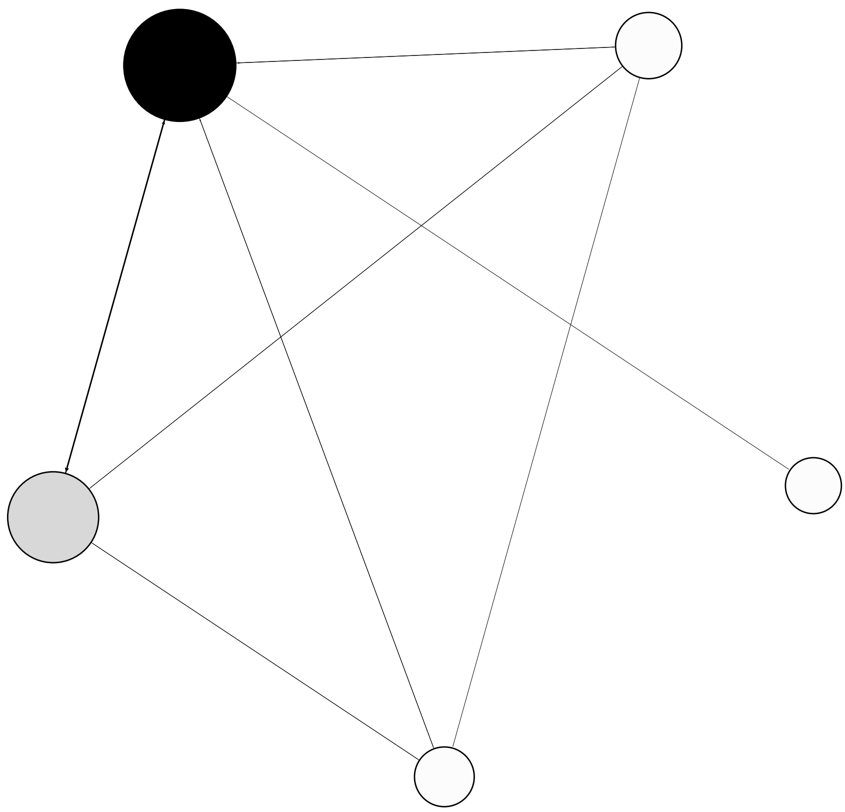


Figure A.22.: Access network modified by time known mapped to communities[4.12(c)]

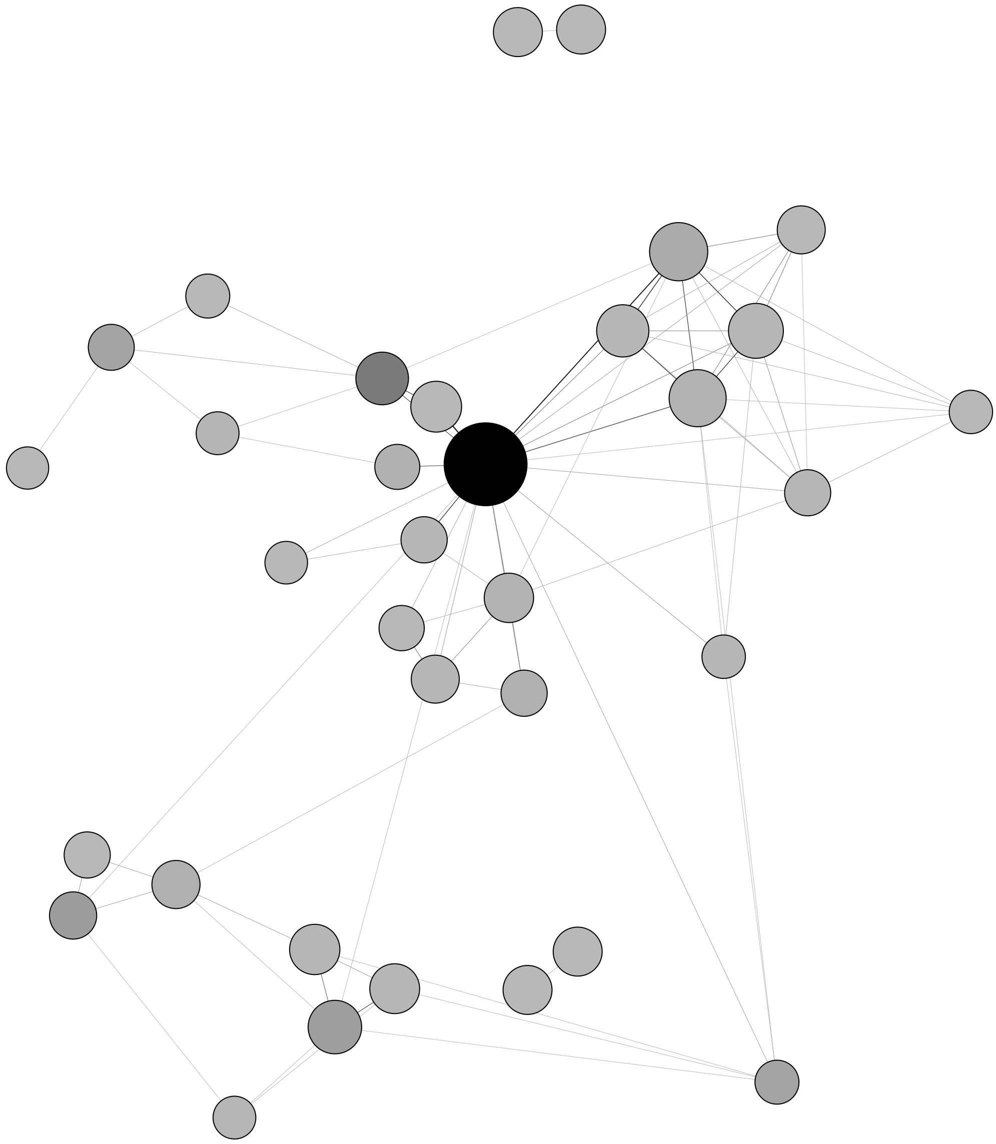


Figure A.23.: Co-Authorship network[4.13(a)]

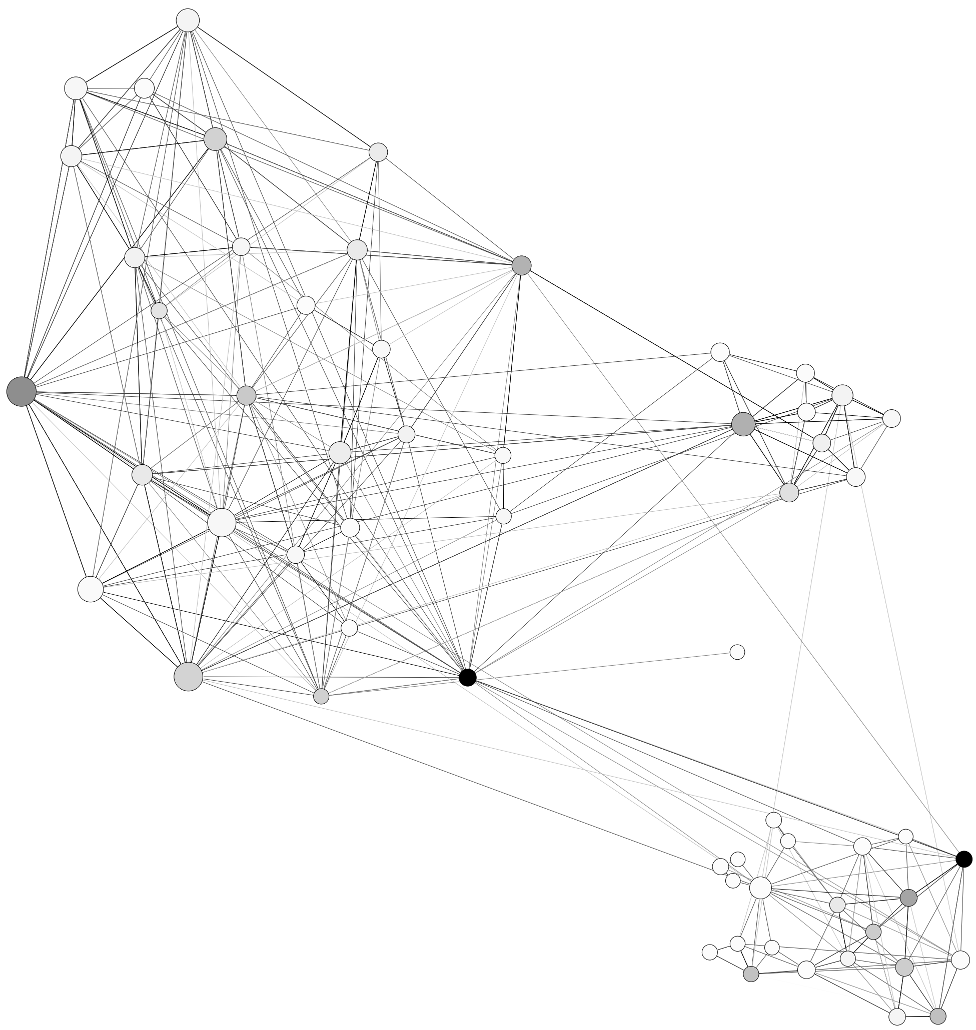


Figure A.24.: Basic-Network: who works with whom modified by amount of interaction[4.13(b)]

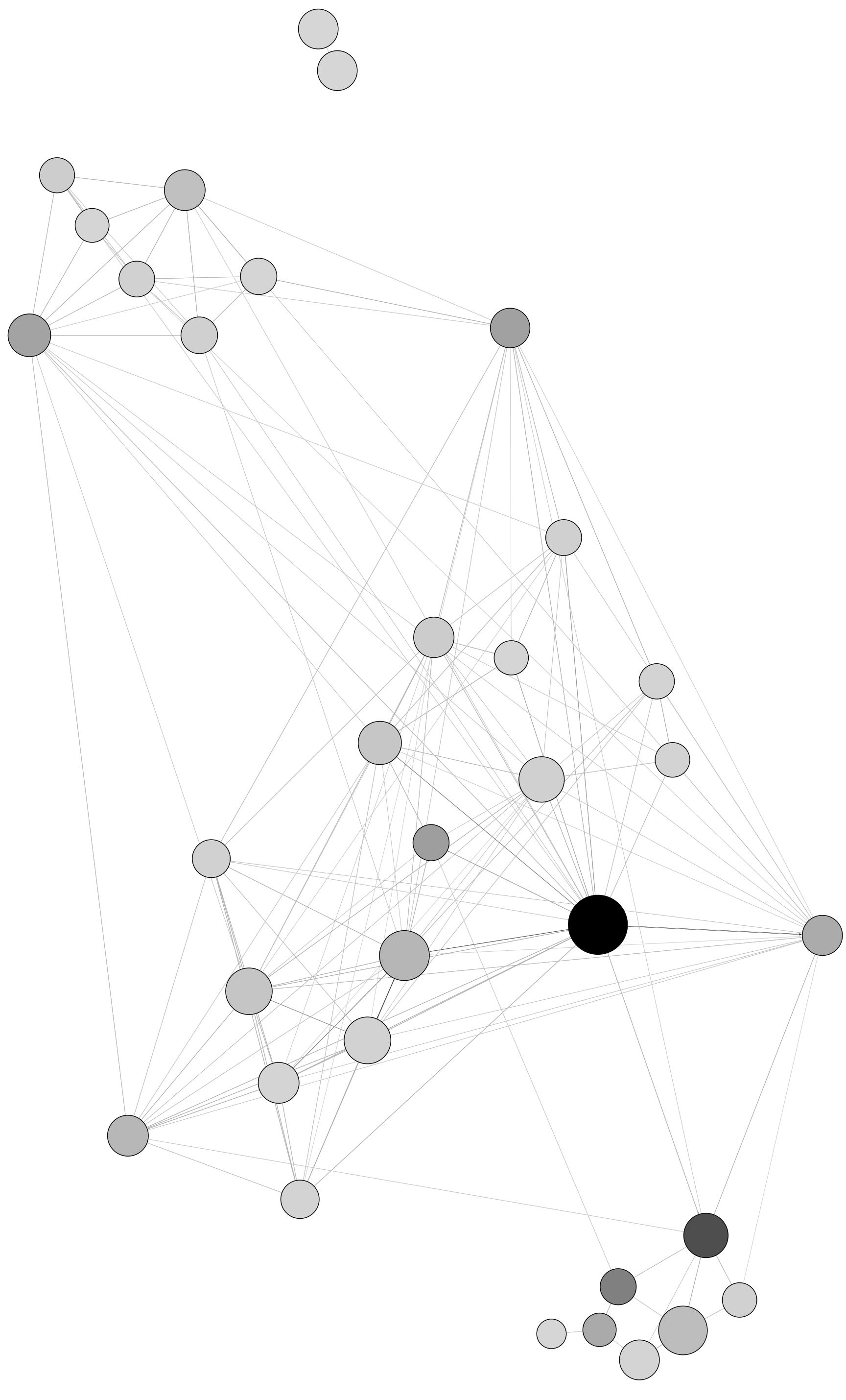


Figure A.25.: Combined network of co-authorship and the interaction-based modification of the basic network[4.13(c)]

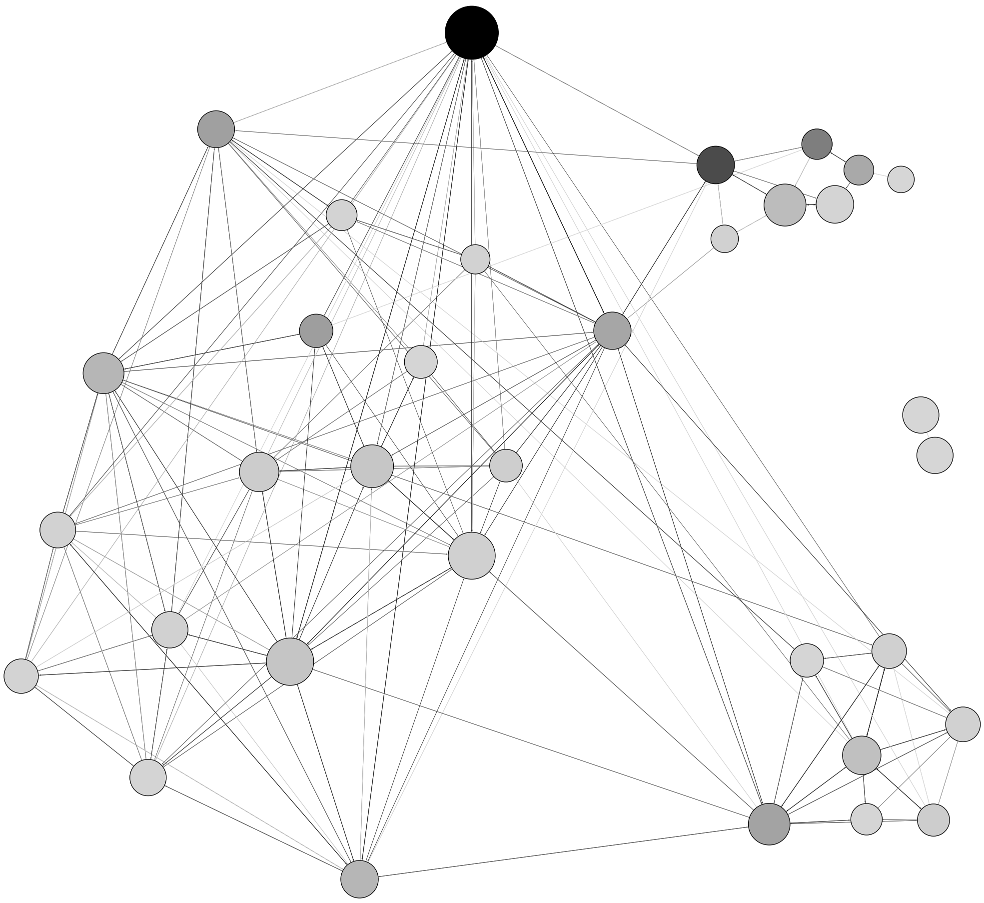


Figure A.26.: Combined network of the co-authorship and the four dimension network[4.13(d)]

Statement of Authorship

I, Wolfgang Schlauch, hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any degree at any educational institution, except where due acknowledgement is made in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

Signature: ....
Date: ....