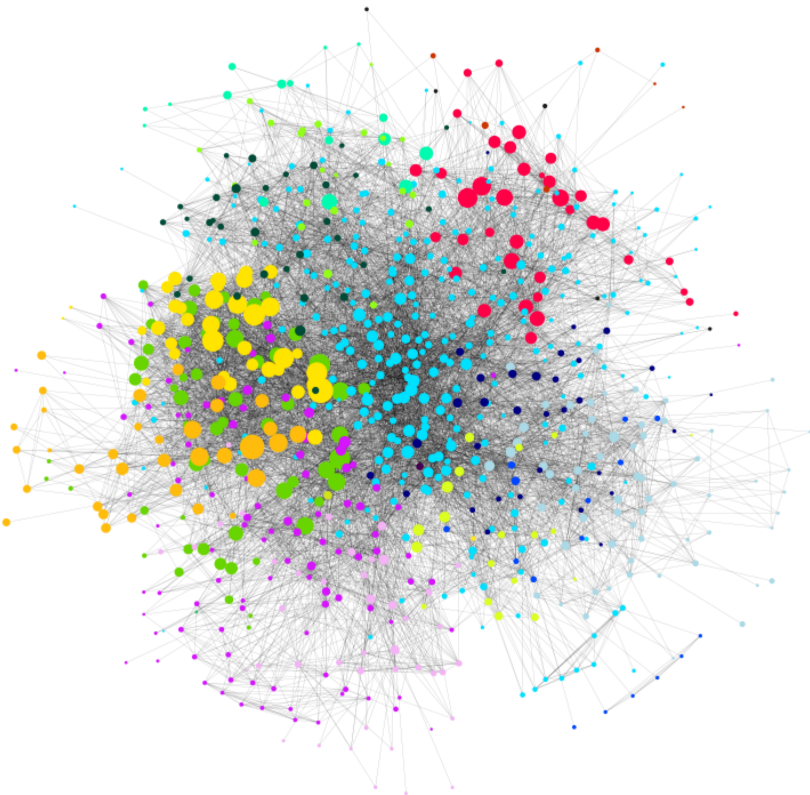
CHALMERS
UNIVERSITY OF TECHNOLOGY



A Network Analysis of a Company's Internal Email Communications

Master's thesis in Complex Adaptive Systems

KLAUDIA MUR

Department of Space, Earth and Environment
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2021

Master's thesis 2021:NN

A Network Analysis of a Company's Internal Email Communications

KLAUDIA MUR

Department of Space, Earth and Environment
Chalmers University of Technology
Gothenburg, Sweden 2021

A Network Analysis of a Company's Internal Email Communications
KLAUDIA MUR

Supervisor: Claes Andersson
Examiner: Claes Andersson

Master's Thesis 2021:NN
Department of Space, Earth and Environment
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: The undirected network of 828 nodes that have sent at least two emails to each other. Colors correspond to locations and node size to total amount of messages sent.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2021

A Network Analysis of a Company's Internal Email Communications
KLAUDIA MUR
Department of Space, Earth and Environment
Chalmers University of Technology

Abstract

This thesis views organisations as complex systems where micro level interactions have consequences at the macro level. While these interactions happen at multiple levels, this thesis attempts to explore the internal communication networks and specifically the email interactions as the core venue for the micro-level interactions. A network analysis of the meta data of a company's internal email data of 90 days is done. Measurement on the network-, as well as location- and node level are performed, to examine connectivity, identify locations and nodes at the core of the network and those at the periphery. A special focus is on comparing nodes at different locations, finding specific patterns of communication and connectivity.

Keywords: social network analysis, digital trace data, organisational network analysis, communication networks, complex social systems

Acknowledgements

As any other work, this thesis is not a lone undertaking, but could not have happened without input, ideas, resources and knowledge that I have gained from my network. I would like to thank everybody that has taught me, inspired and supported me during the course of my education, and in particular during the last few months for the work on this project. I would like to thank my supervisor Claes Andersson for making it possible for me to pursue my own project. Thanks to everybody who took their time and expertise to discuss the topic and the idea with me. Thanks to Birgit Oberkofler and Valentina Thurner.
Thanks to my family for their support, and especially my mum Sabina for being an important hub in my network. And a special thanks to Faheem, for all the discussions and inputs and for always believing in the idea and in me.

Klaudia Mur, Gothenburg, June 2021

Contents

List of Figures

1

Introduction

Social network theory sees individuals as embedded into webs of relations. The structure of those networks as a whole and the position in it matters for individual, team and organisation. The communication network serves as a map of social networks inside the company. Since a significant amount of communication is happening digitally, it is possible to collect and analyse meta data of digital traces. This thesis gives an introduction into social network theory and its applications to organisational research, both from a social science and a physics/complex systems perspective. It further serves as the report of a network analysis done with the internal email data of a company. The analysed data set contains information about sender, recipient and time stamp of every email sent internally by the 933 employees at 16 locations actively using email in a time frame of 90 days. This data is transformed into three networks. First, the directed and weighted network arising from the raw data is measured and described. Then, a undirected network is drawn from the data, with an edge between two nodes having exchanged at least two emails. Third, a network of tight communication is drawn, connecting nodes if they had email contact in at least 9 of the 13 total weeks. The focus of the analysis is especially on comparing locations, in terms of measurements both on the macro- and on the node level. Further, an analysis of connectivity is done, and one of communication between departments.
As expected, nodes mostly form tight cluster within their locations, while several connections also span between locations. While the overall network is quite well connected, three locations stand out, because of high cohesion, clustering and density, while in terms of centrality, nodes of the headquarter are most important.

2

Theory

The scope of this chapter is to explain why social network analysis is relevant to companies, and why analysis of communication is a relevant way of doing that. The first section explains shortly why companies can be seen as complex systems with emerging properties, or rather, as systems where interacting individuals create patterns, functions and identity on the macro level. Networks are the structure on which these interactions happen. The section 2.2.1 explains some of the mathematical properties of networks, grounded in graph theory. Social networks are a set of people that are connected in some significant way. Going further into social network theory, section 2.2.2 describes those ties. It explains types of ties, strength, emergence and types of networks, both between people in general and individuals in a company in particular. Section 2.2.3 zooms out from the focus on single ties and looks at the topology of the network as a whole. It describes three main characteristics of real world networks: scale free degree distributions, clustering, and the Small World model. The remaining sections of this chapter explain why looking at network structure matters: because the network is the structure on which social processes happen; and the structure influences the dynamics and outcome of the process. Section 2.2.4 illustrates the principles of three types of processes, namely contagion-like spreading of beliefs and opinions, the emergence of norms and culture, and knowledge and innovation. What the processes all have in common is that information flows between individuals, rather than from the top down. Through those processes, macro-level patterns and functionalities emerge through the interaction of individuals on the micro level. The last section takes a closer look at network position of individuals, and how those positions affect their access to information and resources, their performance, creativity, well-being and engagement. The section also explains several centrality measurements. The chapter concludes by explaining why the communication network is more than just a flow of information.

2.1 Social Emergence and Complex Systems

In what ways are companies complex social systems with emergent properties? Condorelli [1] describes them in the following way:

> ”Complexity Theory promotes a connectionist and anti-reductionist perspective. From this point of view, the interactive relationship does not simply unite the parts like in an aggregate but connects and mixes them up in a super ordered whole. In other words, they become a system in which and through which components are connected to each other and

are considered a totality rather than separate entities.”

The interaction of components on one scale can lead to complex global behavior on a larger scale that in general cannot be deduced from knowledge of the individual components. Networks as the structure on which this interaction happens are crucial. The company is a system that is created because individuals, with their individual knowledge, tasks, ideas, and properties interact and communicate with each other. The company can be seen as a higher order system in three ways: One, it is an entity on it’s own, second, it shows self-organisation, and third, it shows macro-level functionalities. Further, this system also provides a sense of identity to the individuals, as described in the next sections.

2.1.1 Entity

Companies are to some extent entities on their own. They are legal personas, can sign contracts and take actions. They are seen and talked about as an entity by people inside and outside - for example, "The bank of America is unhappy with the Federal Reserve to reduce rates" [2]. Here, both the *Bank of America* and the *Federal Reserve* are talked about as they were individuals with agency, feelings and beliefs. Companies - or social systems and groups in general - do however not have a brain or consciousness - rather, the decisions they make arise from collective processes involving management and employees - the people on the micro level. However, the company as an entity is to some degree independent of the set of members - even when employees leave, or even when the whole board resigns and is replaced - the company still exists. One way of looking at this is that the higher-level entity arises from a group of interacting people with shared goals and beliefs [2].

2.1.2 Self-organisation

> ”Systems in which organized behavior arises without an internal or external controller or leader are sometimes called self-organizing.” [3]

In other words, self-organizing systems show patterns at the macro level that arise from the interaction of agents at the micro level. The agents have only local information, and nobody is planning the overall pattern. In many ways, companies are not self-organising.

Comparing a company to an ant colony makes this point clearer. In an ant colony, complex patterns and organisation do not arise from central planning, but from quite simple rules, communication and interactions at the micro level. There is nobody controlling the whole operation, nor does anybody have full information over the state of the colony. Furthermore, no ant can reach out to every other ant - they can communicate only locally, with other ants nearby. Still, the amazing functionalities and properties of the ant colony make them a highly functioning and extremely resilient system. Of course, the macro level properties of the ant colony or the rules and interactions between ants on the micro level are far from just being random or arising by coincidence, but are shaped and created by evolution [3].

In a company, on the other hand, information is collected and used on the macro level. Examples are, how many employees are working on what, what is needed to

produce the product, or how much is about to be sold in the next three months. Based on that global information, informed central decisions are made and actions are taken, and the future is predicted and planned. Growth in terms of hiring new employees or new products and sales is planned carefully, by taking not just global information of the company as a whole, but information about the environment outside the company as well into account. Further, collaboration and communication is, to a certain extent, centrally planned. Collaboration does not just happen because two employees meet on the hallway and decide to tackle a problem together. Who should work with whom on what is planned and optimized; the company has a clear hierarchical structure. Also, it is fairly easy and cheap to reach out to every employee, through mail or phone, and collect or distribute information centrally.
However, in some ways, companies as social systems do show self-organisation. Many processes and patterns arise from micro-level interaction, without being centrally planned or even controllable. It can be attitudes towards a certain topic, or the usage of a new tool who's application spreads through the network. Important here is social influence - individuals base their decisions and beliefs on other people's decisions and beliefs. Belief dynamics as an example of this will be described in section 2.2.4.1.
Related to that idea is the idea of the informal structure, "the company behind the charts". Connections between people, and the networks that emerge from the set of those connections, do not always arise in a planned manner. The structure of these networks does however have influence on how the company as whole, the individual and the team function [4]. Individuals usually do not have global information, and they therefore do not know how those networks are structured or what kind of impact they have. Authority, influence, access to resources and information is not always connected to the formal, planned structure - but rather the informal one. Another aspect is that social connections are multiplex - people have different sets of relations to each other. For example, getting a job in the company or getting promoted is more likely if the person knows somebody in the company, or has worked with somebody [5]. Employees from one department ask employees from another department for advice or a favour to get a task done faster. Those are examples for how one type of relation can influence another type of relation between two people. And sometimes, collaborations or new ideas do arise because people met casually in the hallway or by the coffee machine. This, on the system level, leads to both positive and negative outcomes. It can make collaboration more efficient, but also hinder it. Again, many of those processes happen because individuals interact at the micro level. They are not planned, nor controlled or even fully known about, on a global level.
Looking at the communication network is one way of understanding those informal networks better. The sections 2.2.2 - 2.2.4 will look into how those connections arise, what influence they have, and what kind of patterns emerge on the macro level.

2.1.3 Functionalities

On top of showing some self-organisation, companies also have many macro-level functionalities. The individuals do not just meet and interact randomly, but because they share a common goal.

> ”A social group is a certain functional structure and individuals play a role in instantiating that structure” [2]

Page [6] distinguishes between emergence and self-organisation. While self-organisation leads to patterns or forms, emergence leads to functionalities that exist on the macro level, but not on the micro level. Examples are consciousness, culture and social cognition.
An example is guiding a ship into port [7]: no single person has the ability to do that. This ability is a property of the network of communicating individuals.
In a similar way, the company is a system of interacting parts with a shared goal: producing an outcome together. The development, innovation, production and selling of a product is a collective process that cannot be done by one individual. It is more than just the sum of individuals' work and efforts: the functionality exist on the macro level, but not on the micro level. In other words, the group-level properties arise and are different from just the aggregate of the individual properties, because individuals interact and communicate with each other. Communication is then not just an add-on, but a crucial component of the functionality. Not just the content of communication matters, but also the structure - the network - on which it happens, because it is crucial for the higher-level functionality.
Other examples of functionalities are creativity, innovation and new ideas. Very often, innovation does not happen because an individual had an genius idea, but arises because people with different backgrounds or views interact with each other. A well-studied example are structural holes, which will be described more closely in section 2.2.5.5. Employees at specific positions in the network, between groups of highly interacting individuals, do often have great ideas, because they are exposed to different viewpoints and topics. They are not only smart and creative themselves - this capability to come up with good ideas is correlated with the network position they are in, and the flows of ideas and information through the network. Again - interaction and the network structure matter. Innovation is a collective function, and not an individual.
Another important point is that the functionality is uphold even when the configuration of the company itself changes. For example, when one employee leaves the company, a new employee, or an internal reorganisation, will undertake her tasks and her contribution to the macro level functionality. Also, the company as a whole has to adapt all the time, to internal as well as external stresses, and to a changing landscape [8]. To be able to keep up the functionality (producing and selling a product), change and adaption are crucial, for example by innovating new products, introducing new sales methods, or adapting to new forms of internal communication.

2.1.4 Identity

Companies are however not just functionalities. They are social systems as well. Humans are social beings - to be able to do their task, to be motivated and in order to share the common goals, the company has to offer meaningful connections and a sense of community. Networks are formed whose purpose and function transcends the pure collaboration for task execution.

Network theory offers a insightful view on identity: to be part of a system, an individual has to communicate with other parts of the system on a regular basis. The network, and the system, only exists because constant communication takes place from the bottom up, connecting individuals. Closely connected is the idea of structural cohesion, described in section 2.2.1.9. A sense of identity, in turn, leads to shared goals, motivation, and meaningfulness [9] (see section 2.2.6).

2.2 Social Network Theory

A network is a set of items, which are called nodes (or vertices), with connections between them, called edges, or ties [10]. In social networks, the nodes are people, and the edges between them are social connections of some kind, such as friendship, communication, or collaboration [11]. Social network theory sees individuals as embedded in thick webs of social relations and interactions, and seeks answers to questions like, how autonomous individuals can combine to create enduring, functioning societies [12]. Networks capture the patterns of interactions between the parts of a system. The patterns of interactions can have a big effect on the behaviour of the system [11]. The following sections will give some examples on how the structure of the connections between them influences individuals, teams and organisation as a whole. While social network theory and analysis has existed for several decades, gathering of data used to be hard work, since individuals had to be interviewed or observed directly, which requires intensive contact with the group and makes it possible to only study smaller networks of tens to up to a hundred individuals. With communication and interaction taking place more and more digitally, and therefore leaving traces, a new way of collecting data is possible [13]. Analysing meta data of communication, i.e., not the content, but information about who communicates or interacts with whom, makes it possible to gather network data with much less effort and on a much larger scale.

What should be kept in mind is that a networks is a ”simplified representation that reduces a system to an abstract structure or topology, capturing only the basics of connection patterns and little else” [11]. Among others, social network studies do not capture features like detailed behaviours or properties of individual nodes, or the precise nature of the interaction between them.

2.2.1 Mathematical representations and properties of networks

In mathematical terms, a network is a graph: a set of nodes with edges between them, and can be analysed with methods of graph theory.

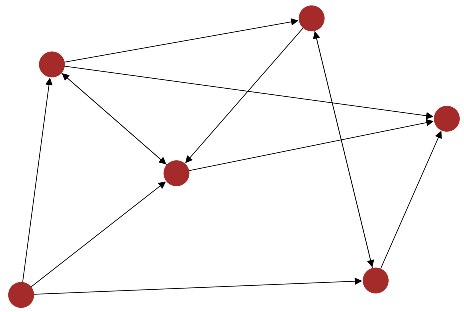


Figure 2.1: A directed network

$$M = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Figure 2.2: Matrix Structure

2.2.1.1 Adjacency Matrix

Given a network with n nodes, labeled with integer labels $1, ...n$, the adjacency matrix is defined as the $n \times n$ matrix with elements $A_{i,j}$ such that

$$A_{i,j} = \begin{cases} 1 & \text{if there is an edge between nodes } i \text{ and } j, \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

The adjacency matrix is a fundamental mathematical representation of the network [11].

2.2.1.2 Directed and Undirected Networks

A directed network is a network where edges have directions. Figure 2.1 shows a directed network and figure 2.2 its adjacency matrix.
The adjacency matrix of undirected networks is symmetric. Examples of undirected social networks are ones where the connection between people are mutual, e.g. friendships, or collaboration networks. Directed social networks are for example communication networks (a message is sent from A to B).
Another example for directed networks is the social network Twitter, where users can follow other users, but they do not necessarily follow them back. Being friends on Facebook, on the other hands, means a mutual connection. This social network is therefore undirected.

2.2.1.3 Weighted Networks

Weighted graphs are graphs where the edges are associated with a certain weight [14]. In the adjacency matrix, weighted networks can be represented with the elements $A_{i,j}$ equal to the weights of the edge between i and j [11]. Figure 2.3 shows a weighted network and figure 2.4 its adjacency matrix. Weighted networks are a representation of networks where the strength of a connection can vary, or for example flows between nodes have different amounts or frequencies, like in an email network.

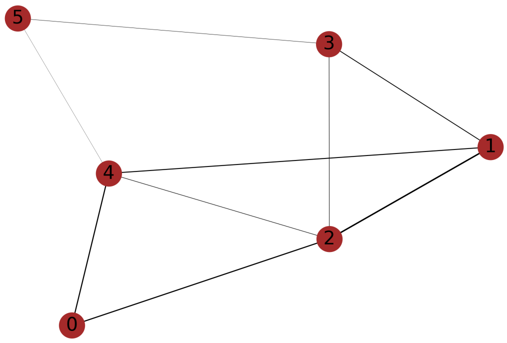


Figure 2.3: A weighted network

$$M = \begin{pmatrix} 0.0 & 0.0 & 0.7 & 0.0 & 0.7 & 0.0 \\ 0.0 & 0.0 & 0.9 & 0.5 & 0.6 & 0.0 \\ 0.7 & 0.9 & 0.0 & 0.3 & 0.3 & 0.0 \\ 0.0 & 0.5 & 0.3 & 0.0 & 0.0 & 0.2 \\ 0.7 & 0.6 & 0.3 & 0.0 & 0.0 & 0.1 \\ 0.0 & 0.0 & 0.0 & 0.2 & 0.1 & 0.0 \end{pmatrix}$$

Figure 2.4: Matrix Structure

2.2.1.4 Density

The density of a network is a measure for how many of the possible edges actually exist. The density for undirected graphs is

$$d = \frac{2m}{n(n-1)}$$

and for directed graphs

$$d = \frac{m}{n(n-1)}$$

, where n is the number of nodes and m is the number of edges in the network.

2.2.1.5 Paths

A path in a network is any sequence of nodes such that every consecutive pair of nodes in the sequence is connected by an edge in the network [11]. The length of a path in a network is the number of edges traversed along the path. A geodesic path, or shortest path, is a path between two nodes so that no shorter path exists.

2.2.1.6 Average Shortest Path Length and Diameter

The shortest path between two nodes is the length of the geodensic path between them. The average shortest path is therefore:

$$a = \sum_{s,t \in V} \frac{d(s,t)}{n(n-1)} \tag{2.2}$$

where V is the set of nodes in G, $d(s,t)$ is the shortest path from s to t, and n is the number of nodes in G. [5] The diameter of a network is the shortest distance between the two most distant nodes in the network. In other words, it is the longest of the shortest path lengths between all possible pair of nodes. Average shortest path length and diameter are defined for both directed and undirected networks and are an indication for the overall connectivity of a network. Small world networks are characterized by a low average path length. In networks with high diameter and average shortest path lengths, it takes many steps to reach the whole network.

2.2.1.7 Connected Components

An undirected network is connected if there exists a path between every two nodes in the network. Many networks are however not connected: they are split up into several parts that are disconnected from another, and there is no path between any two nodes in different components [11]. The components of a network are the distinct maximal connected subgraphs of a network [5].

2.2.1.8 Cliques

A clique is a maximal completely connected subnetwork of a given network. Cliques are generally required to contain at least 3 nodes, otherwise each link could potentially define a clique of two nodes. A given node can be part of several cliques at once [5]. The clique number of a node is the size of the maximum clique containing the node.

2.2.1.9 Structural Cohesion and Connectivity

A k-core of a graph G is a maximal connected subgraph of G in which all vertices have degree at least k. Equivalently, it is one of the connected components of the subgraph of G formed by repeatedly deleting all vertices of degree less than k. k-core decomposition deletes step wise the nodes with lowest degree from the graph, and is therefore a "well-established metric which partitions a graph into layers from external to more central vertices" [15]. For social networks, this can be used as a concept of social cohesion based on node connectivity [16]. If networks have more than one core and fall into several different parts at some point of the k-core decomposition, it means that their cohesion is not very good.

2.2.2 Ties in Social Networks

After describing some of the mathematical properties of networks, this section dives deeper into social networks, by looking at the ties that can arise between people, forming edges of networks.

2.2.2.1 Types of Ties

What kind of ties can exist between two people? Borgatti [17] distinguishes between two main types of ties: states and events. State-type ties can be dimensionalized in terms of strength, intensity, and duration. In contrast, an event-type tie has a discrete and transitory nature and can be counted over periods of time. A further distinction is into four main types of ties: Similarities, Social relations, Interactions and Flows. While similarities and social relations are state-types, interactions and flows are event-type ties [12].
Similarities are based on some shared characteristic of nodes, e.g. same location, membership in the same club or participation in same event, or same attributes, like gender or attitude. People connected by such a tie do not necessarily have to know each other or interact with each other. In companies, there are several possible

kinds of those ties, e.g. same background, same gender, similar position or salary; or member in the same team, department, location, building, office space.

Social relations are roles that are assigned to the relationship between two nodes. They are distinguished into roles (like kinship, e.g. mother of, or other roles, e.g. friend of, boss of), and affective (likes, hates) and cognitive (knows, sees as happy) ties. The former are a formal description of the relationship between two people, while the latter are a description of one persons attitude or mindset towards another person. While roles play a big role in companies, since they constitute the formal structure and hierarchy, also affective and cognitive ties are important. **Interactions** on the other hand describe a tie that arises because two people interact in a certain way with each other (talked to, helped, harmed, gave advice to). They are very important in companies, and also the main focus of this thesis. Those ties are more dynamic than social relations or similarities: people have to actively do something for the tie to hold over time. The last type, **Flows**, are inherently dynamic as well. Those kind of ties are edges in a larger network through which some sort of resource is transported (e.g. information, support, beliefs).

Robertson [9] offers a different classification, by distinguishing between instrumental and expressive ties. **Instrumental ties** involve the exchange of information, advice, and other resources. **Expressive ties** involve the exchange of intangible resources like social support, compassion, and friendship. Expressive ties produce social integration. The distinction here is about the motivation of individuals on why to have an interaction or tie. Interaction based on instrumental ties is goal-oriented, while expressive ties involve some sort of commitment.

An important aspect about social networks is that ties can be multiplex - multifaceted bundles of interactions, roles, affiliations, and exchanges [18] [19]. The different types of ties often overlap, promote or hinder each other. For example, for two people to become friends, they first have to meet somewhere. That might happen because they go to the same event (similarity). Then, they start to talk to each other (interaction), and eventually share some resources, like gossip (flow). For a friendship (role) tie to arise, a lot of those interactions and flows have to happen. The two people probably also have very strong affective and cognitive ties: they like each other and have certain opinions and beliefs about each other. Similarly, a person working in department A that has a friend in department B (expressive tie), might get work-related help and contacts from that department faster (instrumental tie).

Multiplex ties are important for understanding the social networks at a company. An example is that for a supervisor-employee relationship to be actual and not just on paper, certain flows and interactions have to happen on a regular basis. A team is not just a team because they sit in the same office (share a similarity), but because they actively interact with each other. Furthermore, out of regular interaction, positive or negative cognitive and affective ties arise, and also certain roles, e.g. friendships. Also, for a person to get certain resources from another person, they probably have to know each other well, or at least be aware of the others' capabilities.

Flows of resources, knowledge, and help often happen despite the formal organisation (not between formal roles/ties, but informal ones, across departments and functions). They are essential for the functioning of the company [4]. Those ties, their interrelatedness and structure can have both positive and negative effects. They can make work flows more efficient or hinder them. They can create a welcoming social structure at the work place, but can also lead to biases and favoritism.

2.2.2.2 Types of Networks in a Company

Out of all the different ties that exist in a company, various networks arise. Some of them are measurable and visible, some are not. They do however have influence on the functioning and cohesiveness of the company, both on the level of an individual, the team and the system as a whole, as described in later sections.
The most obvious network is the official company structure, the organizational chart. It usually has a tree-like structure and represents the hierarchy in the company. In that sense, it is a network of roles (”reports-to”, ”boss-of” ties). However, another important concept is the informal organization with its various networks: Whom people seek advice or social support from, or whom they actually collaborate with.

> ”If the formal organisation is the skeleton of a company, the informal is the central nervous system driving the collective thought processes, actions, and reactions of its business units” [4]

Those networks are not just a ”dimension of social life complementary to, but separate from, task networks”, but are ”intextricably intertwined”[20].

2.2.2.3 Emergence of ties

After describing the types of ties and networks, this section scratches the question of why ties arise. Why do people connect with each other inside a company? Of course, there are many different reasons for connecting and situations where ties arise, also depending on the type of tie. However, three interesting aspects are to consider from a network perspective. The first is that many people get their jobs through contacts, i.e. through pre-existing ties inside the company [21]. The social systems inside companies are embedded into bigger social systems and networks.
Another important concept is homophily, or assortative mixing. We tend to associate preferentially with people who are similar to ourselves in some way [10]. What is important to distinguish are two different processes that lead to homophily. One is based on choice, and the other on opportunities [5]. Choice means that people consciously or unconsciously prefer to bond with similar individuals. Choice exists independent of the network structure, while opportunities arise because the structure of the network (or of the social system in general): The way the network is formed makes it more likely for similar people to meet. For example, it is more likely to meet and connect with somebody when already having a connection in common. This is the principle of transitivity, or clustering, described in section 2.2.3.2. Homophily can lead to maintaining inequality for minorities within organizations. If one group has more power than an other, homophily will lead to the networks of the less powerful group to consist of mostly less powerful people, limiting their social capital [22].
Another mechanism that depends on the structure of the network is preferential attachment, as described in section 2.2.3.1. Here, individuals are more likely to

connect with others that already have a lot of connections.

2.2.2.4 Strength of ties

Networks do not just express if there is a connection or not, but the connection can also have weights:

> "*Tie strength* captures a gestalt of the time spent, intensity (or psychological closeness), and degree of reciprocity between two social actors. Tie strength is also associated with each actor's intrinsic motivation and overall engagement in a relationship. By definition, strong ties require considerably more resource investment to maintain than weak ties. Together, content and strength capture the quantity, quality, and type of resources exchanged between two individuals." [9]

Tie strength can be expressed through a weighted adjacency matrix (see 2.2.1.3). For communication networks, the strength of a tie can for example be captured by the frequency or amount of communication between two nodes.

2.2.3 Network topology

While the previous sections focused on the description and the types of ties, this section is about characterising the structure of networks as a whole, their topology. Borgatti et. al. [17] make a distinction between "theory of networks", and "network theory". While network theory focuses on the "consequences of network variables", theory of networks refers to "the processes that determine why networks have the structures they do". Network theory in sociological literature explains why, given a certain network structure, people behave in certain ways. Theory of networks, on the other hand, is more focused on investigating, comparing and explaining the network structure as a whole. Theory of networks is not interested in one particular network, but finding features of real-world networks that manifest themselves in networks of different types.

> "One surprising discovery is the universality of the network topology: Many real networks, from the cell to the Internet, independent of their age, function, and scope, converge to similar architectures." [23]

> "Natural, social, and technological evolution seem to have produced organisms, communities, and artifacts with similar structure."[3]

The goal of theory of networks is therefore to find properties of network topology that transcend the specific network, and even the field of study, and find mechanisms that explain how those universal properties come into being. In particular, research investigates how real-world networks are not like random graphs. [10]

The random graph model defines a graph with n nodes and a probability p of an edge between two nodes: $G_{n,p}$. Another definition is a graph with n nodes and m edges that are distributed randomly between the nodes: $G_{n,m}$ [24].

The following paragraphs describe three of the fundamental properties of network topology: the scale free model, clustering, and small world models.

2.2.3.1 Scale free model

Scale free is a property of the degree distribution of networks. In those networks, the probability for a node to have exactly k links follows a power law distribution [25]:

$$P(k) \sim k^{-\gamma} \tag{2.3}$$

Those networks have a few nodes with very high degree (called hubs), while most of the nodes have low degree. The degrees of a random graph are Poisson-distributed [26]. The scale-free distribution, on the other hand, has a fat tail: it is very unlikely for nodes to have such a high degree in the random graph, as it is in the scale-free graph.

Networks across many different fields show a scale free structure: Hollywood actors acting in the same movie [25], the world wide web [27], citations between scientific papers [28], metabolic networks [29], twitter [30], and the email network of a university [31].

Scale-free networks show a high degree of robustness. Their functioning in terms of overall connectivity is very resilient to the random removal of nodes, since most nodes have very low degree and the removal of a random node is very unlikely to significantly affect the function of the network as a whole. However, strategic attacks on hubs have devastating effect [32]. Well-studied examples are the robustness of scale-free infrastructure networks [33] and the world wide web [11].

The discovery of universal scale-free properties in many networks led to the study of mechanisms and models explaining why the networks look like they do. One mechanism is preferential attachment, or the "rich-get-richer" phenomenon. It is quite simple: Nodes that already have a lot of connections are more likely to get even more. Barabási and Albert [34] combine preferential attachment with the growing nature of networks to define a simple model that has a scale free network as an outcome. The network grows at every time step, with a new node arriving and attaching to m nodes that already are in the network. However, the nodes to link with are not chosen randomly, but based on the already existing links. The probability of connecting to a node s is proportional to the number of it's already existing links k_s. An interesting aspect of this model is that it shows that structure and the evolution of networks are inseparable. How the network looks like depends on its past structure, and how it will evolve depends on the structure it has right now.

> "To explain a system's topology we first need to describe how it came into being." [23]

Noteworthy is also that the emerging networks of this model show a scale-free distribution at every point in time.

The preferential attachment mechanism is quite intuitive for social networks: a person who already knows a lot of people is more likely to meet new people. Also, they are easier to find, for example in advice networks, since they have more visibility [5].

More recent literature [35] makes use of the waste amount of network data now available and claims that strongly scale-free structure in networks is empirically rare, challenging the universality of network topology and advocating structural diversity of networks.

2.2.3.2 Clustering

Many real networks show transitivity: If node A is connected to node B and to node C, then there is a heightened probability that B is also connected to C. A, B, and C then form a triple. In social networks, this means for example that the friend of a

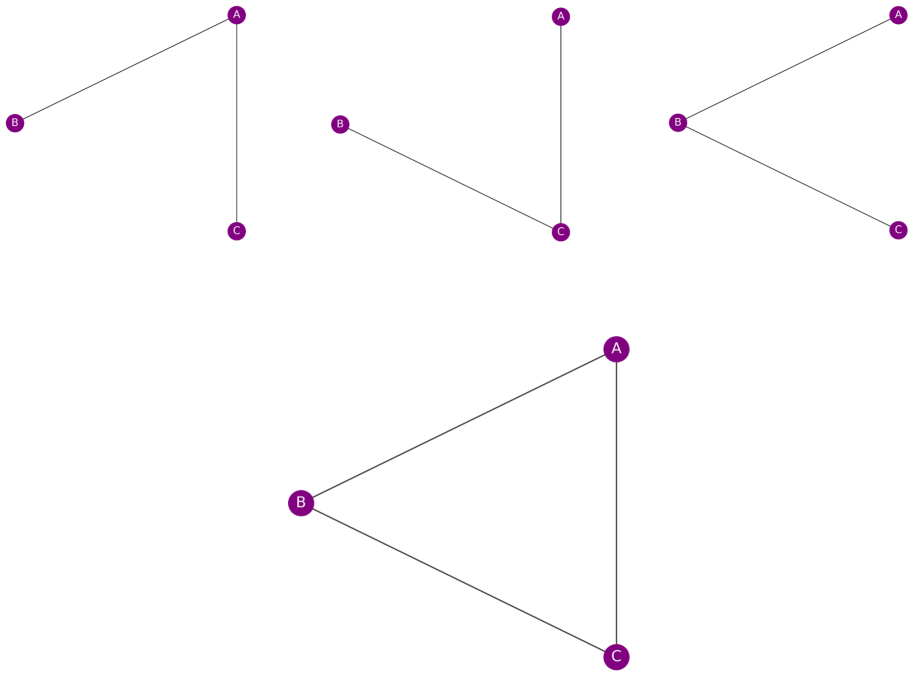


Figure 2.5: Three connected triples with the respective triangle

friend is more likely to be a friend as well [10]. In the random graph, the probability for the existence of an edge between B and C is independent of the edges between B and A and C and A. Therefore, the clustering property is a clear distinction from the random graph.
There are two measurements for clustering. One calculates transitivity for the whole network: It measures the number of triangles in the network over the number of connected triples, as shown in 2.5. In other words, it counts how many three-node-lines are closed, or for any three nodes A, B, C: if A is connected to B and C, are B and C connected as well. [10]

$$C = \frac{3 \times \text{ number of triangles in the network}}{\text{number of connected triples of nodes}} \tag{2.4}$$

Another measurement is the local clustering coefficient C_i. It calculates how connected the neighbours of the node i are. k_i is the degree of node i, and the neighbours have e_i edges between them, then the local clustering coefficient can be calculated as [36]:

$$C_i = \frac{e_i}{k_i\,(k_i - 1)\,/2} \tag{2.5}$$

If all the nodes that have an edge with i are connected with each other as well, C_i is 1. If there are no connections between the neighbouring nodes, i.e. nobody of i's friends knows each other, it is 0. C_i measures in other words how many of the possible connections between the neighbours of i are actually there, see figure 2.6. It is measurement for the density of the ego network, and is therefore also called"network density" in sociological literature [37]. A closely connected measurement are the

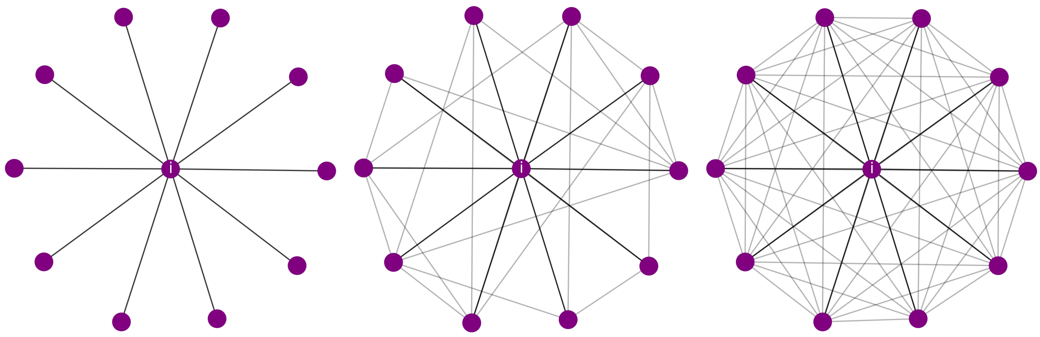


Figure 2.6: A network with various local clustering coefficients: (a) $C_i = 0$ (b) $C_i = 0.3$ (c) $C_i = 1$

structural holes, as described in section 2.2.5.5.
The clustering of the whole network is the average:

$$\overline{C} = \frac{1}{n} \sum_i C_i \tag{2.6}$$

Compared with transitivity (eq. 2.4), this measurement tends to weight the contributions of low-degree nodes more heavily [10].
Many social networks show clustering, which is quite intuitive. People often meet because they are introduced to each other by a third person; in professional as well as private setting. It is more likely to become friends with a person who is the friend of a friend, than with a stranger.

2.2.3.3 Small World

Small world networks combine the random graph's low average path length with clustering. Many real networks have a low average shortest path length (also called characteristic path length), see 2.2.1.6: one does not need many steps to reach everybody in the network. A well-known first experimental study that gave those networks their name is the "Small world problem" study by Stanley Milgram in 1967 [38]. By letting randomly selected people try to reach some targets by sending letters to others over several steps, he showed that the characteristic path length for the social network of the United States is about 6: The famous 6 degrees of separation, or 6 handshakes, that separate everybody from everybody else. With an abundance of social network data now available, newer studies use global information and quantify the characteristic path length even more: Studies on the facebook graph show that the average path length of the 721 million people using the social network in 2011 was 4.74 [39], and for the twitter graph and its 175 million active users it was 4.17 in 2012 [40].

When thinking about average shortest paths, it is important to keep in mind that the random graph actually does have a relatively low average shortest path length. Since the nodes are randomly connected with each other, it is not unlikely that the connections of one node span over a large part of the network. What needs to be explained, then, is not a low average shortest path length, but a deviation from it. The reason for the deviation in many real networks is clustering. Since nodes are clustered and more likely to connect with nodes that are already at a short distance from them, by keeping the amount of nodes and edges the same, the overall connectivity decreases, because the outreach over the whole network is reduced.

The small world model by Watts and Strogatz [36] combines those two properties (clustering and low average shortest path length): The model starts with a lattice, where n nodes are positioned on a circle and connected with their m nearest neighbours. This network has high clustering and a high average shortest path length, because one has to hop around half the circle to reach every node. Then, shortcuts between two random nodes are introduced. Every shortcut reduces the average shortest path length significantly, while the average clustering almost stays the same. The study then analyses how the clustering coefficient and the average shortest path length behave when increasing the amount shortcuts, starting from the regular lattice, and ending at the random graph. Even though the model is not very close to the topology of real networks, among others because of its degree distribution, it offers important thoughts and findings about networks: First, it does not need many shortcuts to make a small world network out of a highly clustered network, and second, global information is needed to estimate the average shortest path length. Further, complex networks are in realm between being completely regular (the lattice) and completely random.

Small world structures allows a specific kind of information processing, for example in the brain: while local, segregated areas process information, the shortcuts and low average shortest path length between those areas allow for effective communication between them [3]. As described in section 2.2.4.3, the small world structure has advantages for knowledge and creativity production, since it allows for individuals to express, modify, develop ideas and know how in the relatively safe space of clusters, which allow feedback from people who are familiar with the topic, while at the same time, ideas and new approaches can spread quickly between clusters through the shortcuts [41].

2.2.4 Processes on Networks

Given a certain network topology, what are the processes that happen on it? The structure alone does not describe the function of the system; it does not describe how the system will behave [11]. What is the nature of dynamical processes taking place on networks? The purpose, or the function, of an individual ties does not necessarily match the functions or patterns on the arising overall network. For example, two people connect with each other because they are friends, because they like each other, because they need something from each other. At the same time, without having that in mind, they influence each others opinion on a topic, or they serve as a pipe of some information flow, which in turn has influence on the spreading over the whole network. This section presents models of processes and functions of the network as a whole.

2.2.4.1 Contagion and Diffusion

Flow models assume implicitly or explicitly that something is transported, or diffused, over the edges of the network [17]. In social networks, there are several things that can flow between nodes, for example used goods, money, gossip, e-mail, attitudes, infection, and packages [42]. What this section will focus on is information, beliefs and attitudes, since those are most relevant to the communication network. Of interest is not information that is centrally distributed from above, but rather the processes on the micro scale that lead to spreading patterns on the macro scale. The nodes either receive a piece of information from the neighbours, or are influenced in their own opinion or belief towards something by the beliefs of the neighbours. Models for contagion or diffusion assume that the nodes can have certain states and adapt them based on updated rules that takes the state of neighbors into account. The nodes are connected on a network, and the structure of this network has influence on the outcome of the process.
Usually, the network structure is considered static, to be able to examine the dynamical processes taking place on networks independent on the evolution of the network structure itself. However, in many cases, those two are interrelated, and the dynamic on the network influences the network structure. One example is a child that does not go to school because it is sick. Because of the infection, the usual network structure of interactions changes, and this influences the disease spreading dynamics [43]. Often, the states of the nodes in models are binary, and the update rule might also depend on some random component, making the model stochastic.
Two main types update rules need to be distinguished. For the first, only one neighbour needs to have state 1 for the node in state 0 to change to 1. An example is gossip, or information spreading. The node does only have to hear it from one other person to know about it, and be able to spread it to others. The models are contagious-like: information spreads in similar ways as diseases would. For the second class of update rules, the node takes the states of all of it's neighbors into consideration before changing the own state. Examples are beliefs, opinions or attitudes. The important point is here is social influence: individuals do, consciously or unconsciously, base their own beliefs or attitudes on the beliefs of others. Or, in other words, the flow process is "an influence process in which, through interaction, individuals effect changes in each other's beliefs or attitudes" [42]. "We are all, in a sense, mining the collective experience in order to make judgements and choices." [44]
What kind of pattern on the macro level do the interactions on the micro level lead to? Since those processes depend on interacting agents and are therefore complex, it is often hard to predict the exact outcome. However, there are several classes of outcomes that can be distinguished [6]: **Steady states**, where the states are distributed in a way that they do not change anymore with the given update rule. This happens on the one hand if the whole network has adapted one belief, or has gotten the information. On the other hand, there could also arise patterns where

several states coexist on the network. The state of the whole network can however also **oscillate** between different steady states, or be **random** or in **chaos**, and never settle to a steady state. Computer models can help to identify the conditions in which a certain belief or information spreads over network. One interesting aspect are cascading behaviours, that sometimes are analysed like phase transitions in physics [43]: Beliefs, or other properties on the macro level, do not change linearly and gradually, but suddenly ”tip” from one to another, in a non-aggregate fashion. On top of the initial frequency of beliefs and intrinsic preferences, what is important to the outcome of the processes is also the network structure [45]. For example, for spreading processes that are contagious-like, Jackson [5] claims that what is really important is if the network is connected or not. The information will eventually spread over the whole connected component, but there is no spreading between unconnected parts of the network. Another property of network models is that highly influential individuals can be identified. Individuals with high eigenvector centrality (see section 2.2.5.2), for example, have high influence, since their opinion, their beliefs or information might spread to a large section of the network [42].
Such considerations are relevant for companies, especially in the context of organisational changes. Beliefs, attitudes, adaption to new tools often depend on the beliefs, attitude and adaptions of coworkers and other peers in the company. Even when change is imposed from above, for example by introducing new forms of organisation or new technical tools, the actual implementation, adaption and attitude that employees have towards it, depends on micro decisions and behaviour happening on the network, leading to macro patterns and processes.

2.2.4.2 Norms and Culture

Not just information spreads on networks: norms, modes of behaviour, and culture does as well. The idea here is that organisational culture is not just a set of values and rules of behaviour imposed from above, but emerges through interactions on the micro level, and can spread over the network.

> ”Organizational culture is not a cohesive, organization-wide control system implemented by top management, but an emergent property of informal relationships within work groups. Researchers within this tradition have investigated how norms, beliefs, attributes, behaviors, and other aspects of organizational culture are controlled through the informal networks of coworkers” [46]

Norms, and modes of behaviour, arise because groups tightly interact with each other. They arise from micro-level processes and over time and are not controlled by individuals. Further, they can spread over the network. However, norms do not just have positive effects on individuals and groups. An example for a negative norm is mobbing [47].

2.2.4.3 Knowledge and Innovation

Knowledge and know-how is produced, stored and accessed as a collective process. Powell [48] claims that networks, as opposed to hierarchies or markets, can process information in multiple directions, and have an ability to disseminate and interpret new information. On top of that, network structure is ideal for utilizing and enhancing intangible assets, like tacit knowledge or innovation:

> ”Knowledge-intensive activities [...] are based on know-how and detailed knowledge of the abilities of others who possess similar or complementary skills. Know-how typically involves a kind of tacit knowledge that is difficult to codify. These assets exist in the minds of talented people whose expertise cannot be easily purchased.” [48]

Knowledge and know-how is therefore not something that can be stored on paper, or accessed by one person alone, but is ”distributed in different minds, and to make use of it effectively, individuals need to know who knows what. In addition, [...] individuals need to have certain kinds of relationships (e.g., mutual accessibility, low partner-specific transaction costs) in order to utilize each others' knowledge” [22]. Knowledge creation and usage is a social process and depends on the structure of social relations - i.e., the structure of the network.

A group that is good at producing, accessing and utilizing knowledge is not necessarily the group with the smartest individuals:

> ”Surely the best groups will be made up of the epistemically best individuals, and the epistemically best individuals will necessarily make the best groups. This is not always the case, however” [49]

Again, the whole is different than the sum of its parts. What kind of group level functionalities and properties a group of people has does not just depend on the aggregate properties of the individuals. For example, diversity can lead to superadditive effects: diverse groups of people can outperform groups of best individuals [8]. Among others, because they have multiple ways to look at a problem, which can be beneficial.

Not only knowledge is produced, stored and transmitted in network, but the network structure and the processes happening on it have an influence on innovation, creativity and the emergence of new ideas.

> ”We know that creativity is spurred when diverse ideas are united or when creative material in one domain inspires or forces fresh thinking in another. These structural preconditions suggest that creativity is not only, as myth tells, the brash work of loners, but also the consequence of a social system of actors that amplify or stifle one another's creativity.” [41]

Creativity is a collective process: For one, what ideas an individual comes up with depends on the information, point of views, problems and solutions she is confronted with. That does in turn depend on who she interacts with, and what position in the network she has. And second, problem solving is, as the guiding the ship to the port example in section 2.1.3, often a collective endeavor. It needs the minds, ideas, backgrounds and skills of several interacting individuals. New ideas arise when different people meet. Associated is the theory of structural holes, as described in section 2.2.5.5. Uzzi [41] claims that networks showing a small world structure are ideal for innovation and new ideas. The dense clusters or cliques offer a safe space with similar people who understand each other and can work well together on known grounds. The shortcuts and connections between clusters, on the other hand, allow for ideas and novel material to spread and transform into new applications in other clusters. In a similar fashion, studies show that the co-authorship networks of scientific papers show small world properties, and that those structures are related to higher productivity [50]; and that knowledge diffusion on R&D collaboration networks is most efficient on a small-world network structure [51].

2.2.5 Network positions

After examining the structure of networks as a whole, this section goes back to "network theory" [17] and examines how, given a certain network structure, the position of the individual has impact on their performance, well being, creativity and power.

> "Position in a network both empowers and constrains action" [52]

> "A fundamental axiom of social network analysis is the concept that structure matters. For example, teams with the same composition of member skills can perform very differently depending on the patterns of relationships among the members. Similarly, at the level of the individual node, a node's outcomes and future characteristics depend in part on its position in the network structure" [12]

What does position in a network mean though? How can it be captured and measured? One way of expressing the position is the centrality in the network. This section describes four measurements for centrality: degree centrality, eigenvector centrality, betweenness centrality and closeness centrality. Another measurement for the quality of network positions, namely network constraint or structural holes, is described.

2.2.5.1 Degree Centrality

The most obvious centrality measure in networks is degree centrality. A node's degree centrality is the number of other nodes it is connected with. In social context, this is the number of friends, contacts or connections a person has, and therefore a measurement for direct influence [11].

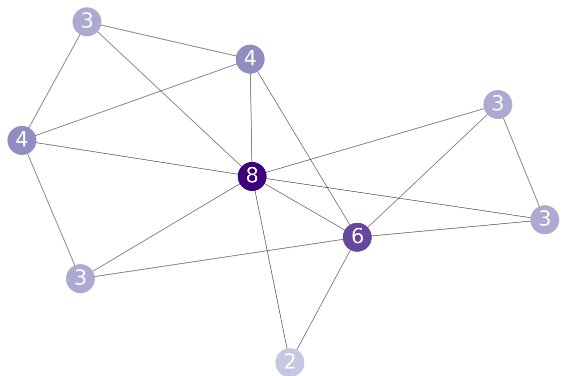


Figure 2.7: A undirected network with nodes and degrees

2.2.5.2 Eigenvector centrality

Degree centrality is an important characteristic for nodes, but the networked structure allows for less obvious (i.e., less local) measurements as well. Eigenvector centrality is based on the idea that a node's importance is determined by how important its neighbors are. It does not only account for how many nodes a node is connected to, but how close it is to other important nodes: The centrality of a node is proportional to the sum of the centrality of its neighbors [5]. It is therefore an indication for how good the node is positioned as a spreader, since it indicates how many nodes the node can reach after not only one, but several steps. This could be spreading of new information, beliefs, or early adoption of tools and methods. Eigenvector centrality is calculated, as the name says, with the eigenvector of the adjacency matrix A:

$$Ax = \lambda x \tag{2.7}$$

where λ is the largest eigenvalue of A. [11]

2.2.5.3 Betweenness centrality

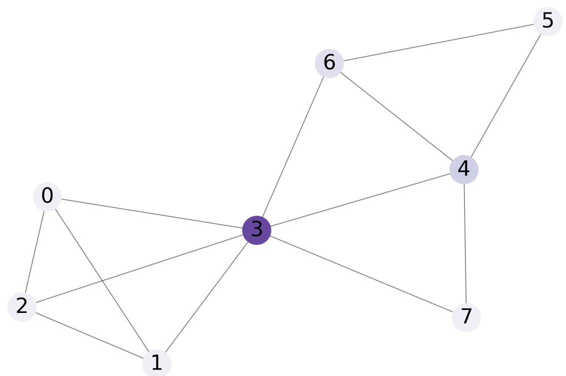


Figure 2.8: A network with node betweenness centrality measures. Node 3 connects two components of the network and does therefore have high betweenness centrality.

Betweenness centrality measures the extend to which a node lies on paths between other nodes [11]. Betweenness centrality of a node v is the sum of the fraction of all-pairs shortest paths that pass through v.

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t \mid v)}{\sigma(s,t)} \tag{2.8}$$

where V is the set of nodes, $\sigma(s,t)$ is the number of shortest (s,t) -paths, and $\sigma(s,t \mid v)$ is the number of those paths passing through some node v other than s,t. If $s = t$, $\sigma(s,t) = 1$, and if $v \in s,t$, $\sigma(s,t \mid v) = 0$ [53].
Nodes with high betweenness centrality often function as bridges: they connect parts of the network that otherwise would not be connected. When a node with high betweenness is removed from the network, the diameter of the network might significantly increase. In social networks, betweenness centrality can be an indication

of how important a person is for connections between different social groups, and for the network as a whole. For example, people connecting several departments that otherwise would not be connected at companies have high betweenness centrality.

2.2.5.4 Closeness centrality

Closeness centrality measures the mean distance from a node to other nodes [11]. Closeness centrality of a node u is the reciprocal of the average shortest path distance to u over all $n-1$ reachable nodes.

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)} \tag{2.9}$$

where $d(v,u)$ is the shortest-path distance between v and u, and n is the number of nodes that can reach u [54].
Closeness centrality is an indicator for how central a node in the overall network is.

2.2.5.5 Structural holes

Burt's [55] theory of structural holes concerns the structure of ego networks. A structural hole is the absence of a tie among a pair of nodes in the ego network. A person who's ego network spans over many structural holes is also called broker:

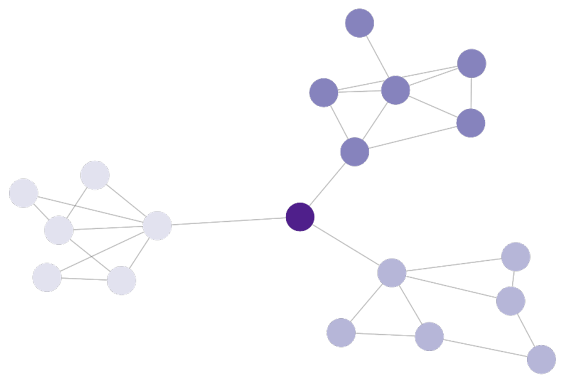


Figure 2.9: A broker spanning structural holes

she mediates between two or more dense clusters of interacting nodes, see figure 2.9. Such positions have advantages: people occupying them are more likely to come up with good ideas, novel solutions to problems, have higher compensation, positive performance evaluations, and promotions [56].
The structural holes measurement takes only local information into account (how the ego-network is structured), in contrast to the betweenness centrality measurement, which, even though capturing a related idea, uses global information and measures the importance of the node for the connectivity of the whole network.
One mechanism explaining the structural hole theory are flow processes through the network. Brokers sit at positions in the network that enable them to access novel, and different kinds of, information flowing through the network.
Borgatti et. al. [12] add another mechanism to the picture:

> ”The *binding mechanism* is similar to the old concept of covalent bonding in chemistry. The idea is that social ties can bind nodes together in such a way as to construct a new entity whose properties can be different from those of its constituent elements.”

A set of nodes that forms a cluster, communicate a lot with each other and act, in fact, as one. The broker then connects several such clusters, and can take several advantages of that position:

> ”People specialize within clusters and integrate via bridges across clusters. A theme in this work is that behavior, opinion, and information, broadly conceived, are more homogeneous within than between groups. People focus on activities inside their own group, which creates holes in the information flow between groups, or more simply, structural holes. Given greater homogeneity within than between groups, people whose networks bridge the structural holes between groups have earlier access to a broader diversity of information and have experience in translating information across groups.” [56]

This idea is also connected to the description of creating of identity (see section 2.1.4), and of diffusion of norms (section 2.2.4.2).
Since sitting at the intersection of several groups, brokers are exposed to various ideas, points of views, problems and solutions. This can have a synergy effect: combining elements from several groups leads to new behaviour or beliefs. Another point is creation of new ideas and knowledge by combining information from several groups.
Brokers do also have important social functions in networks:

> ”The simplest act of brokerage is to make people on both sides of a structural hole aware of interests and difficulties in the other group. People who can communicate these issues between groups are important because so much conflict and confusion in organizations results from misunderstandings of the constraints on colleagues in other groups.” [56]

Burt defines *network constraint* to measure brokerage. The constraint is a measure of the extent to which a node v is invested in those nodes that are themselves invested in the neighbors of v. Formally, the constraint on v, denoted $c(v)$, is defined by

$$c(v) = \sum_{w \in N(v) \setminus \{v\}} \ell(v, w) \tag{2.10}$$

where $N(v)$ is the subset of the neighbors of v that are either predecessors or successors of v and ℓ(v,w) is the local constraint on v with respect to w. Formally, the local constraint on u with respect to v, denoted $\ell(v)$, is defined by

$$\ell(u, v) = \left(p_{uv} + \sum_{w \in N(v)} p_{uw} p_{wv} \right)^2$$

where $N(v)$ is the set of neighbors of v and p_{uv} is the normalized mutual weight of the (directed or undirected) edges joining u and v, for each node u and v. The mutual weight of u and v is the sum of the weights of edges joining them (edge weights are assumed to be one if the graph is unweighted) [56]. To end with, what should be kept in mind about the structural hole theory is that action, ideas and performance of brokers do not just come from their own genius. They are implicitly and explicitly empowered and leveraged by the network position they occupy.

2.2.6 Embeddedness into the Network: Motivation, Engagement, Well-Being

As described in the previous section, the position of a person in a network can both constrain and empower action. Through their direct and indirect peers, individuals can access resources, knowledge, collaborations and inspiration to perform their work related tasks. However, this is not the only function of (social) networks at work. The relations individuals have to their peers, and how those relations are structured in the network as a whole, have an impact on their motivation, their engagement, their well being and the meaningfulness they connect with their work [20]. One aspect here is that coworkers seek out to others not just for their task knowledge, but also for affective rewards of social interaction, including friendship and emotional support. Norms and culture in the company often arise from informal networks [46], see section 2.2.4.2. In turn, how those networks are structured, and what position the individual is in, matters.
One line of research looks at network position and job satisfaction. Raile et. al. [57] examined how the position in the friendship network at the workplace correlates with job satisfaction, and found that closeness in the friendship network is positively correlated with job satisfaction, while degree and betweenness centrality do not have any impact. Another study [58] studied the communication network of a health care provider and found that individuals at the highly connected core had significantly greater job satisfaction than those who were on the periphery.
Another line of studies relates turnover to network positions. Krackhardt and Porter [59] found that turnover does not occur randomly, but in structurally equivalent clusters. Further, turnover of friends affects attitude of stayers: they actually get more committed to the organization. Gloor et. al. [60] did a network study of the email communication network of managers and found that on average, managers who quit had lower closeness centrality and less engaged conversations compared to managers who stayed at the company. Feeley [61] examined a communication network and found that both degree centrality and betweenness centrality had a negative effect on turnover. Another study [62] shows that employees at the periphery of the social network of a company are at high risk of turnover. Soltis [63] studied advice networks and turnover intention, both for advice seeking and advice giving. They found that especially the ability to seek advice from coworkers outside of regular required workflow ties decreases turnover intentions. This can be seen as an embeddedness into a bigger network, or ties that span departments and teams, is positive. On the other hand, they found that individuals with high in-degree in the advice network (those who are sought out heavily for advice) do not perceive themselves to be adequately rewarded for their extra efforts and are more likely to intend to leave.
Another studied variable is engagement. Engaged employees "have a sense of energetic and affective connection with their work activities and exhibit greater task

and contextual performance. Engaged employees often perceive a sense of autonomy, task variety, task significance, job complexity, and social support in the workplace. They are characterized by high levels of energy and identification with their work." [64]

Halgin et. al. [65] show that highly engaged employees not only profit from their network, but they actively try to connect with others and strengthen their relationships in order to achieve certain goals. They actively try to shape and build their networks.

Network measurements can even be used to predict engagement. Guha et. al. [66] studied a company's internal social networking service and the correlations with results from an annual employee engagement survey. They found that both degree and eigenvector centrality in the internal social network correlate with higher engagement, as well as clustering coefficient. They conclude that both the position in the social network is important for employees, since high engaged employees arrange themselves in specific ways to achieve enterprise objectives, as well as tightness of connections. They use their findings not only to analyse, but also predict employee engagement based on the network positions.

Another important aspect is that engagement as a behavioural state can spread like a contagion (see section 2.2.4.1). It is a social influence process: when somebody a person works with is very engaged and motivated, that might make the person more engaged as well. At the same time, if peers are disengaged, it can have negative effects on the ego's motivation. A longitudinal study on a large corporation found that engagement and disengagement spread from one employee to another with direct peers exerting the strongest influence [67].

A more in-depth study tries not just to find correlations, but explain them with mechanisms. Robertson et. al. [9] investigate meaningfulness at work. They claim that an important aspect to perceive one's work as meaningful are the various connections and networks at work. Not just the direct connections to peers play a role, but "how those relationships are interconnected in larger social structure relative to the self", i.e., what kind of overall network position the self is in.

> "Our central argument is that whether individuals have the opportunity to successfully engage in the various purposeful actions that can lead to meaningfulness of work will ultimately depend on the resources available to them via their social networks."

To sum up, those studies have to be compared with caution, since they having different research objectives, and look at different types of both organizations and network data. However, a common finding is that network structure matters, and the better integrated the individual into the structure, the better.

2.2.7 The Communication Network - More Than Just a Flow of Information

Why is communication, and the network that arises from the structure on which communication happens, important? First of all, for a system with macro-level functionalities and identity to emerge of a set of individuals, communication is needed. Communication is not just an add-on, but a crucial feature. The system only exists because individuals communicate with each other. The sum is more than the parts, because interaction leads to non-aggregate effects.

Internal communication has several functions that go beyond pure sharing of information, both on the micro and on the macro level: coordination, collaboration, planning; creating, storing and accessing knowledge and know-how; coming up with new ideas and developing them further; sharing purposes and goals and creating an identity. For employees, embeddedness into communication networks means access to information and resources, but also motivation, engagement, or even meaningfulness in their work. On top of that, the company is a social system, with its own norms, culture and modes of behaviour. While a tie in a communication network is inherently of the *interaction* or *flow* type (see section 2.2.2.1), they do capture other ties as well: for example, both instrumental and expressive ties require communication. Further, to constitute an affective or cognitive tie, communication is needed. However, what does the structure of the communication network tell us about their functions? What does the structure of a good communication network look like? While other networks, for example traffic networks or blood vessel networks, are optimized either by evolution or engineering in regards to a specific function - allow a flow to as many points as possible in the shortest possible time [11] - it is hard to find such a function for human communication networks. A naïve approach would be to optimize the transmission of bits of information. This is naïve because first of all, nowadays, transmission of information inside a company is cheap. Everybody can be easily reached by just sending them an email directly, there is no need to go over several steps. Important information does not need to diffuse through the network through several people, through time intense and expensive phone calls or conversations or letters. However, this is just about *transmission* of information - to be actually perceived by the recipient, it needs more. For example attention and time: one cannot read and react to infinite amount of emails. As already stated, on top of that, internal communication has more functions than just pure transmission of information. To actually communicate in an effective way, often times there has to be a certain *relation*, or a certain position in the network, between sender and recipient. To access resources, they have to know each other, or have certain affective tie with each other, or have to be introduced to each other by a shared contact. To share knowledge and know-how, the two have to speak the same language and share some background. And, generally, they have to know about each other, have to know what kind of knowledge or support or ideas they can get from the other person. However, for such ties to arise and to be maintained, a lot of communication is needed. In short: the communication is embedded into social networks, and vice versa. One cannot exist or be understood without the other. In some ways it is naïve to look at communication data and nothing more to understand social network inside a company. It does not say anything about the type of tie, or the relation between the two people. However, it is a relatively cheap and unbiased source of data. As opposed to surveys, where people are asked whom they trust or seek advice from or whom they communicate with, the data captures not just the perception of individuals, but the actual situation, and takes much less time and energy to collect. Further, it allows to examine the actual patterns of communication and collaboration. What has to be kept in mind is that any network data is always just a map of the real network: A map that can be close or far apart from reality, and an abstract view of real relations and social structure.

3

Methods

3.1 The data set

The company analysed is operating at 16 locations in 10 different countries, with 1.467 employees in total. The data set contains meta data of all the email conversations taking place internally in a range of 90 days, from February to end of April 2021. It includes sender, recipients and timestamp for every message exchanged. An additional data set is a list of all email-addresses, names and locations, and is used to assign the location to every node. Only personal accounts are taken into consideration. The first step is therefore to clear the data from technical and other accounts. From this data, a network is constructed, with employees as the nodes and connections between them if emails are sent. Figure 3.1 shows the locations with user per location, and the color legend that is used for every plot.

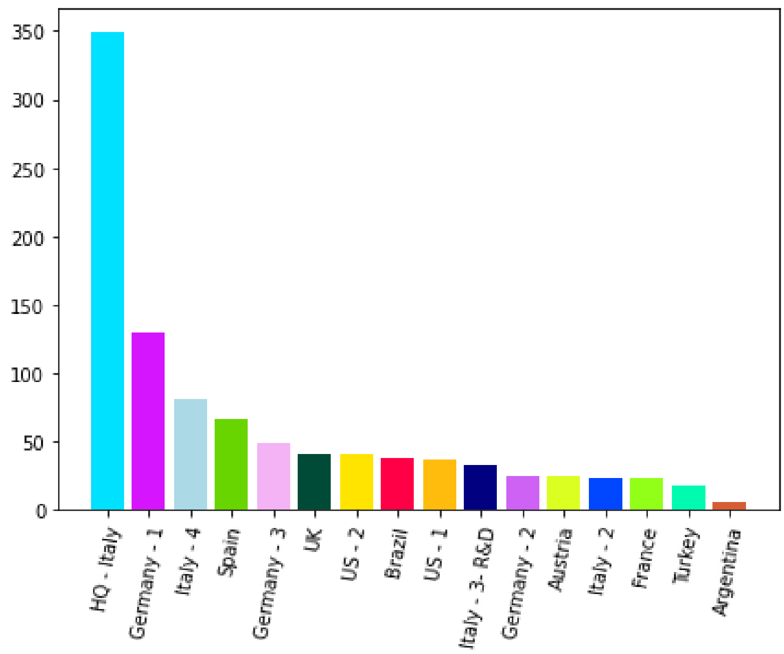


Figure 3.1: Locations with users per location

3.2 Network Analysis

Network analysis is done in Python, using networkX package [68].
Three different networks are constructed and examined. One is the **directed and weighted network G dir**, containing all the internal emails. Analysis is done not just on the network, but also on the total amount of messages sent and received.
However, it is difficult to handle and analyse directed and weighted networks. Furthermore, not all connections in this network are equally significant. Pure frequency of emails is not necessarily a good measurement for weight, since two people might exchange many short emails about a not so important thing, while two other people exchange only 2 long emails about something important. Therefore, two undirected and unweighted networks are constructed out of the raw data:
The **loose communication network, G1**, is a undirected network from the same data: There is an edge between two nodes if they exchanged at least 2 emails. This to make sure that there has actually been an interaction between the two people.
The **tight communication network, G2**, has edges when two nodes have exchanged emails in at least 9 of the 13 total weeks, making it a network of tighter collaboration. This is to distinguish it from just sparsely exchanged emails.
For all the networks, number of nodes, number of edges, density, average degree, diameter, average shortest path length, transitivity and clustering are measured. From the networks, subgraphs are constructed for every location, containing the nodes and the connections between them. The subgraphs are then analysed compared in terms of clustering, density, average degree and size.

3.3 Communication between Locations

To examine the communication between locations, first, the overall amount of emails sent inside and between locations is analysed and visualized. Next, networks and adjacency matrices are created that have the locations as nodes and the normalized amount of connections between the nodes of those locations as edges; for both G1 and G2. The networks are visualized and analysed.

3.4 Connectivity and Cohesion

To examine connectivity and cohesion, a k-core decomposition is done for both the loose network G1 and the tight network G2, as described in section 2.2.1.9. The results are compared with measurements for closeness centrality. The focus is especially on understanding the differences between locations: what locations make the core, and what the most central nodes are.

3.5 Node-level Measurements

On the node level, several measurements are done for every node: the total amount of messages that were sent and received by the user (tot-msg-sent, tot-msg-received), the in- and out-degree for the overall directed network, and several measurements for both the loose (G1) and the tight (G2) network: degree centrality 2.2.5.1, betweenness centrality 2.2.5.3, closeness centrality 2.2.5.4, eigenvector centrality 2.2.5.2. Further, the core number of the node is measured, the maximum clique number of the node, the constraint to determine structural holes (see 2.2.5.5), and the diversity of interaction, which is the number of departments the node interacts with. Nodes with low constraint are closer examined by plotting their ego networks, to identify structural holes.

4

Results

4.1 Network measurements

4.1.1 The overall directed network

This network contains the total amount of emails exchanged internally between the employees of the company. Figure 4.1 shows a plot of the network, with different colors for different locations. It contains 933 active users, i.e. people who sent or received at least one email to/from another person inside the company. 78 of them did not send any emails and 1 did not receive any. In total, 549,722 messages were exchanged. On average, every user sent and received 589 messages (around 9 per working day). Figure 4.2 and figure 4.3 show the distribution of messages. Most people sent and received less than 1000 messages, while a few sent a significantly higher amount. The most active users sent 8055 messages and received 7401. The network has 40,562 edges, which makes it on average 13.6 messages per edge.

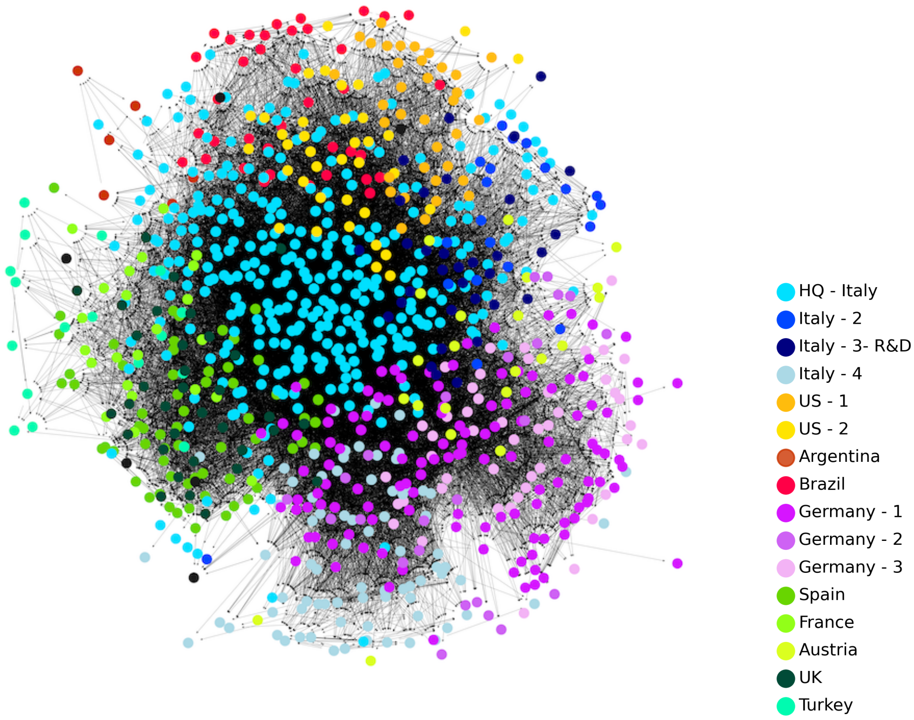


Figure 4.1: The overall directed network with color labels

	Messages sent	Messages received	out-degree	in-degree
mean	589.2	589.2	43.47	43.47
std	880.5	850.9	35.75	28.98
median	297	292	37	38
min	0	0	0	0
max	8055	7401	240	185

Looking at the degree distributions of the network, the average user sent and received messages from 43.5 different users. The out-degree distribution has a higher standard deviation than the in-degree. Since 78 users did not sent any messages at all, the in-degree-distribution is shifted to the left (Figure 4.5). The highest out-degree is 240, meaning that the user with highest out-degree centrality sent messages to 240 individuals, while the user with the highest in-degree centrality got messages from 185 different individuals.

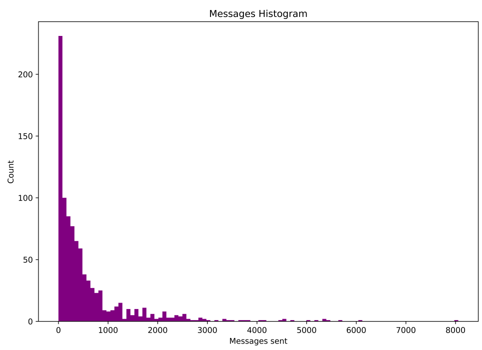


Figure 4.2: Histogram of sent messages per user

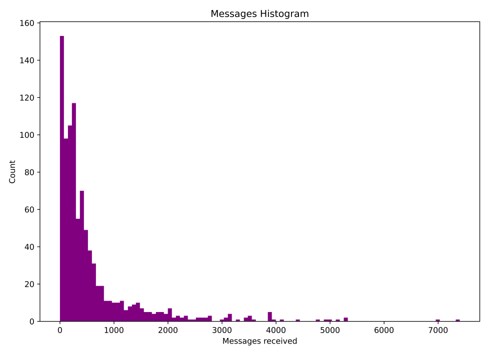


Figure 4.3: Histogram of received messages per user

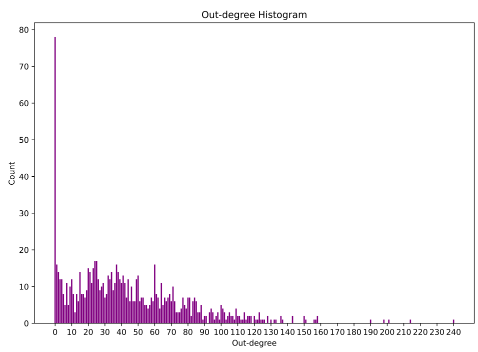


Figure 4.4: Out-degree distribution

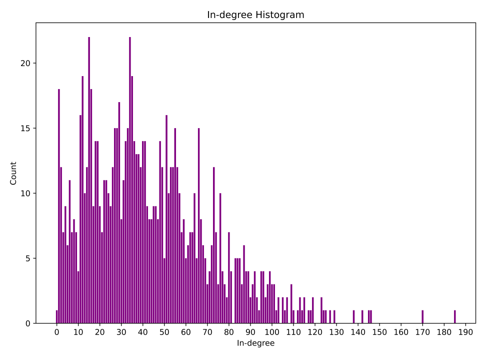


Figure 4.5: In-degree distribution

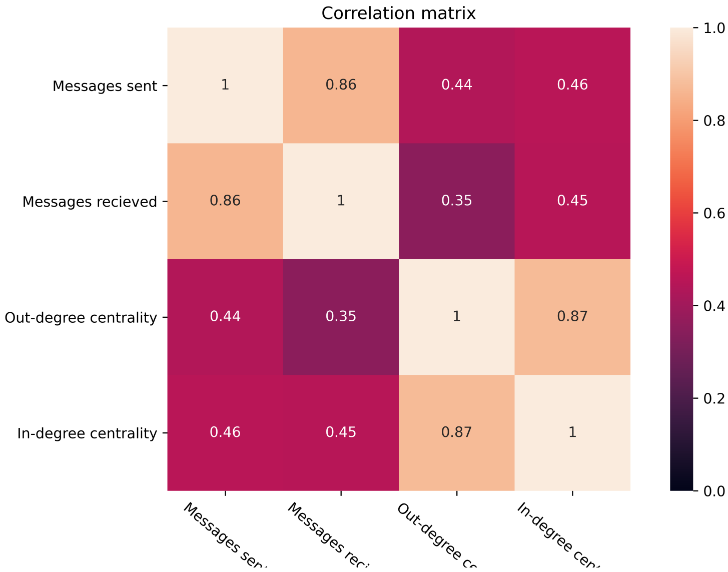


Figure 4.6: The correlation matrix of the overall directed network

Figure 4.6 shows how amount of messages sent, received, and out- and in-degree centrality correlate. Amount of messages sent and amount of messages received correlate highly: a user that sent many messages is also likely to get many. As expected, there is a positive correlation between sending and getting many messages, and having a higher out- and in-degree centrality. Interesting to notice is that messages sent correlates slightly higher with in-degree-centrality than out-degree-centrality. That means that sending more messages is a better indicator for getting messages from many different people, than sending messages to many different people.
The following table shows results of measurements on the network level:

n nodes	n edges	density	transitivity	avg clustering	avg shortest path length
933	40562	0.046	0.28	0.46	2.27

Since the graph is not strongly connected, it is not possible to calculate the diameter.

4.1.2 Comparison of subgraphs

Figure 4.7 shows how the subgraphs per location scale in terms of nodes and edges. The scale is almost linearly, the fit is plotted as the dashed line. The number of possible edges however increases with n^2. Locations above the line have more edges between nodes than average, and nodes below the line have less. All German locations lie below the average, while the headquarter is slightly above. Looking at figure 4.8, showing the density of network, this is seen more clearly: The density is especially high for the subgraphs in Spain, U.S. 2, and Argentina. It is very low for Germany 2 and the headquarter, and low for all the German quarters. However, the density of every location is higher than the density of the whole network. It is significantly more likely for a connection to exist between two nodes of the

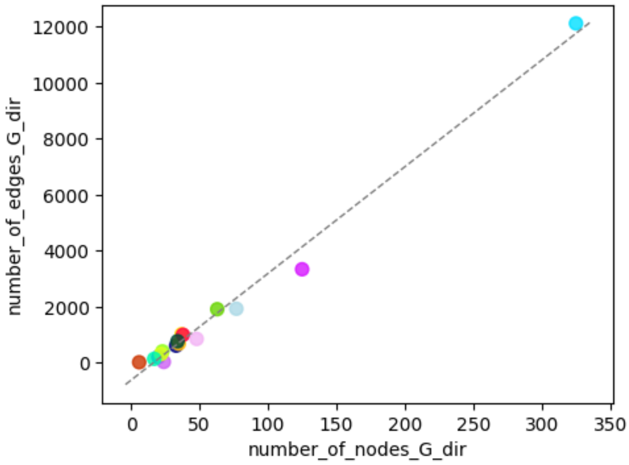


Figure 4.7: Number of nodes vs. number of edges for subgraphs

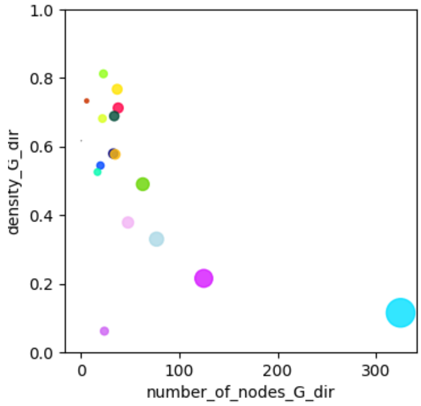


Figure 4.8: Number of nodes vs. density of subgraphs

same location than between different locations. This is also related to the findings described in section 4.2: intra-location communication is more intense than inter-location communication.

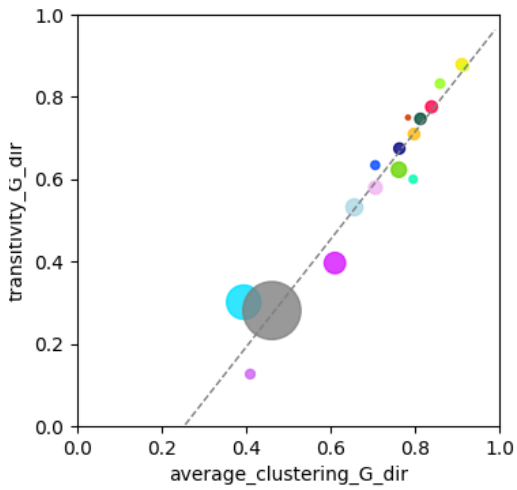


Figure 4.9: Average clustering vs. transitivity of subgraphs

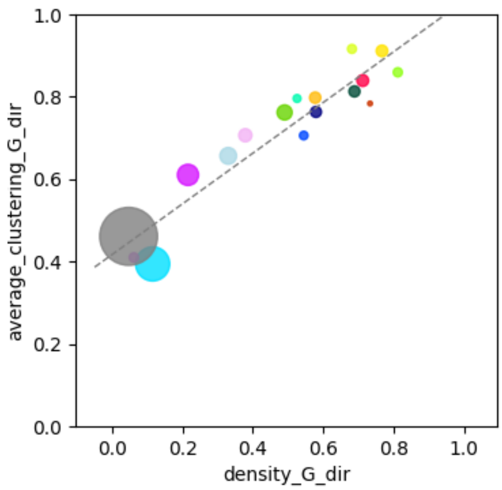


Figure 4.10: Density vs. average clustering of subgraphs

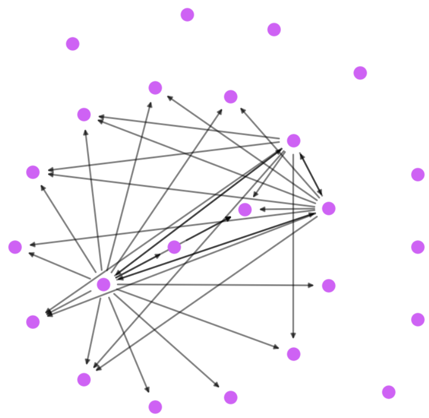


Figure 4.11: Subgraph Germany - 2 with low clustering

Figure 4.9 shows how the subgraphs scale in terms of clustering and transitivity. The grey node is the average for the total network. The sizes of the nodes are proportional to the size of the subgraphs. Figure 4.10 shows how density and clustering correlate. The more edges there are in the network relative to the number of nodes, the more likely it is that a triple is connected: the higher the density, the higher the clustering. This can also be seen in the plot, with some deviations though. Relative to their density, the U.S. offices and especially Spain show high clustering, while the Italian offices, especially the head quarter, show low clustering. Apart from that, there are several interesting aspects to note here. First, Germany 2 shows both very low clustering and transitivity. Looking at the plot of the subgraph (figure 4.11), this can be seen clearly: of the 24 nodes, there are actually only 3 connected triples. It is a small location, has low density and does not really act as one entity, since many nodes are not even connected with each other in the overall network.
Further, the headquarter has the lowest average clustering and second-lowest transitivity. This is an indication for the fact that at the locations that are not the headquarter, nodes are more likely to interact in connected triples than just 1-1 (see 2.2.3.2). At the headquarter, on the other hand, many interactions are not happening in triples. If person A is talking to both B and C, B and C are not necessarily talking to each other. Further, both U.S. 1 and Spain are very highly clustered. There are almost no non-connected triples. This is connected to findings in later sections (see 4.3, 4.4), where nodes from those locations have high core number and degree. In this figure, the difference between transitivity and average clustering can be seen. In locations where transitivity is relatively higher than clustering (HQ, Italy - 4, Argentina, Spain), nodes with higher degree are in many triples. In locations where clustering is higher (the German locations, Turkey), it is especially nodes with low degree (see 2.2.3.2).

4.1.3 The loose communication network

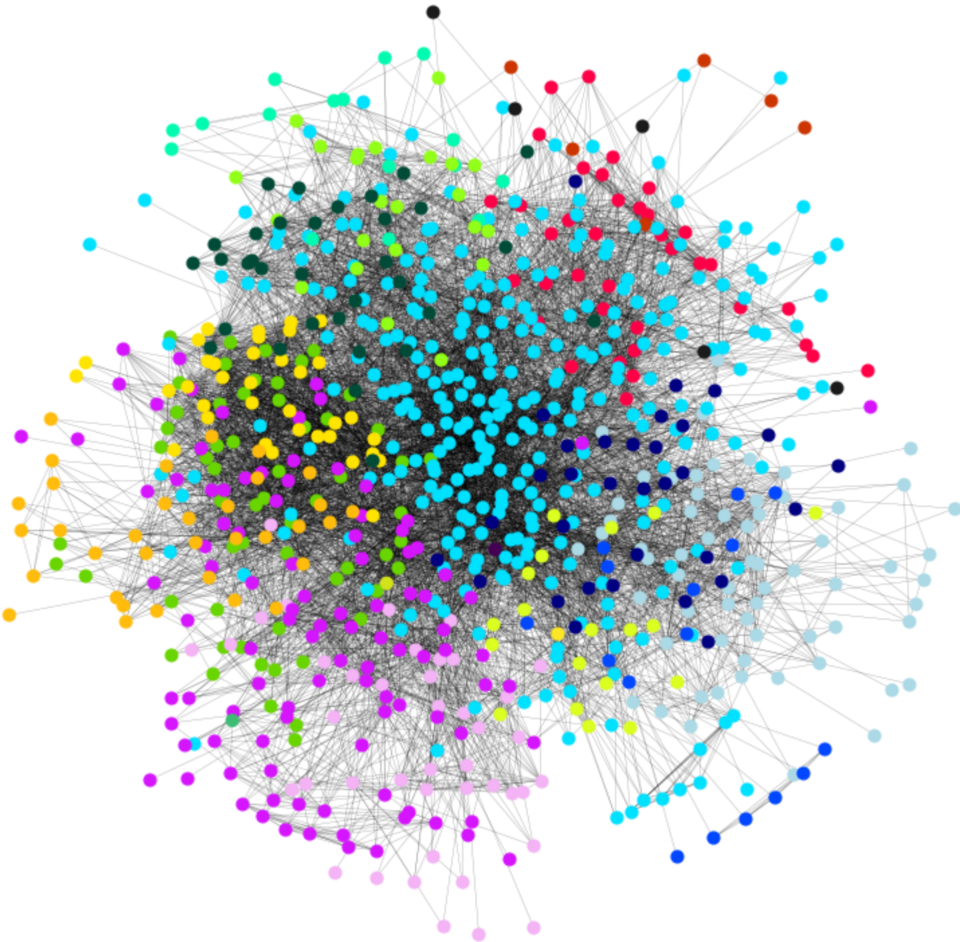


Figure 4.12: The loose communication network

This network is the network of all employees that exchanged at least 2 emails over the three months. It contains 828 nodes and 10,070 edges, that means that on top of the 78 users that did not have any out-connection and the 1 user that did not have in-connections, there are 26 users that fall away, since they did not exchange at least 2 emails with somebody else. The amount of edges is reduced by around one half (considering that the overall network is directed). That means that half of the users that send emails to each other actually get emails back and exchange more than 1 email. The density is around $\frac{2}{3}$ of the density of the overall network. The

n nodes	n edges	density	avg degree	transitivity	avg clustering	diameter	avg shortest path length
828	10070	0.029	12.16	0.345181	0.472129	6	2.903279

network is connected: it has only one component. The diameter is however 6, and the average shortest path length is 2.9. Figure 4.13 shows the degree distribution of the network. With the relatively high clustering and transitivity, and low average path length and diameter, compared with the size of the network, this network could count as a Small world network, as discussed in section 2.2.3.3.

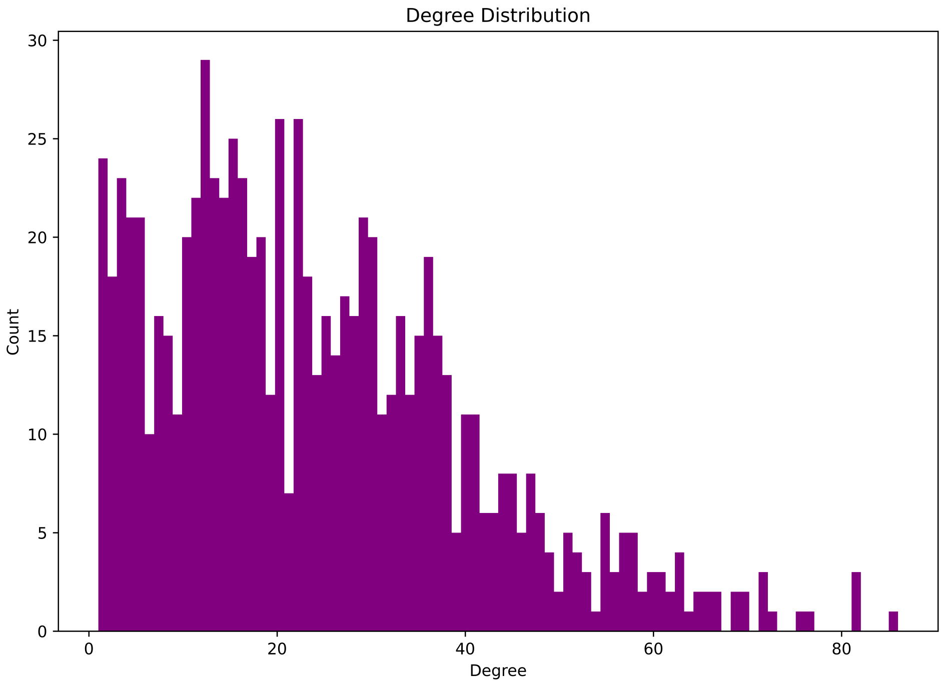


Figure 4.13: Degree distribution undirected network

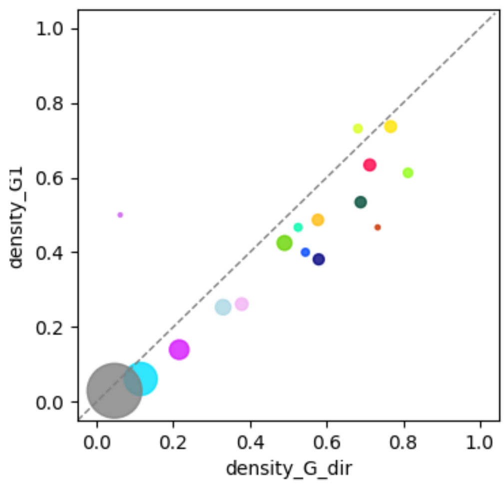


Figure 4.14: Density G dir and G1, subgraphs

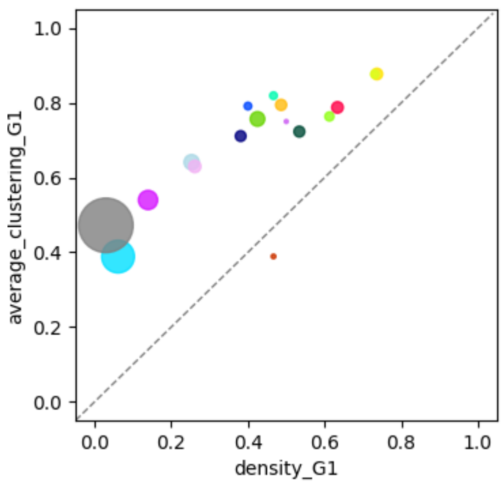


Figure 4.15: Density vs. average clustering of G1, subgraphs

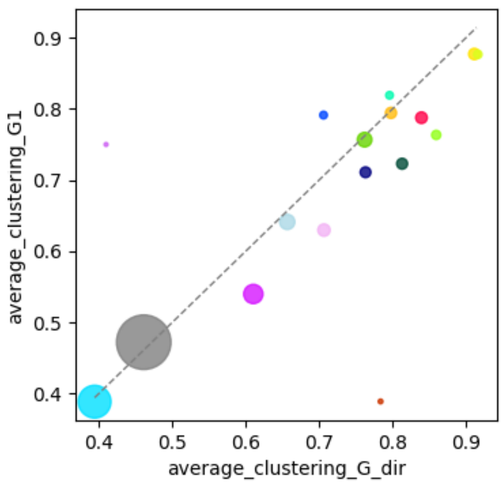


Figure 4.16: Average clustering Gdir vs G1, subgraphs

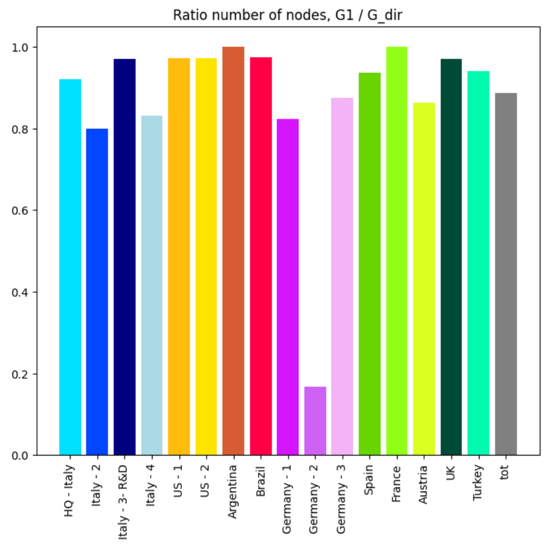


Figure 4.17: Ratio number of nodes, G1 / G dir

Further, the subgraphs per location of this network are analysed and compared with the subgraphs of G dir. Figure 4.14 shows how density of the subgraphs in G dir and G1 correlates. The size of the points corresponds to the size of the sugraphs in G1 (e.g., number of nodes at the location that are in G1). For almost all locations, the density has decreased. However, for some (namely Austria and Germany 2), it has increased. This is because at those locations, the loosely connected nodes are no longer part of the network. Figure 4.15 shows how density and clustering correlate for the subgraphs of G1. Looking at figure 4.16 showing how clustering in G dir correlates with clustering in G1, and compared with the same plot for G dir (see figure 4.10), the density / clustering behaves similar for both G dir and G1. Comparing also with figure 4.16, most locations have slightly less density and less clustering in G1 than in G dir. However, the total network is slightly more clustered in G1 than G dir. The very small subgraphs in G1 (Italy 2, Germany 2, Turkey) are an exception, since they are more clustered. Figure 4.17 shows the ratio of nodes from G dir that are still part of G1. Only very few nodes from Germany 2 are still part of the loose communication network subgraph. This can already be seen from the results in the previous section, since those nodes are not very well connected.

4.1.4 The tight communication network

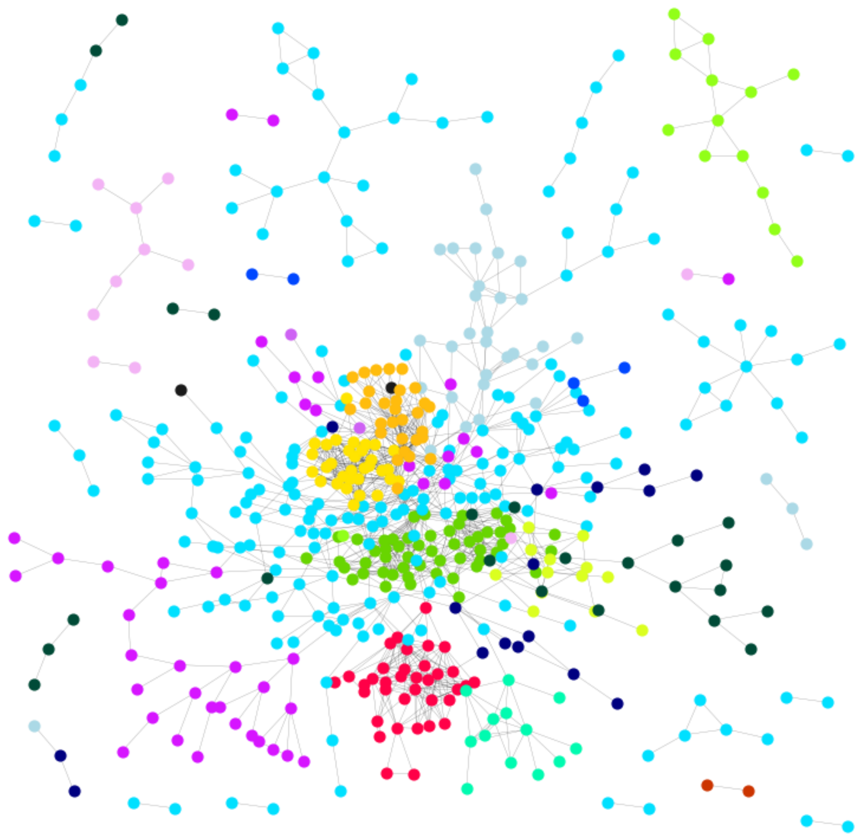


Figure 4.18: The tight communication network

The tight communication network consists of all the nodes that have exchanged emails in at least 9 of the 13 weeks. It is therefore a map of stronger collaboration than the loose network. In figure 4.18, it is already possible to see that most nodes are close to nodes from the same location. An exception is the head quarter (in turquoise), which is spread over the whole network, and whose nodes have connections with other nodes from different locations. Since the head quarter contains corporate positions and functions, this makes sense. This is also shown in figure 4.25, showing connections between and inside locations.

n of nodes	n of edges	density	avg degree	transitivity	avg clustering
524	1295	0.0094	4.94	0.39	0.30

524 of the 933 total users have at least one such intense connection; the ratio of active users is therefore 0.56. Figure 4.19 shows the degree distribution. The node with the highest degree has 24 connections, while most nodes only have one connection.

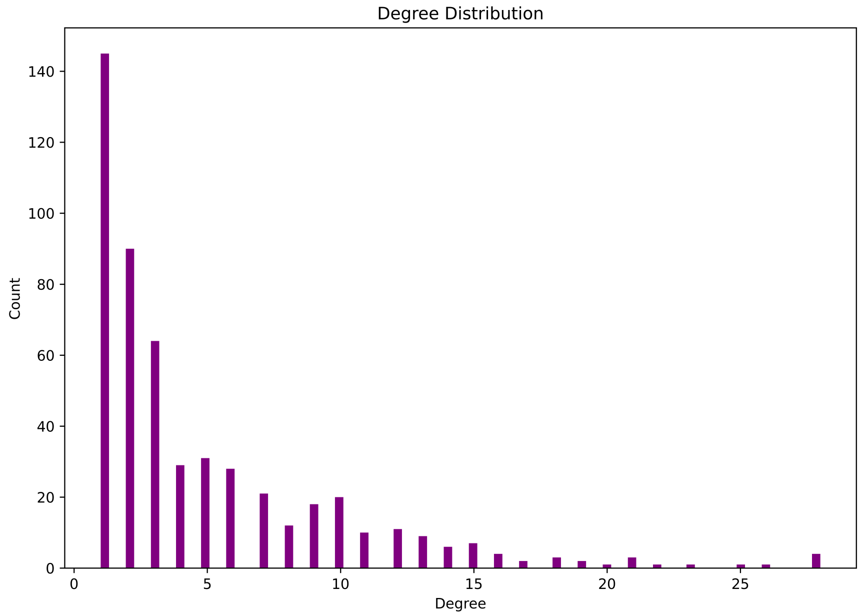


Figure 4.19: Degree distribution of tight network

The network is not connected: it has 25 components. The biggest component contains 421 nodes, and has the following characteristics:

n of nodes	n of edges	diameter	avg shortest path length
421	1206	17	5.86

It has an average shortest path length of 5.9, and a diameter of 17, which is significantly higher than the loose network with 2.8 and 6. It is to be expected though, since the average degree is lower, and tight connections are not spread out so much over the whole network: nodes work closely with only a few other nodes that are close to them. Even clustering and transitivity is slightly lower: close collaboration seems to happen in pairs rather than in groups of three. While the degree distribution resembles a scale-free distribution, with most nodes having only 1 or 2 other connections, but a few nodes having very high degree, with its high diameter and average shortest path length, and low clustering, it is not really a small world network.

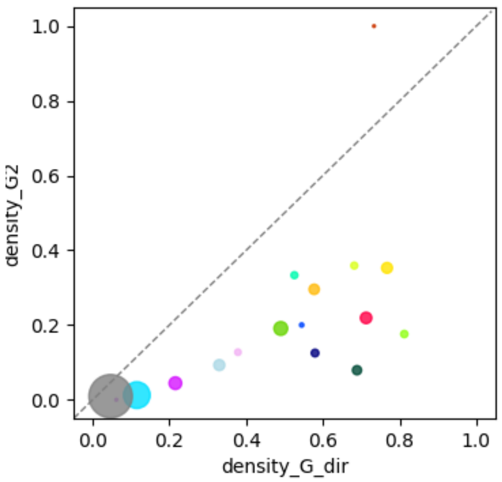


Figure 4.20: Density G dir and G2, subgraphs

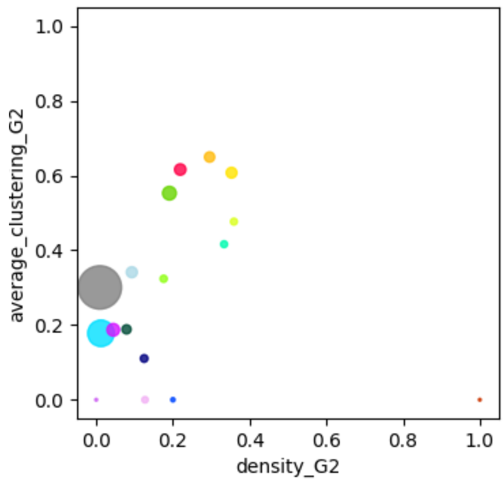


Figure 4.21: Density vs. average clustering of G2, subgraphs

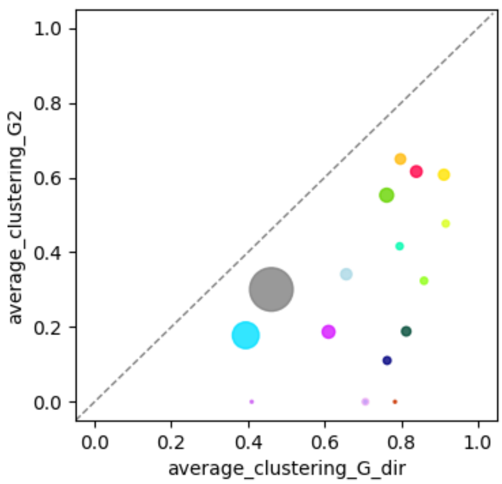


Figure 4.22: Average clustering Gdir vs G2, subgraphs

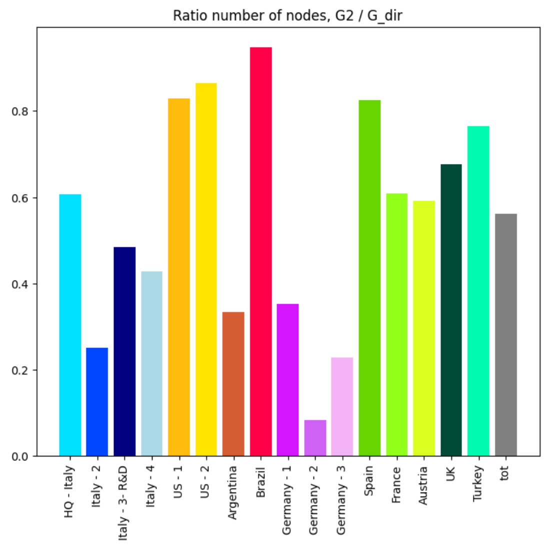


Figure 4.23: Ratio number of nodes, G2 / G dir

Figure 4.23 shows the ratio of nodes from the G dir subgraphs that are still active in G2. The highest ratio has Brazil, followed by the two US-offices and Spain. This is also related to later findings, since those are the locations at the core of both networks (see 4.3). Locations with a small ratio are all three German offices, Italy 2 and Argentina. This means that nodes at those locations do not exchange emails as intensely as at other location. It could however also depend on different collaboration styles or other modes of communication.

Figure 4.20 shows how density of subgraphs relates between G dir and G2. For all subgraphs except Argentina, the density is significantly lower in G2. Argentina has only very few nodes left in G2, and they all are connected to each other, that is why the density is 1. Especially lower is the density of the UK-office. Figure 4.21 shows how density and clustering correlate. Compared with the same plots for G dir (see 4.10) and G1 (4.21), the clustering of subgraphs is higher relatively to their density. Highest clustering show the nodes with the highest ratio (US offices, Spain, Brazil). However, clustering per subgraph is significantly lower for G2 than G dir (see figure 4.22). This is especially true for locations that have significantly decreased in size (Italy - 3, German offices, Austria).

4.2 Network between Locations

Figure 4.24 shows a weighted network of the total amount of emails between locations, and a matrix showing the total amount of emails between and inside locations on a log-scale. The rows are the senders and the columns are the recipients. Element $A_{i,j}$ stands for the total amount of emails sent between location i and location j. While most communication happens between nodes of the same location (the highest being Spain, followed by Brazil and U.S. 2), it is also interesting to notice that every location has a quite intense communication with the head quarter. Most emails between locations are exchanged between the Spain offices and the head quarter. There are also other clusters of high communication - between the two U.S. offices (also described later in 4.3 and 4.4), the three German locations, Italy 3 and 4, and Argentina and Brazil. This figure also shows locations that are not connected with many others - Argentina and Turkey, who both sent emails to 5 other locations, but only get emails back from 2. Also, Germany is very far from Argentina and Brazil. Figures 4.27 shows a plot of those networks on a world map.

Figure 4.25 and figure 4.26 show the weighted networks between locations, derived from G1 and G2, and the adjacency matrix. The weights are the amount of connections between the two locations, normalized by the number of total connections. I.e, for a node from location i, $A_{i,j}$ stands for the probability that a connected node is at location j. In figure 4.25, the clusters identified in figure 4.26 (between U.S. offices, Argentina and Brazil, and Germany) are still visible. However, most connections are inside locations, and between the locations and the headquarter. The strongest connection with the head quarter have Italy 2 and Italy 3.

In figure 4.26, standing for the tight communication, it is clearly visible that most tight interaction happens between nodes of the same location. An exception being Germany 2, who has more connections with Germany 1 than themselves (as already described earlier). However, close collaboration also exists between the U.S. offices and the German offices, and almost all locations (except Germany 3 and Argentina) are still connected to the head quarter. Relatively many degrees, to be precise, 5, have Spain, Italy 3 and U.S. 2, making them hubs in the tight communication network, and for close collaboration. Figure 4.28 shows a plot of this network on a map of Europe.

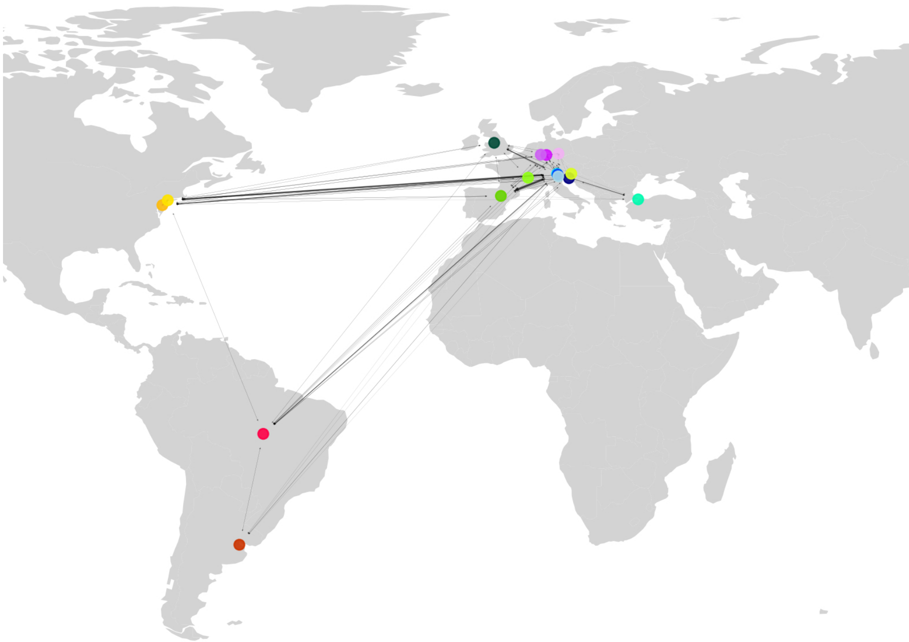


Figure 4.27: Total communication on world map

Figure 4.28: Tight communication on Europe map

4.3 Connectivity and Cohesion

The k-core decomposition of the networks show that the cohesion is high, since at every step of the decomposition, the networks still have only one core. It further shows that the tightest knit communities are the two US-offices, for both the loose and the tight network.

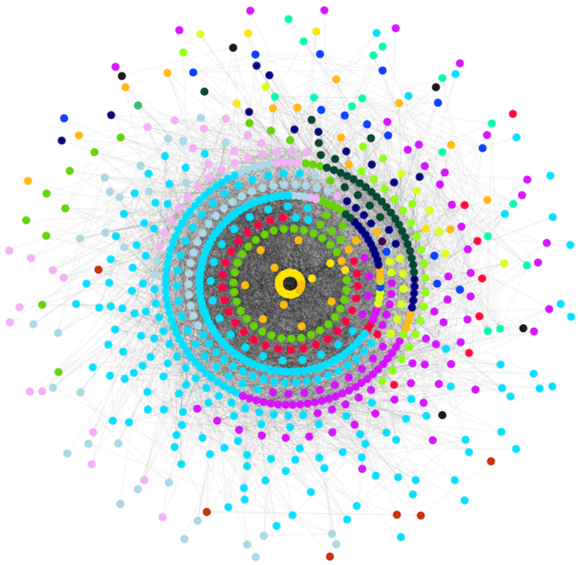


Figure 4.29: k-cores G1

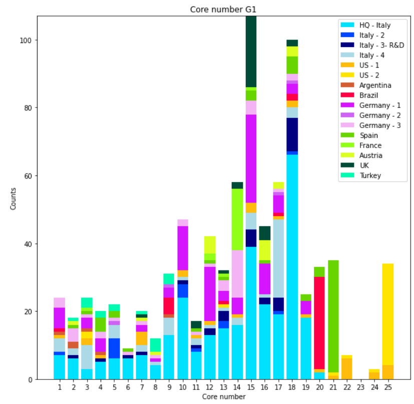


Figure 4.30: Core number histogram G1

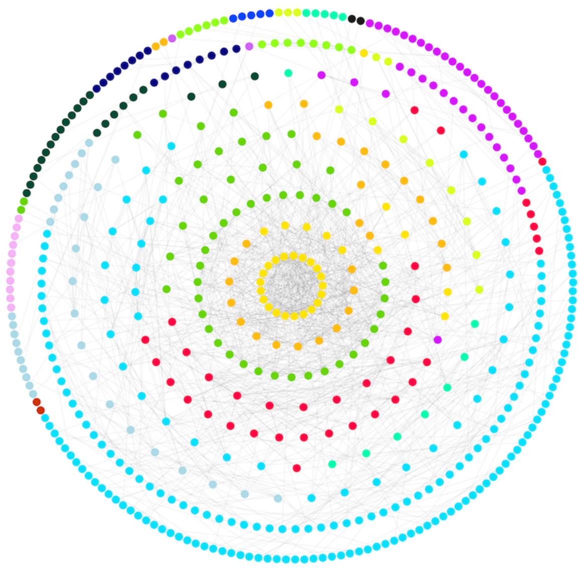


Figure 4.31: k-cores G2

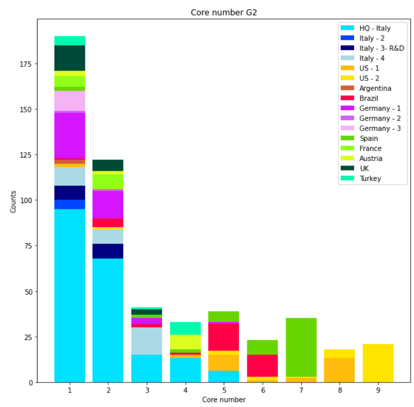


Figure 4.32: Core number histogram G2

Figure 4.29 and figure 4.31 show a plot of the shells of the networks G1 and G2. The core of both networks are the nodes located in the two U.S.-offices. Figure 4.30 and figure 4.32 show a histogram of the same measurement: The highest k-core number for G1 is 25, and for G2 it is 9. This means that the core of the networks consists of nodes having at 25 respective 9 connection with each other, and are forming a connected component. For both networks, the most central shells are composed

purely of nodes in the U.S.-offices, followed by nodes located in the Spain-office, and in Brazil.
Nodes in the headquarter are, especially in the tight network G2, mostly located at the less central shells of the networks. This is due to the fact that nodes at other locations (especially U.S.) are very tightly communicating with each other, (which can also be seen in the higher clustering and density, see 4.1.1, and high communication between the two locations, see 4.25), while the nodes at the headquarter are more spread out.

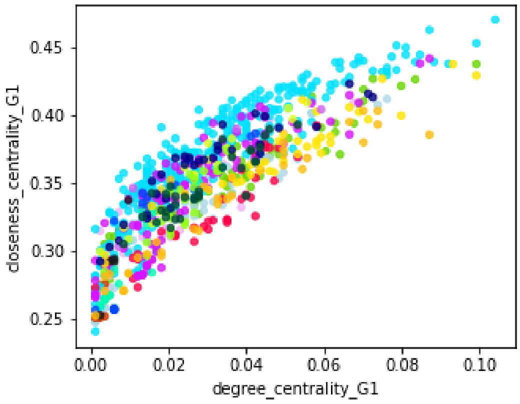


Figure 4.33: Degree vs. closeness centrality, G1

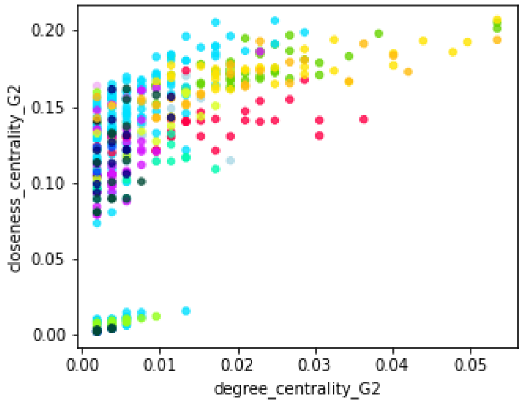


Figure 4.34: Degree vs. closeness centrality, G2

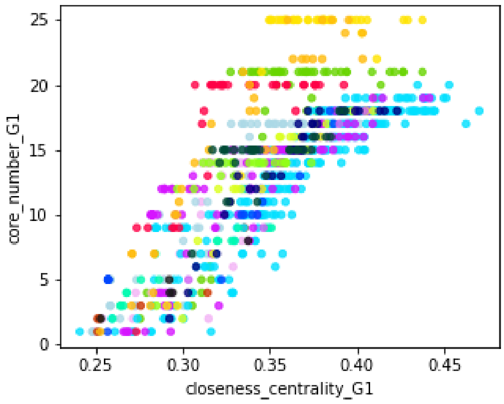


Figure 4.35: Closeness centrality vs. core number, G1

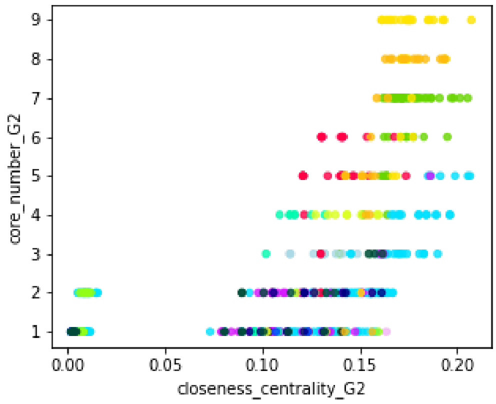


Figure 4.36: Closeness centrality vs. core number, G2

Figure 4.33 shows how degree centrality and closeness centrality per node correlate in G1. The colors stand for the locations. What can be seen here is that the nodes in the headquarter (turquoise) are the ones with highest closeness centrality overall, in relation to their degree centrality; while the nodes that have a high core number (U.S. locations and Spain) have relatively low closeness centrality compared to nodes with similar degrees. The nodes with highest closeness centrality are without exception nodes from the headquarter. They have the lowest distance to all other nodes in the network, as described in section 2.2.5.4. In the plot for the tight network G2 (figure 4.34), the U.S. and Spain nodes come closer to the center, while also having higher degree than the nodes from the head quarter. The networks portrait in figure 4.12 (G1) and 4.18 (G2) are plotted with nodes with higher closeness centrality in the center of the plot. Here again, the most central ones are the nodes from the headquarter. Figure 4.35 and 4.36 show how closeness centrality and core number correlate in network G1. While closeness and core number do in general correlate (see also section 4.4), the nodes in the headquarter have higher closeness, the nodes in the U.S. offices and Spain have higher core number, as already described.

4.4 Node-level measurements

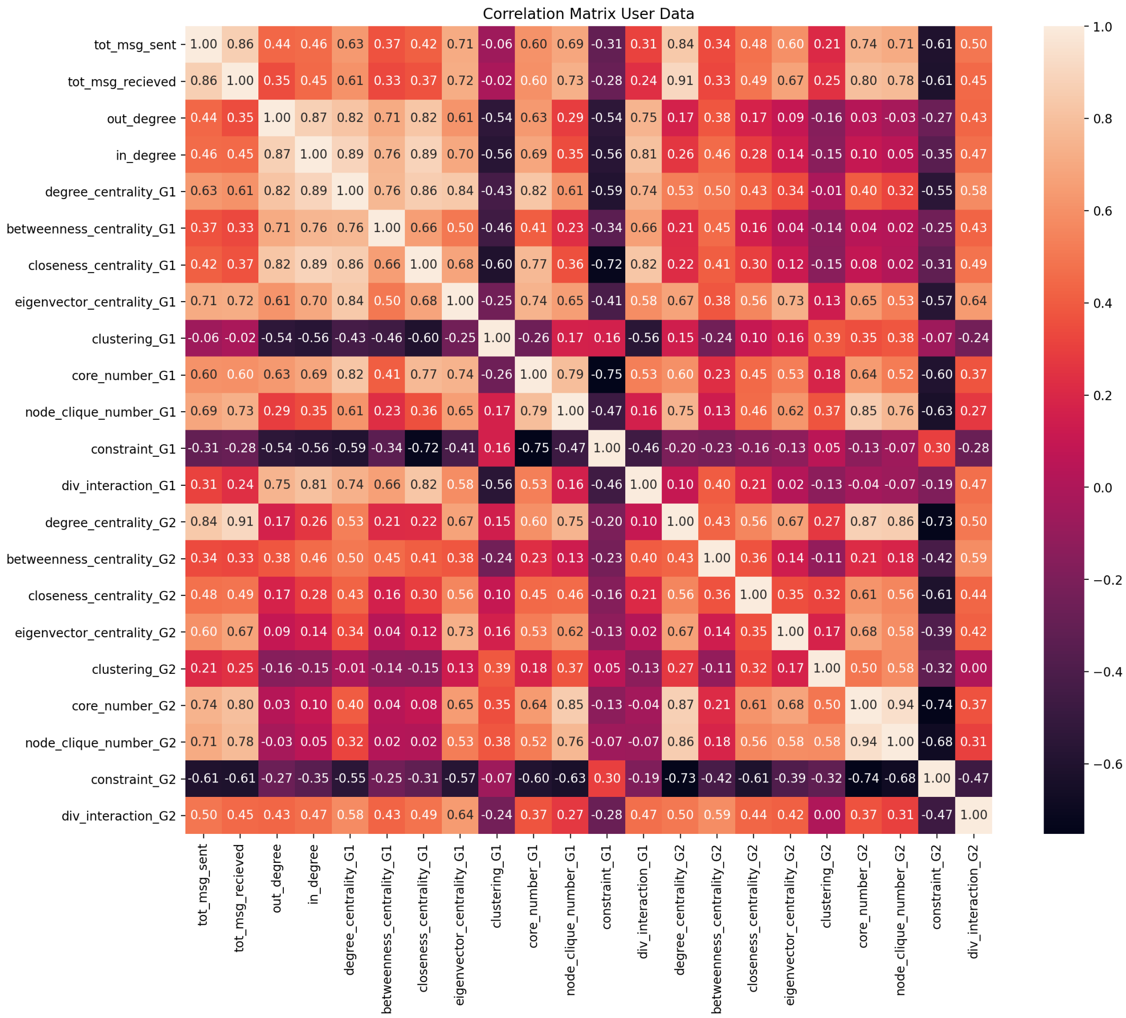


Figure 4.37: Correlation Matrix Users Data

Figure 4.37 shows the correlation matrix of user data measurements. It is actually total messages received, and not total messages sent, that indicates how central the node is in G1 and G2. The correlation is particularly high between amount of messages received and the degree centrality in G2, and eigenvector centrality in both G1 and G2. Core number correlates highly with constraint, both for network G1 and for G2. That means that nodes in the core of the network, determined by the k-core decomposition (see section 4.3), are more likely to have structural holes in their ego networks, and therefore advantages, as described in section 2.2.5.5. Core number further correlates with node clique number.

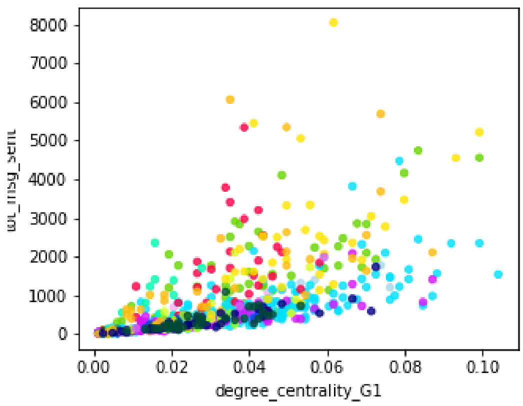


Figure 4.38: Degree vs total messages sent, G1

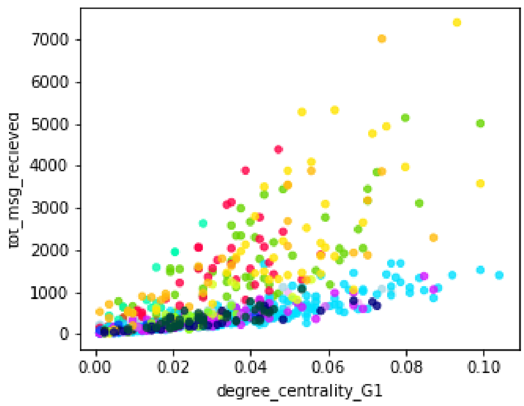


Figure 4.39: Degree vs total messages received, G1

Figure 4.38 and 4.39 show how total amount of messages sent (resp. received) correlates with degree centrality in the network G1. It is clearly visible that it is especially nodes from U.S. offices and Brazil that received and sent a lot of emails, even when they have a relatively low degree centrality. In the headquarter, nodes exist with very high degree centrality, but a low amount of total messages sent and received.

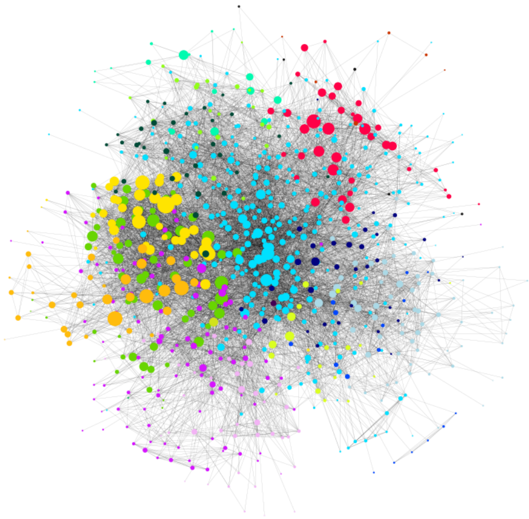


Figure 4.40: Total messages sent, G1

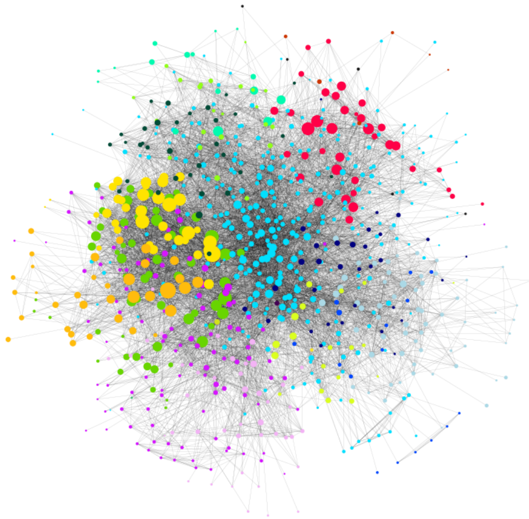


Figure 4.41: Total messages received

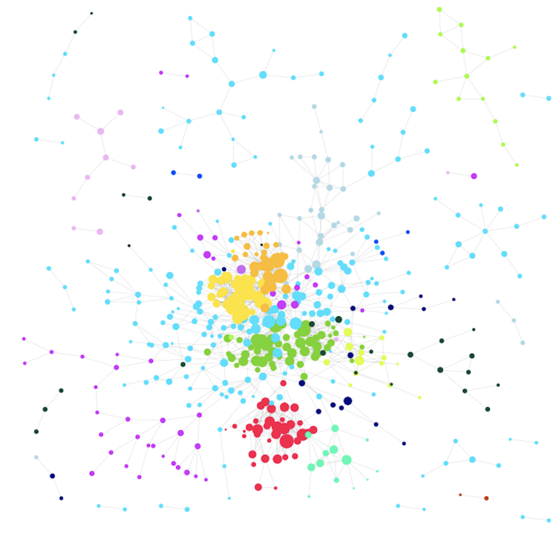


Figure 4.42: G2: Total messages sent

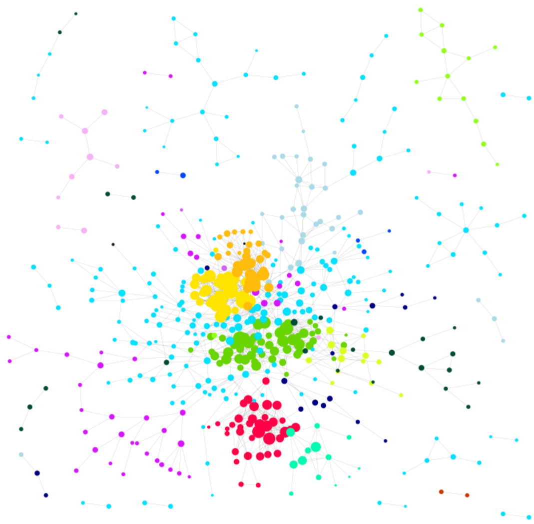


Figure 4.43: Total messages received

Figures 4.40 and 4.41 are plots of the network G1, with node size proportional to total messages received and sent by this node. It is clearly visible that the nodes with most messages received and sent are from the US, Spain, Brazil, which are the locations at the core of the network (see 4.3), and also the one with highest node ratio in G2 (see 4.1.4). Nodes from the headquarter sent more messages than they received.

Figure 4.42 and 4.43 show the same plots for network G2. Again, US, Brazil and Spain are important.

4.4.1 Centralities

Figure 4.44 and 4.45 show networks G1 and G2, with sizes of nodes proportional to their degrees. Figure 4.46 and 4.47 show a degree distribution with locations. While the highest degree, 86, in the network G1 is hold by a node in the head quarter, nodes with degrees above 60 are quite diverse. In G2, the highest degree is 28, hold by a node in Spain and in the U.S., apart from that, degrees above 20 are purely hold by nodes in the U.S. The tight collaboration of those nodes also leads to them being the core of the network, as discussed in 4.3.

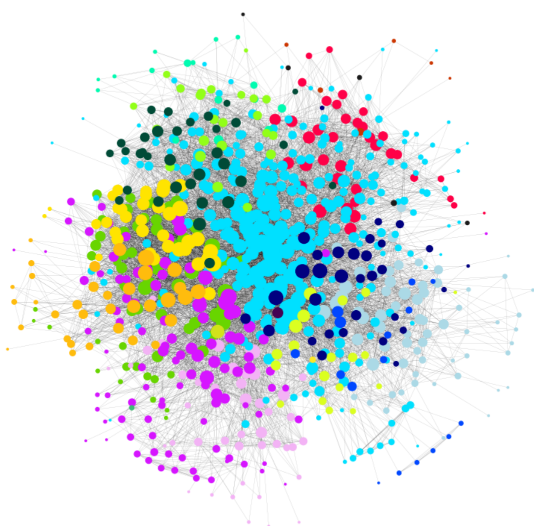


Figure 4.44: Degree G1

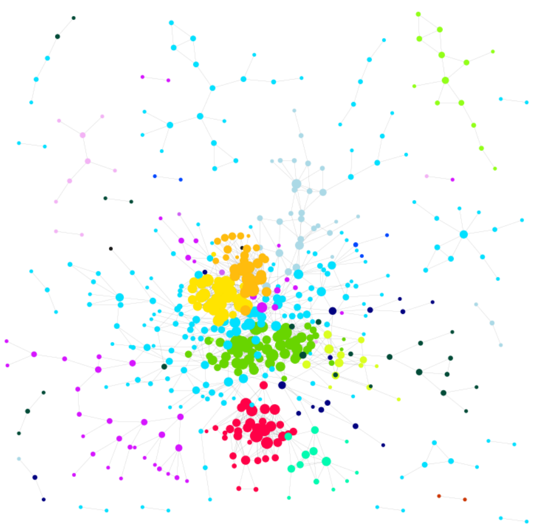


Figure 4.45: Degree G2

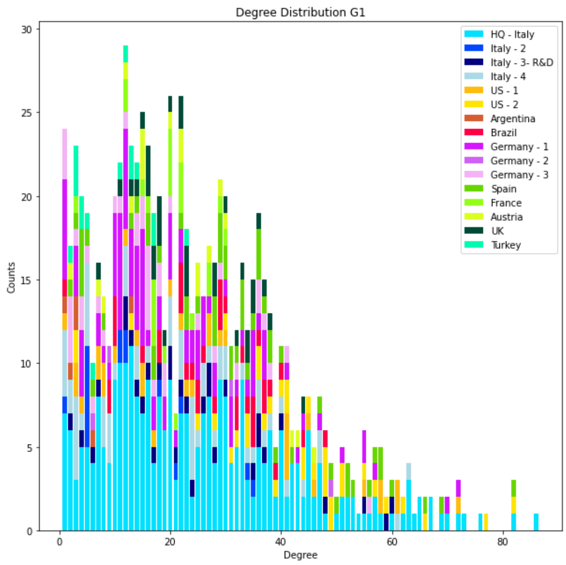


Figure 4.46: Degree distributions with locations, G1

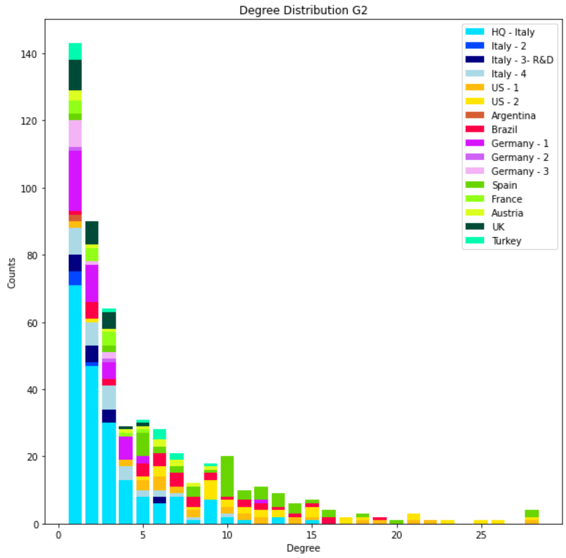


Figure 4.47: Degree distributions with locations, G2

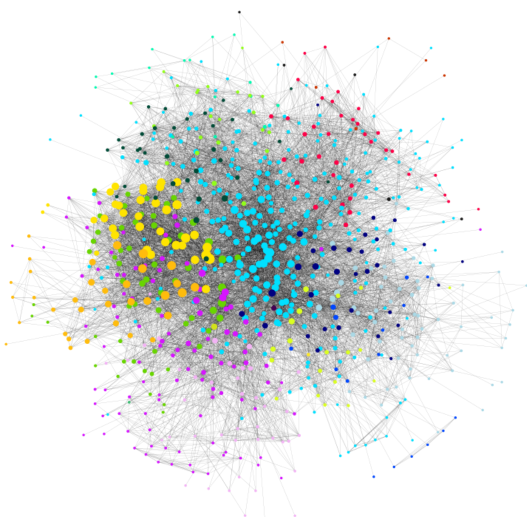


Figure 4.48: Eigenvector centrality G1

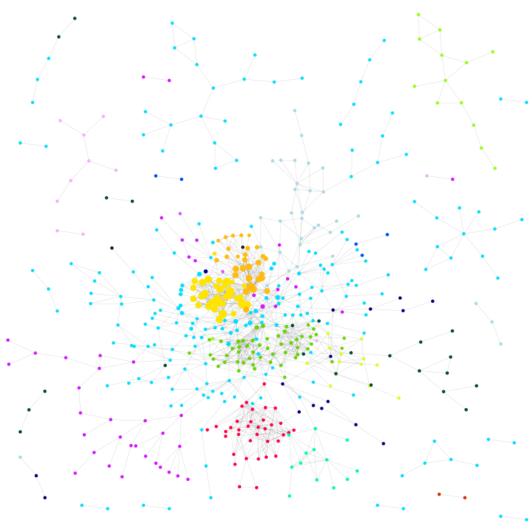


Figure 4.49: Eigenvector centrality G2

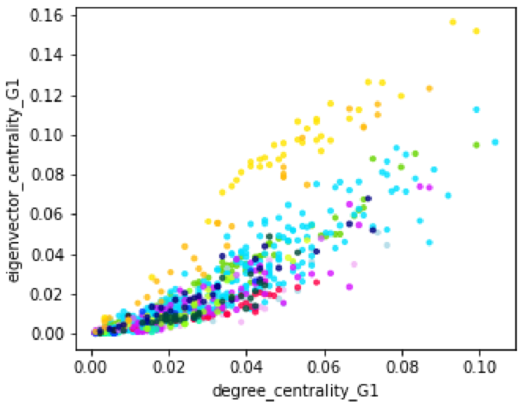


Figure 4.50: Degree centrality vs. eigenvector centrality G1

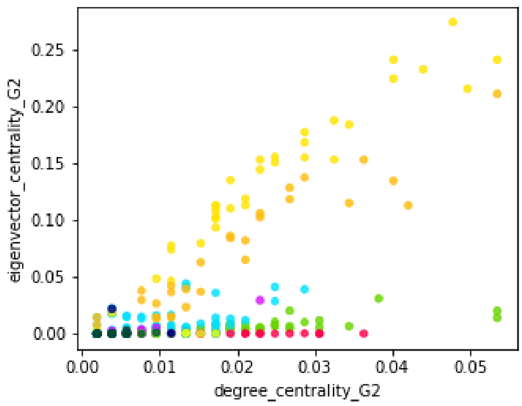


Figure 4.51: Degree centrality vs. eigenvector centrality G2

Figure 4.48 and 4.49 show networks G1 and G2 with node size proportional to eigenvector centrality of the node. Figure 4.50 and figure 4.51 show correlations between degree and eigenvector centrality for G1 and G2. The nodes with high eigenvector centrality in G1 are solely nodes from the U.S. offices, followed by the nodes in the headquarter.

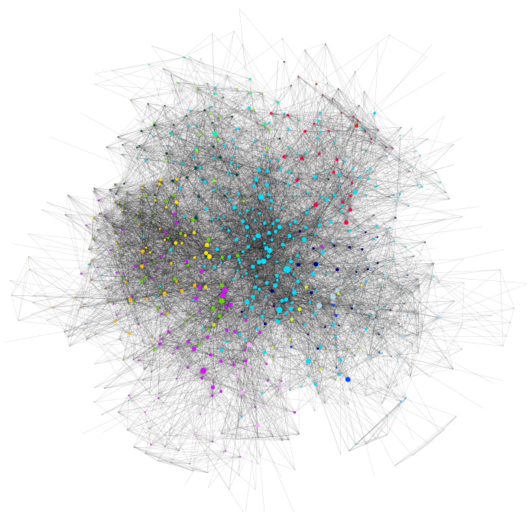


Figure 4.52: Betweenness centrality G1

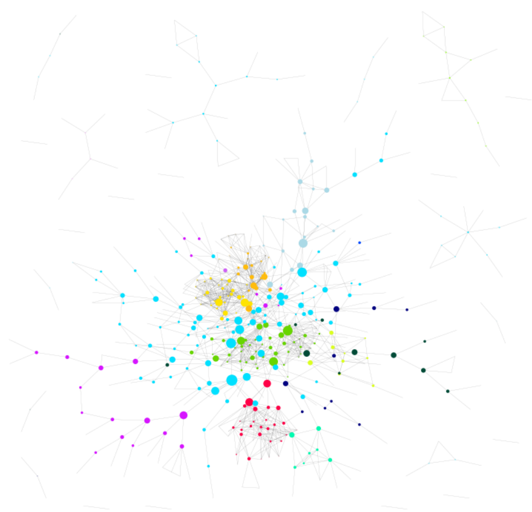


Figure 4.53: Betweenness centrality G2

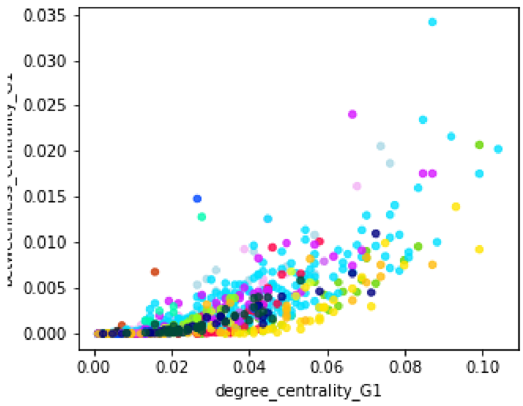


Figure 4.54: Degree centrality vs. betweenness centrality G1

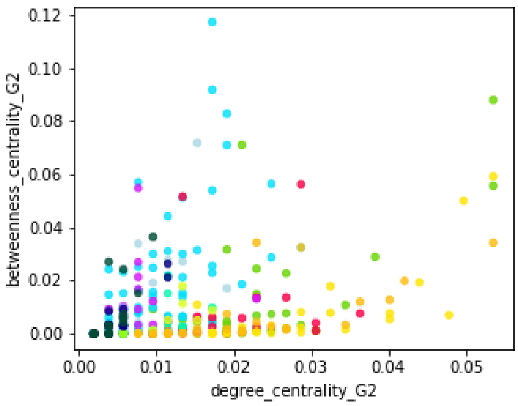


Figure 4.55: Degree centrality vs. betweenness centrality G2

Figure 4.52 and 4.53 show networks G1 and G2 with node size proportional to betweenness centrality of the node. Figure 4.54 and figure 4.55 show correlations between degree and betweenness centrality for G1 and G2. In general, nodes from the headquarter have highest betweenness centrality in relation to their degree, compared with nodes from other locations. However, there are single nodes that have very high betweenness centrality, because they probably function as a bridge between different parts of the network. In G1, they come from the headquarter, from Italy 2 and 4, France, Germany and Argentina. In G2, they come from the headquarter, Spain and Italy 4. In figure 4.53, it is possible to see pretty clearly what parts of the network they connect.
For discussion of closeness centrality, see section 4.3.

4.4.2 Clustering

Looking at the correlation matrix (Figure 4.37), node-level clustering in G1 (or density of the ego-network), correlates negatively with degree centralities and diversity of interaction in G1. The non-central nodes and the ones that do not interact with many different nodes are more likely to be in tight clusters. There is almost no correlation between clustering in G2 and centralities in G1 and G2. But there is high correlation between clustering, core number and node clique number in G2. Nodes in big cliques are more likely to have a dense ego network (since many of the neighbors are probably in the same big clique), and they are also in the core of the network. As already discussed in 4.3 and 4.1.4, nodes with high core number come from very highly clustered locations.

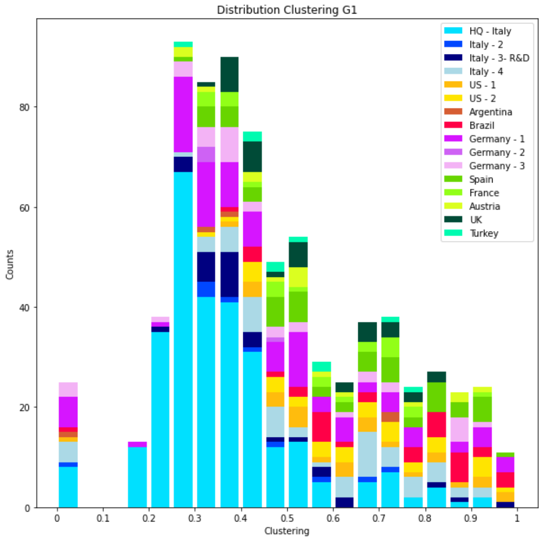


Figure 4.56: Distribution node-level clustering G1

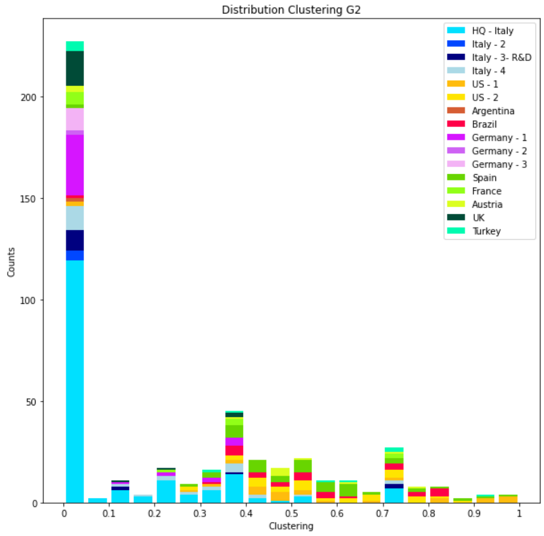


Figure 4.57: Distribution node-level clustering G2

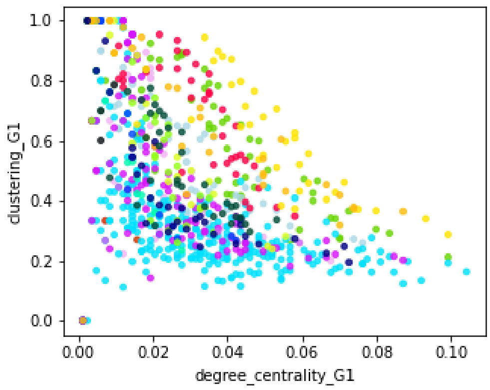


Figure 4.58: Degree centrality vs. clustering G1

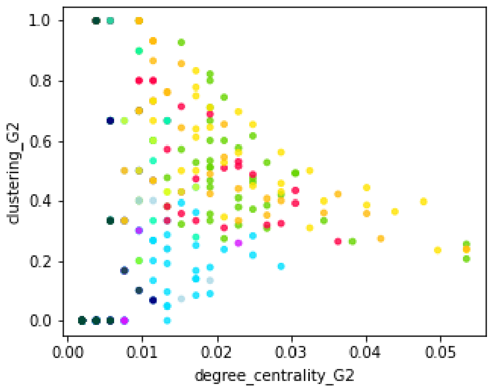


Figure 4.59: Degree centrality vs. clustering G2

Figure 4.56 and 4.57 show how the node-level clustering (or the density of the ego-network (see section 2.2.3.2)), are distributed in network G1 and G2. Figure 4.58 and 4.59 show how clustering and degree correlate on the node level, for network G1 and G2. Interesting here is that some nodes in both networks have density 1 (all their neighbors are connected with each other), but they all have very low degree (only a few neighbors). The distribution of locations is similar to the k-core picture (see figure 4.3). U.S. and Spain nodes, followed by Brazil, tend to have higher clustering with the same degree (as opposed to other nodes), as they tend to have a high core number. The nodes from the headquarter, on the other hand, have relatively low degree and clustering. This pattern is visible for both G1 and G2. In

G2, it is visible that nodes with a significant degree centrality (higher than 0.015), only show high clustering if they are part of offices in the U.S., Spain or Brazil.

4.4.3 Diversity of Interaction

Diversity of interaction, which is the number of different locations the node communicates with, correlates highly with (out- and in-) degree of the overall directed network and G1, but not with the degree of G2, or with other centralities of G2 (see figure 4.37) This is probably because many of those diverse interactions are not very tight, and fall away for the tight network. Communicating with a diverse set of nodes loosely does therefore not lead to more central positions (degree, eigenvector, betweenness or closeness) in the tight communication network. Figure 4.60 shows how the diversity of interactions is distributed over the number of nodes, and the locations. The highest diversity for both G1 and G2 is hold by the nodes in the headquarter. This is reasonable, since the head quarter contains corporate functions which demand contacts with different locations. No node has contacts with all 16 locations. For the tight network, most nodes only communicate closely with nodes from one location, while only a few communicate with nodes from more than 3 locations.

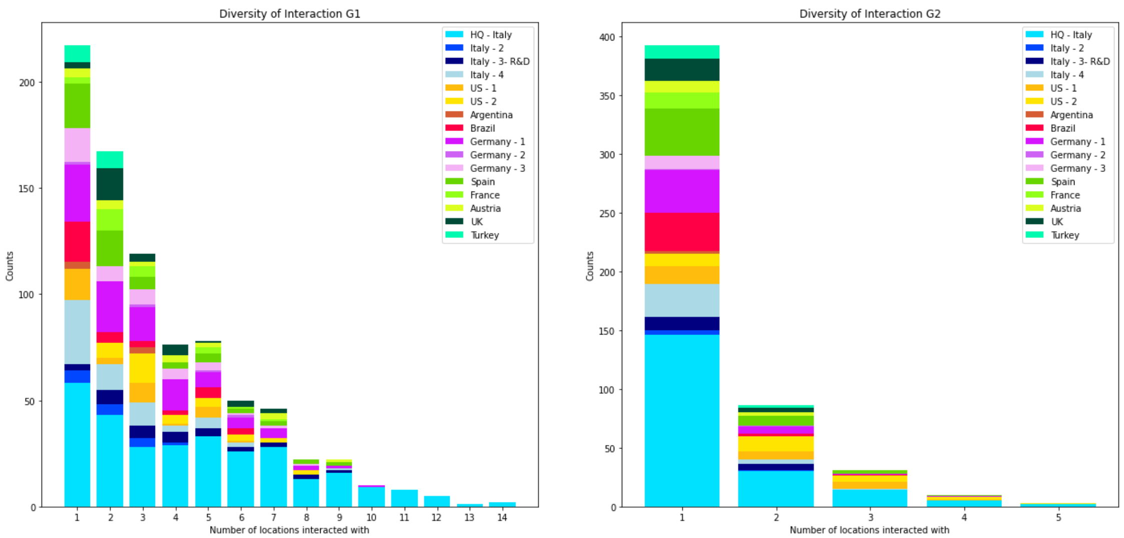


Figure 4.60: Diversity of interactions, histogram for G1 and G2

4.4.4 Structural holes

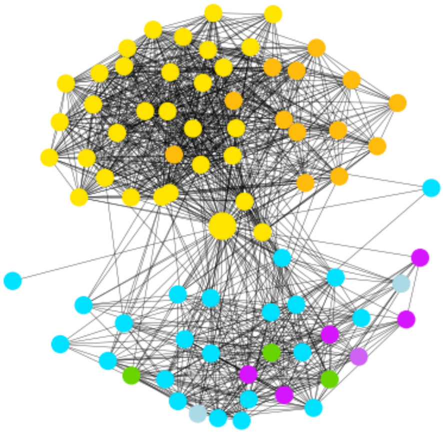


Figure 4.61: Ego network node with lowest constraint

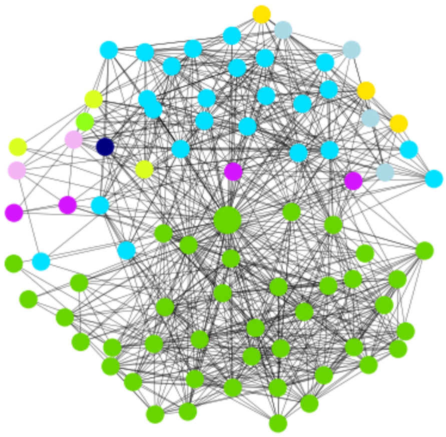


Figure 4.62: Ego network node with second lowest constraint

After calculating the constraint of all the nodes, how do the ego networks of nodes with low constraint, i.e., the ones that have many structural holes in their network (see 2.2.5.5), look like?
Figure 4.61 and 4.62 show the ego networks of the two nodes with lowest constraint in the tight network (G2). It shows pretty clearly that there are cliques of intense interaction that the node connects. As described in the previous section, a low constraint correlates with high core number.

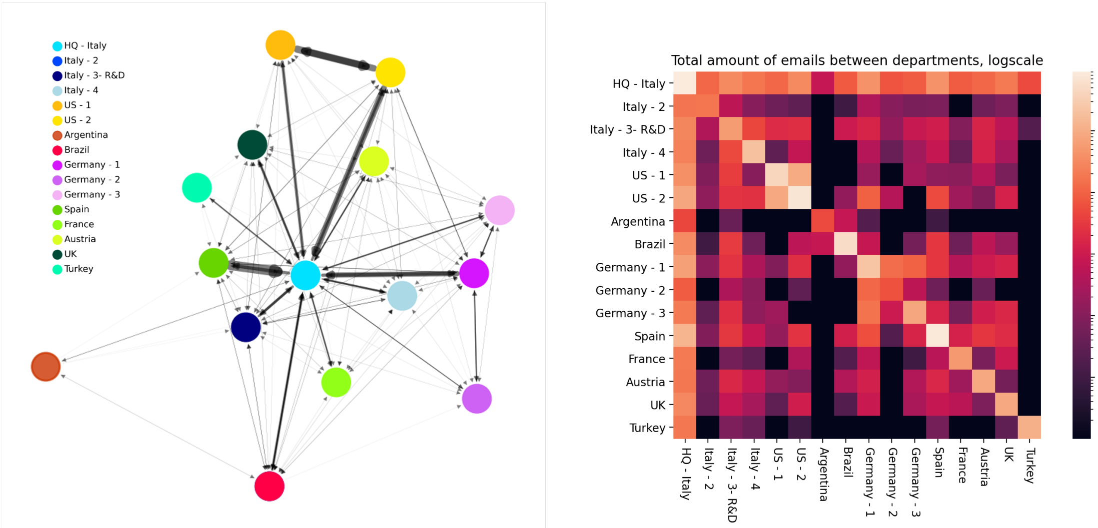


Figure 4.24: Total communication between locations

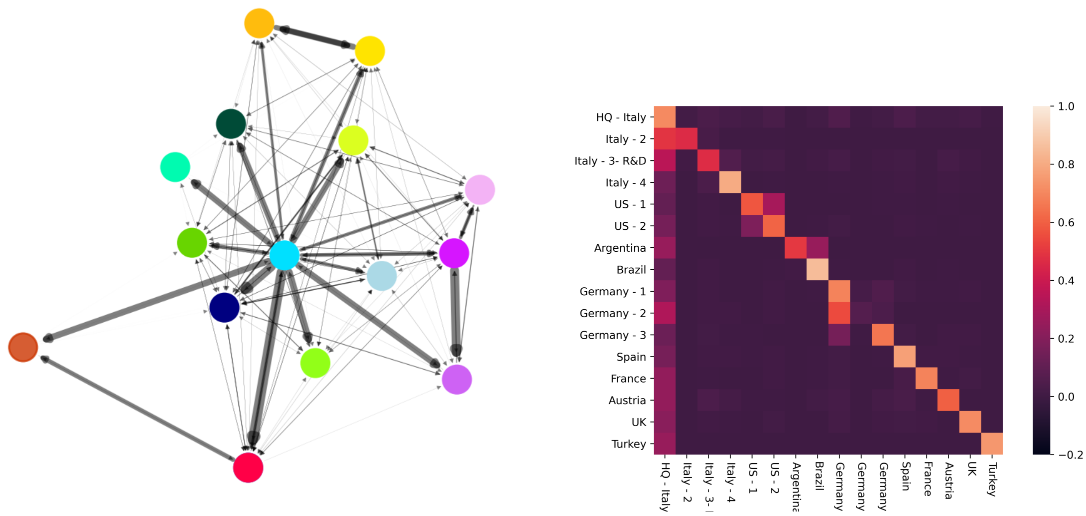


Figure 4.25: Network of locations, G1

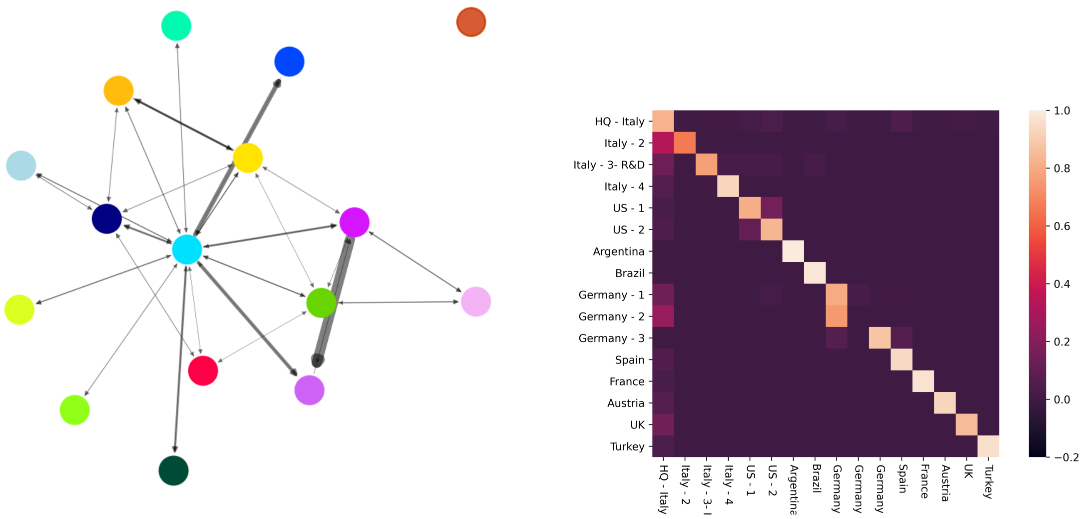


Figure 4.26: Network of locations, G2

5

Conclusion

The chapter 2 describes why networks, and network structure matters in organisations. In particular because of knowledge creation, storage and application, distribution and access to resources, innovation and new idea creation are collective processes and happen on networks. Further, network position matters for the individual, not least because of engagement and motivation. Chapter Methods describes the network analysis that is done, and chapter Results the results of the analysis.

5.1 Summary of Results

The analyses of the three networks: the overall directed network, the loose network and the tight network, all show that the organisation's 900 active email-users, and it's 16 locations, are connected quite well with each other. Despite spanning several functions, locations and time zones, the networks show low average shortest path lengths and high connectivity. The analysis is especially focused on comparing the different locations with each other, and examining the connectivity between locations. The head quarter has both loose and strong connections with almost every other location. This is to be expected, since it is the location with the most nodes, and holds corporate functions which demand collaboration and communication between various parts of the network and the company. The nodes of the head quarter have highest closeness centrality, making them the center of the network, and relatively high betweenness centrality, making them important bridges that connect unconnected parts of the network.
A special case are the U.S. offices, for the following reasons: The nodes there have relatively high degree, and receive and sent more messages than the average user. The subgraph also shows high density and clustering. Further, they are the highly connected core of the network, identified by k-core decomposition. On top of that, they show significantly higher eigenvector centrality than other nodes, meaning that they have a high outreach over the overall network. Apart from the H.Q. and the U.S. offices, Spain and Brazil are important locations in the network. They function as hubs in the tight communication network, get and receive many emails, and have a relatively high core numbers.
Germany - 2, on the other hand, is a small location and has very low clustering and density. It acts almost as if it were part of the close by location Germany - 1 than its own entity. Measurements on the node-level identify individuals with important functions for the overall network. Individuals with high betweenness centrality function as bridges that connect several otherwise poorly connected parts. Individuals

with low constraint have structural holes in their ego networks. They connect several tight clusters and sit a special positions for receiving novel information, function as translators and are exposed to different point of views.

5.2 Ethical considerations

Email data is personal data, and requires therefore care in handling. As opposed to other quantitative social studies, network data can by its very nature not be anonymised, since the connections between specific people matter. Further, it is hard to deal with missing data when for example several people decide not to take part in the study, because the structure and connectivity of the network is significantly altered by removing nodes [69]. However, there are a few considerations with that. First, the scope of network analysis studies is not the content of communication. The important aspect is whether communication happened or not, i.e., the meta data, which is much less invasive in privacy matters. Second, the focus is not on the actions of an individual. Not the exact communication patterns of individuals matter or are examined, but rather the structure that arises from the system level communication patterns. The scope of a organisational network analysis should not be to surveil, control or punish individual behaviour. Rather, it should be used to understand and improve connectivity, collaboration and communication patterns on a macro level. Zooming into the node level (which was above the scope of this thesis) can be used to identify people with high social influence, at central positions, or people acting as bridges. Those can be utilized as for example change makers, or be considered in promotions or for reorganisations. This does however also mean that the analysis could potentially be used to punish people with low centrality, and has therefore to be done with care.

5.3 Limitations and Future Applications

Due to technical reasons, it was only possible to get email data from the past 90 days. The analysis is therefore a mapping of the communication and collaboration that has happened in those specific 90 days. Analysing data over a longer time span might have led to more balanced results, and would have allowed to take a dynamical perspective. Email data alone might not capture all the communication that has happened. Among others, it does not include communication acts by phone, online meetings, or talking to each other in person. However, people who have an online meeting are also likely to send emails to each other. And, given the different locations spread over Europe and the Americas, email data is after all a good approximation for the overall communication happening.
As a further application, the technical accounts that were filtered out for this analysis could be included. However, that would require a lot more effort in preprocessing of the data, and also manual selection and more information about the nature of the accounts in question. Another future application is to compare the network structure with the official hierarchy of the company. Even though this data was available, it was not available in a format that would have allowed a comparison

without putting a lot of effort in the mining and preprocessing of the hierarchy data. Another application is to look closer at node-level data and identify people based on that. This can also be done in combination with some other measurements, for example engagement, HR data or performance reports, to find correlations between network positions and those values. However, this raises more ethical questions because of the use of personal data, and also, it was not the scope of this thesis. This network can also be used as a structure to model and understand for example spread of beliefs, or adaption to new technology, or conflicts in the organisation.

Bibliography

[1] R. Condorelli *et al.*, “Complex systems theory: some considerations for sociology,” *Open Journal of Applied Sciences*, vol. 6, no. 07, p. 422, 2016.

[2] D. Tollefsen, “Social ontology,” *Philosophy of Social Science: A New Introduction*, pp. 85–101, 2014.

[3] M. Mitchell, *Complexity: A guided tour*. Oxford University Press, 2009.

[4] D. Krackhardt and J. R. Hanson, “16 informal networks: The company behind the chart,” *Creative Management*, p. 202, 2001.

[5] M. O. Jackson, *Social and economic networks*. Princeton university press, 2010.

[6] S. E. Page, “What sociologists should know about complexity,” *Annual Review of Sociology*, vol. 41, pp. 21–41, 2015.

[7] E. Hutchins, *Cognition in the Wild*. No. 1995, MIT press, 1995.

[8] S. E. Page, *Diversity and complexity*, vol. 2. Princeton University Press, 2010.

[9] K. M. Robertson, J. O’Reilly, and D. R. Hannah, “Finding meaning in relationships: The impact of network ties and structure on the meaningfulness of work,” *Academy of Management Review*, vol. 45, no. 3, pp. 596–619, 2020.

[10] M. E. Newman, “The structure and function of complex networks,” *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[11] M. Newman, *Networks*. Oxford university press, 2018.

[12] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, “Network analysis in the social sciences,” *science*, vol. 323, no. 5916, pp. 892–895, 2009.

[13] J. Kleinberg, “The convergence of social and technological networks,” *Communications of the ACM*, vol. 51, no. 11, pp. 66–72, 2008.

[14] K. R. PM, A. Mohan, and K. Srinivasa, *Practical Social Network Analysis with Python*. Springer, 2018.

[15] W. Khaouid, M. Barsky, V. Srinivasan, and A. Thomo, “K-core decomposition of large networks on a single pc,” *Proceedings of the VLDB Endowment*, vol. 9, no. 1, pp. 13–23, 2015.

[16] J. Moody and D. R. White, “Structural cohesion and embeddedness: A hierarchical concept of social groups,” *American sociological review*, pp. 103–127, 2003.

[17] S. P. Borgatti and D. S. Halgin, “On network theory,” *Organization science*, vol. 22, no. 5, pp. 1168–1181, 2011.

[18] J. R. Methot and E. Rosado-Solomon, “Multiplex relationships in organizations,” *Social Networks at Work*, 2019.

[19] E. R. Crawford and J. A. LePine, “A configural theory of team processes: Accounting for the structure of taskwork and teamwork,” *Academy of Management Review*, vol. 38, no. 1, pp. 32–48, 2013.

[20] T. Casciaro, "Networks and affect in the workplace," *Social Networks at Work*, 2019.

[21] R. M. Fernandez, E. J. Castilla, and P. Moore, "Social capital at work: Networks and employment at a phone center," *American journal of sociology*, vol. 105, no. 5, pp. 1288–1356, 2000.

[22] S. P. Borgatti and P. C. Foster, "The network paradigm in organizational research: A review and typology," *Journal of management*, vol. 29, no. 6, pp. 991–1013, 2003.

[23] A.-L. Barabási, "Scale-free networks: a decade and beyond," *science*, vol. 325, no. 5939, pp. 412–413, 2009.

[24] P. Erdös and A. Rényi, "On the evolution of random graphs," in *The structure and dynamics of networks*, pp. 38–82, Princeton University Press, 2011.

[25] A.-L. Barabási and E. Bonabeau, "Scale-free networks," *Scientific american*, vol. 288, no. 5, pp. 60–69, 2003.

[26] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical processes on complex networks*. Cambridge university press, 2008.

[27] R. Albert, H. Jeong, and A.-L. Barabási, "Diameter of the world-wide web," *nature*, vol. 401, no. 6749, pp. 130–131, 1999.

[28] D. d. S. Price, *Networks of scientific papers*. Princeton University Press, 2011.

[29] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[30] S. Aparicio, J. Villazón-Terrazas, and G. Álvarez, "A model for scale-free networks: application to twitter," *Entropy*, vol. 17, no. 8, pp. 5848–5867, 2015.

[31] H. Ebel, L.-I. Mielsch, and S. Bornholdt, "Scale-free topology of e-mail networks," *Physical review E*, vol. 66, no. 3, p. 035103, 2002.

[32] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *nature*, vol. 406, no. 6794, pp. 378–382, 2000.

[33] A. H. Dekker and B. Colbert, "Scale-free networks and robustness of critical infrastructure networks," in *7th Asia-Pacific Conference on Complex Systems*, pp. 685–699, 2004.

[34] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[35] A. D. Broido and A. Clauset, "Scale-free networks are rare," *Nature communications*, vol. 10, no. 1, pp. 1–10, 2019.

[36] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[37] J. Scott and P. J. Carrington, *The SAGE handbook of social network analysis*. SAGE publications, 2011.

[38] S. Milgram, "The small world problem," *Psychology today*, vol. 2, no. 1, pp. 60–67, 1967.

[39] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna, "Four degrees of separation," in *Proceedings of the 4th Annual ACM Web Science Conference*, pp. 33–42, 2012.

[40] S. A. Myers, A. Sharma, P. Gupta, and J. Lin, "Information network or social network? the structure of the twitter follow graph," in *Proceedings of the 23rd International Conference on World Wide Web*, pp. 493–498, 2014.

[41] B. Uzzi and J. Spiro, "Collaboration and creativity: The small world problem," *American journal of sociology*, vol. 111, no. 2, pp. 447–504, 2005.

[42] S. P. Borgatti, "Centrality and network flow," *Social networks*, vol. 27, no. 1, pp. 55–71, 2005.

[43] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics," *Reviews of modern physics*, vol. 81, no. 2, p. 591, 2009.

[44] P. Ball, *Why society is a complex matter: Meeting twenty-first century challenges with a new kind of science.* Springer Science & Business Media, 2012.

[45] M. Galesic and D. L. Stein, "Statistical physics models of belief dynamics: Theory and empirical tests," *Physica A: Statistical Mechanics and its Applications*, vol. 519, pp. 275–294, 2019.

[46] D. Krackhardt and M. Kilduff, "Friendship patterns and culture: The control of organizational diversity," *American anthropologist*, vol. 92, no. 1, pp. 142–154, 1990.

[47] B. Pauksztat and D. Salin, "Targets' social relationships as antecedents and consequences of workplace bullying: a social network perspective," *Frontiers in psychology*, vol. 10, p. 3077, 2020.

[48] W. W. Powell, B. Staw, and L. Cummings, "Neither market nor hierarchy," 1990.

[49] K. J. Zollman, "Network epistemology: Communication in epistemic communities," *Philosophy Compass*, vol. 8, no. 1, pp. 15–27, 2013.

[50] A. Ebadi and A. Schiffauerova, "On the relation between the small world structure and scientific activities," *PloS one*, vol. 10, no. 3, p. e0121129, 2015.

[51] H. Kim and Y. Park, "Structural effects of r&d collaboration network on knowledge diffusion performance," *Expert Systems with Applications*, vol. 36, no. 5, pp. 8986–8992, 2009.

[52] L. Smith-Doerr and W. W. Powell, "Networks and economic life," *The handbook of economic sociology*, vol. 2, no. 3, pp. 379–402, 2005.

[53] U. Brandes, "On variants of shortest-path betweenness centrality and their generic computation," *Social Networks*, vol. 30, no. 2, pp. 136–145, 2008.

[54] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215–239, 1978.

[55] R. S. Burt, *Structural holes.* Harvard university press, 1992.

[56] R. S. Burt, "Structural holes and good ideas," *American journal of sociology*, vol. 110, no. 2, pp. 349–399, 2004.

[57] A. N. Raile, R. K. Kim, J. Choi, K. B. Serota, H. S. Park, and D. W. Lee, "Connections at work: How friendship networks relate to job satisfaction," *Communication Research Reports*, vol. 25, no. 2, pp. 168–178, 2008.

[58] M. P. Mundt and L. I. Zakletskaia, "Professional communication networks and job satisfaction in primary care clinics," *The Annals of Family Medicine*, vol. 17, no. 5, pp. 428–435, 2019.

[59] D. Krackhardt and L. W. Porter, "When friends leave: A structural analysis of the relationship between turnover and stayers' attitudes," *Administrative science quarterly*, pp. 242–261, 1985.

[60] P. A. Gloor, A. F. Colladon, F. Grippa, and G. Giacomelli, "Forecasting managerial turnover through e-mail based social network analysis," *Computers in Human Behavior*, vol. 71, pp. 343–352, 2017.

[61] T. H. Feeley, "Testing a communication network model of employee turnover based on centrality," 2000.

[62] J. Yuan, "Predicting employee turnover from network analysis," in *Journal of Physics: Conference Series*, vol. 1168, p. 032030, IOP Publishing, 2019.

[63] S. M. Soltis, F. Agneessens, Z. Sasovova, and G. Labianca, "A social network perspective on turnover intentions: The role of distributive justice and social support," *Human Resource Management*, vol. 52, no. 4, pp. 561–584, 2013.

[64] W. B. Schaufeli, A. B. Bakker, and M. Salanova, "The measurement of work engagement with a short questionnaire: A cross-national study," *Educational and psychological measurement*, vol. 66, no. 4, pp. 701–716, 2006.

[65] D. S. Halgin, G. M. Gopalakrishnan, and S. P. Borgatti, "Structure and agency in networked, distributed work: The role of work engagement," *American Behavioral Scientist*, vol. 59, no. 4, pp. 457–474, 2015.

[66] S. Guha, M. Muller, N. S. Shami, M. Masli, and W. Geyer, "Using organizational social networks to predict employee engagement," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 10, 2016.

[67] T. Mitra, M. Muller, N. S. Shami, A. Golestani, and M. Masli, "Spread of employee engagement in a large organizational network: a longitudinal analysis," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, pp. 1–20, 2017.

[68] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using networkx," in *Proceedings of the 7th Python in Science Conference* (G. Varoquaux, T. Vaught, and J. Millman, eds.), (Pasadena, CA USA), pp. 11 – 15, 2008.

[69] S. P. Borgatti and J. L. Molina, "Ethical and strategic issues in organizational social network analysis," *The Journal of Applied Behavioral Science*, vol. 39, no. 3, pp. 337–349, 2003.