**University of Zürich**

Department of Business Administration  
Institute for Network Science  
Prof. Dr. Claudio J. Tessone

Master thesis

**Mapping information spreading to   
improve communication:   
The case of ONE smart solution**

**Made by:**

Denis Krebs  
Herzogstrasse 9  
5000 Aarau  
077 411 00 14  
denis.krebs@uzh.ch  
Student ID: 11-747-912  
Faculty of Business, Economics and Informatics  
Number of Semesters: 13

Supervisor: Prof. Dr. Claudio J. Tessone  
Advisor: Dr. Zhao Yang

Submitted on: 31/01/2019

* 4 Bewertung einer wissenschaftlichen Arbeit

Bei der Bewertung der Arbeiten werden die folgenden 3 Bereiche geprüft:

1) Formaler Aufbau

Beurteilt werden Zitierweise, Abbildungen, Literaturverzeichnis, Formatierungen und Orthographie/Interpunktion die Vorgaben der Wegleitung.

2) Struktur

Ist ein roter Faden in der Arbeit erkennbar und wurde systematisch auf die Beantwortung der

Fragestellungen hingearbeitet? Sind die einzelnen Kapitel und Unterkapitel ausgewogen?

3) Inhalt

Wurde das formulierte Ziel der Arbeit erreicht und die daraus abgeleiteten Fragestellungen beantwortet? Wurde der aktuelle Forschungsstand anhand wissenschaftlicher Beiträge aufgezeigt?

Sind die Gedankengänge klar und die Wortwahl einer wissenschaftlichen Arbeit entsprechend?

Wurde die relevante Literatur zum Thema der Arbeit verarbeitet?

TIPP: Eine detaillierte Checkliste für wissenschaftliche Arbeiten, welche die 3 genannten Bewertungsbereiche noch konkretisiert und mit deren Hilfe man die eigene Leistung genau überprüfen kann, findet sich auf der nächsten Seite (Abbildung 8).

**Abstract**

…

# Table of contents

[Table of contents IV](#_Toc534648368)

[Table of figures VI](#_Toc534648369)

[List of abbreviations VII](#_Toc534648370)

[1 Introduction & Problem Statement - 1 -](#_Toc534648371)

[1.1 Scientific Research Statement - 1 -](#_Toc534648372)

[1.2 Research context and relevance - 1 -](#_Toc534648373)

[1.2.1 Scientific context - 1 -](#_Toc534648374)

[1.2.2 Business relevance - 2 -](#_Toc534648375)

[1.3 Terminology / structure of the thesis - 3 -](#_Toc534648376)

[2 Theory - 4 -](#_Toc534648377)

[2.1 Network Science - 4 -](#_Toc534648378)

[2.1.1 Characteristics of Network Science - 4 -](#_Toc534648379)

[2.1.2 Networks at the Heart of Complex Systems - 6 -](#_Toc534648380)

[2.1.3 Two Forces Helped the Emergence of Network Science - 6 -](#_Toc534648381)

[2.2 Tools - 7 -](#_Toc534648382)

[3 Research State & gaps - 9 -](#_Toc534648383)

[3.1 Email network analysis - 9 -](#_Toc534648384)

[3.2 Connectivity patterns - 9 -](#_Toc534648385)

[4 M-Industry & Project ONE - 12 -](#_Toc534648386)

[4.1 Migros Group - 12 -](#_Toc534648387)

[4.2 Migros Industry - 12 -](#_Toc534648388)

[4.3 ONE - 12 -](#_Toc534648389)

[5 Hypothesises - 14 -](#_Toc534648390)

[6 Concept & Method - 15 -](#_Toc534648391)

[6.1 Data - 15 -](#_Toc534648392)

[6.1.1 Email log - 15 -](#_Toc534648393)

[6.1.2 Questionnaire - 16 -](#_Toc534648394)

[6.2 Operationalization of the research question - 16 -](#_Toc534648395)

[6.3 Execution - 17 -](#_Toc534648396)

[6.3.1 Static network - 18 -](#_Toc534648397)

[6.3.2 Dynamic network - 19 -](#_Toc534648398)

[6.3.3 In-block nestedness - 20 -](#_Toc534648399)

[7 Results - 21 -](#_Toc534648400)

[7.1 Description of population & sample - 21 -](#_Toc534648401)

[7.1.1 Questionnaire - 21 -](#_Toc534648402)

[7.1.2 Social network analysis - 22 -](#_Toc534648403)

[7.2 Research questions & answers - 23 -](#_Toc534648404)

[7.2.1 SNA helps to map and understand the flow of communication - 24 -](#_Toc534648405)

[7.2.2 Hypothesis 1 & hypothesis 2 - 29 -](#_Toc534648406)

[7.2.3 Hypothesis 3 - 30 -](#_Toc534648407)

[7.2.4 SNA helps to understand the connections between individuals - 30 -](#_Toc534648408)

[7.2.5 Insights to improve communication - 31 -](#_Toc534648409)

[7.3 Summary of results - 33 -](#_Toc534648410)

[8 Conclusion - 35 -](#_Toc534648411)

[9 Recommendations for ONE - 37 -](#_Toc534648412)

[Bibliography - 38 -](#_Toc534648413)

[Appendix - 44 -](#_Toc534648414)

[A: Network science terms - 44 -](#_Toc534648415)

[B: Survey results - 47 -](#_Toc534648416)

[C: SKRIPT R - 54 -](#_Toc534648417)

[Achtung: Manuell Zeilenumbruch zwischen letztem Kapitel und Literaturverzeichnis einfügen]

# Table of figures

[Figure 1 Bridges of Konigsberg - 5 -](file:///C:\Users\Denis%20Krebs\Desktop\Thesis\Thesis_Working%20Document_20190107.docx#_Toc534621697)

[Figure 2 Visualization of an example social network. - 7 -](#_Toc534621698)

[Figure 3 Vertices in three communities - 9 -](file:///C:\Users\Denis%20Krebs\Desktop\Thesis\Thesis_Working%20Document_20190107.docx#_Toc534621699)

[Figure 4 Number of emails per day during the period under consideration - 23 -](file:///C:\Users\Denis%20Krebs\Desktop\Thesis\Thesis_Working%20Document_20190107.docx#_Toc534621700)

[Figure 5 Figure 5 Slicing the ONE dataset - 24 -](file:///C:\Users\Denis%20Krebs\Desktop\Thesis\Thesis_Working%20Document_20190107.docx#_Toc534621701)

[Figure 6 Dynamic evolution of the number of mails, edges and vertices - 25 -](file:///C:\Users\Denis%20Krebs\Desktop\Thesis\Thesis_Working%20Document_20190107.docx#_Toc534621702)

[Figure 7 Complete ONE network (simplified) - 26 -](file:///C:\Users\Denis%20Krebs\Desktop\Thesis\Thesis_Working%20Document_20190107.docx#_Toc534621703)

[Figure 8 Node level SNA measures on the DynSnap timescales - 28 -](file:///C:\Users\Denis%20Krebs\Desktop\Thesis\Thesis_Working%20Document_20190107.docx#_Toc534621704)

[Figure 9 Centralization of DynSnap networks - 29 -](file:///C:\Users\Denis%20Krebs\Desktop\Thesis\Thesis_Working%20Document_20190107.docx#_Toc534621705)

[Figure 10 Centrality measures - 29 -](file:///C:\Users\Denis%20Krebs\Desktop\Thesis\Thesis_Working%20Document_20190107.docx#_Toc534621706)

[Figure 11 Mutual information measure for core team members - 30 -](file:///C:\Users\Denis%20Krebs\Desktop\Thesis\Thesis_Working%20Document_20190107.docx#_Toc534621707)

[Figure 12 Mutual information measure for site group members - 31 -](file:///C:\Users\Denis%20Krebs\Desktop\Thesis\Thesis_Working%20Document_20190107.docx#_Toc534621708)

[Figure 13 Timeseries and boxplot of in-block nestedness - 32 -](file:///C:\Users\Denis%20Krebs\Desktop\Thesis\Thesis_Working%20Document_20190107.docx#_Toc534621709)

[Figure 14 ONE project plan - 33 -](file:///C:\Users\Denis%20Krebs\Desktop\Thesis\Thesis_Working%20Document_20190107.docx#_Toc534621710)

[Figure 15 In-block nestedness and number of holidays trend - 34 -](file:///C:\Users\Denis%20Krebs\Desktop\Thesis\Thesis_Working%20Document_20190107.docx#_Toc534621711)

[automatisches Abbildungsverzeichnis, rechte Maustaste 🡪 Felder aktualisieren]

[neue Abbildung im Text einfügen: Verweise 🡪 Beschriftung einfügen]

List of abbreviations

SNA Social Network Analysis

MGB (FME) Migros-Genossenschafts-Bund (Federation of Migros Cooperatives)

M-Industry Migros Industry

ONE One smart solution

ERP Enterprise Resource Planning

PMO Project Management Office

IBN In-block nestedness

NMI Normalized mutual information

BINMATNEST Binary matrix nestedness temperature calculator

NTC Nestedness temperature calculator

NODF Nestedness metric based on overlap and decreasing filling

FCM Fitness-complexity metric

HR Human Resources

[Tipp: Bei der ersten Verwendung der Abkürzung im Text diese in Klammern ausschreiben und kurz erklären]

# 1 Introduction & Problem Statement

The success of companies greatly depends on understanding the informal web of contacts because these hidden connections drive how work gets done. Already 2005 Kate Ehrlich and Inga Carboni state “A social network analysis examines the structure of social relationships in a group to uncover the informal connections between people” and therefore is an effective method for revealing the hidden connections between people. (Ehrlich & Carboni, 2005)

## **1.1 Scientific Research Statement**

This thesis applies social network analysis on an empirical dataset to unveil the hidden patterns of human interaction. We conduct basic SNA measures, community detection with the Louvain algorithm and the relatively recent approach of in-block nestedness from Solé-Ribalta, Tessone, Mariani, & Borge-Holthoefer (2018). Beside answering the following research questions, theoretical models will be applied on a real-world dataset, proved and eventually extended (not the primary goal of the master thesis).

1. How does network analysis help to map and understand the flow of communication in a project in a real business environment?
2. How does network analysis help to understand the connection between individuals working in a big business project?
3. With which insights about how information in an organization spread – generated with network analysis out of the data from the project ONE smart solution – can the communication within the core project team be improved? Whereas improved means: The collaborator know where they can get the relevant information (for themselves) about the subprojects of ONE smart solution.

## 1.2 Research context and relevance

How the thesis distinguishes itself from existing literature and how it contributes to bring the research state a step ahead?

### 1.2.1 Scientific context

Network analysis is a relatively new research field (Newman, Barabási, & Watts, The structure and dynamics of networks, 2006, p. 4), with an increasing relevance since the beginning of the 21st century and especially in recent years (National Research Council (U.S.) 2005, p. 11). This boom won’t be over soon (Contreras Arias 2017, p. 39). Not only is the internet a giant network and as well contains them (social media networks), but networks are everywhere (Barabási and Pósfai 2014). There is an increasing number of theoretical studies in the field of network science, which mostly address scientific problems in biology (genetic networks), artificial intelligence (neural networks), geography, transportation networks and sociology (diseases spreading) (Otte and Rousseau 2002, p. 441). The work addressing social networks is mostly focusing either on co-authorship networks in the scientific world such as the work of Cheong & Corbitt (2009) or on social media network analysis (Himelboim 2017). An interesting real-world application of SNA are emerging patterns of collaboration, because a better quality of communication within organizations is linked to higher levels of performance and service (Hargie and Tourish 2009). Nevertheless, social networks included in email logs thought to date have been little explored. Many studies testing the approach of social network science in the organizational context analyse the Enron dataset (Tran and Khaw 2006). This thesis applies existing methods to a new real-world dataset and providing further evidence for universal communication patterns.

### 1.2.2 Business relevance

The master thesis shows how information in business flows and maps the actual communication using a network analysis approach. The question “Who communicates with who” occupies every project manager (Welch and Jackson 2007) and I want to answer it for M-Industry’s project ONE smart solution. An analysis like this was never made inside the “Migros universe” before. Formal relations represented in organizational charts are important and structure certain kinds of communications, but informal relationships and employee interaction defines how work is accomplished (Borgatti and Molina 2002). In the past, companies interested in tracking the flow of information might conducted a survey and interviewed employees. These methods can be biased and people only reveal the information they want to share (King and Horrocks 2010). Email logs provide reliable data, systematically collected by the system, regardless of the study conducted, so the behaviour of the examinee is unbiased. Every mail is stored in internal servers of the IT department of Migros (MITS). It took hard work to get access to this treasure of information, unbiased and real. The master thesis helps to tackle a problem almost every business project has – track how information flows and find the best communication path to spread information – in a new way.

## 1.3 Terminology / structure of the thesis

According to Newman (2001) a social network is *“a collection of people, each of whom is acquainted with some subset of the others. Such a network can be represented as a set of points denoting people, joined in pairs by lines denoting acquaintance”*.

Social networks can represent any community no matter what size it has. Such communities can be a company or a firm, a school or even the entire world. (Newman, 2001)  
I use graph related representation is used to systematically describe the relations between people (points). Graphs were existent long before social network analysis became a thing and originate from mathematics and computer science. Graphs can represent all kind of networks, so that SNA directly benefits from the theoretical work in this two research fields. Figure 2 shows the interacting elements of a social network (individuals) as points. The fundamental units of which graphs are formed, are called nodes, vertices or actors. The second fundamental unit are the connections between objects, called edges. The connections, usually lines between nodes, can represent friendship, mail communication or other types of social interactions. Vertices connected by an edge are called neighbouring or adjacent. (Mariani 2017, pp. 9, 10) Edges can be directed (e.g. in online social networks you can follow a celebrity, which doesn’t necessarily mean that the celebrity also follows you) or undirected (friendship between two actors normally is reciprocal). The interaction intensity between two vertices is called weight (e.g. number of mails sent between two people) (Kunc n.d., p. 220). Graphs appear in several varieties, defined by the type of relationships they represent. In this text the words ‘network’ and ‘graph’ are synonymous even if network contains additional information about the vertices compared to a graph. Such additional information as age, sex, etc. can be display in a graph as well (e.g. colouring the nodes according to the age).

The rest of the thesis is structured as followed: An introduction to the theory of network science, the tools to model social interactions and the scientific research state mark the beginning of the main part. After maintaining a basic scientific understanding, the practical environment is outlined. Further parts describe the analysis conducted, starting with the hypothesises and the underlying data. The method chapter describes the data and the operationalization of the research question. These parts are brought together in the results section, where the research questions are answered. The conclusion connects the theory and the conducted analysis. In the last part we bridge them to the real world by elaborating recommendations for ONE smart solution.

# 2 Theory

Facebook, Twitter, LinkedIn, Cisco, Apple and Google base their technology on networks. The business models either make networks visible (Facebook, Twitter, LinkedIn) or use them to facilitate fast and simple access to desired information (Google, Cisco).

Attention towards networks and in particular social network analysis experienced a remarkable growth in recent years (Garton, Haythornthwaite, and Wellman 2006). Networks are applied in many different research disciplines such as computer science (World Wide Web) , artificial intelligence (neural networks), geography and transportation networks, economics and disease transmission (Barabási, Newman, Watts, Barabasi, & Watts, 2006). SNA answers social research questions as for example co-authorship networks, collaboration structures and other forms of social interactions (Otte & Rousseau, 2002) based on precise formal mathematical definitions (Kirchhoff 2010).

Social network analysis (SNA) focuses on the structure of ties within a set of social actors. The guiding assumption of SNA is, that the way of communication used by the members of a group affects some important characteristics of that group, e.g. efficiency when performing a task, moral satisfaction, leadership (Knoke, 2008; Scott, 2012).

## 2.1 Network Science

Network science has its origin in mathematical graph theory, which dates back to the 18th century. 1736 Leonard Euler published the first paper in history of graph theory – “Geometry is unimportant, only degree matters”. Euler tried to solve the puzzle of the bridges of Konigsberg (today Kaliningrad). Figure 1 shows the city of Konigsberg, consisting of four landmasses and seven bridges. He tried to answer the question “is there a trail that transverses each bridge exactly once?”, by using graph theory. Landmasses were represented as nodes and the bridges as edges. (Mariani, 2017)

### 2.1.1 Characteristics of Network Science

The study object – a network – but also the methodology defines the approach of network science. The following four aspects define the nature of network science methodology. In combination it offers multi-faceted tools and perspectives necessary to understand the properties of real networks. (Barabási and Pósfai 2014)

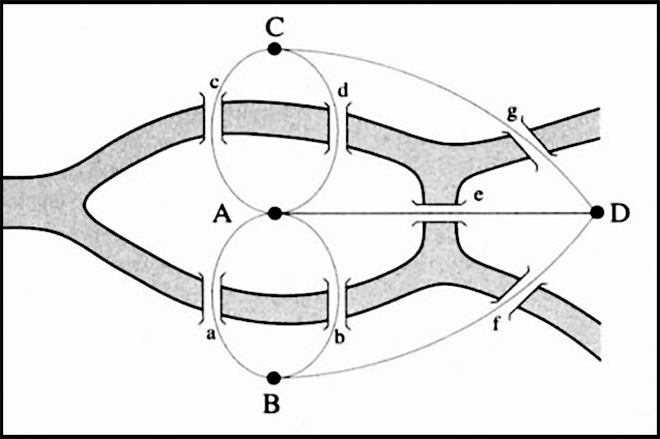
**Source: Carroll, 2014**

Figure 1 Bridges of Konigsberg. Four landmasses connected by seven bridges.

**Interdisciplinary**

Network science is not limited to one discipline, it rather offers a language allowing different disciplines to interact. The common nature of many issues cell biologists, brain scientists and computer scientists deal with, has led to a cross disciplinary fertilization of tools and ideals. Each trying to characterize the wiring diagram behind their system, too extract information of a noisy and incomplete dataset or trying to understand the robustness to failures.

**Empirical, Data Driven**

Graph theory, a field of mathematics, is the root of many key concepts in network science. Unlike graph theory, network science is empirical and is never satisfied with abstract mathematical tools to describe network property. It focuses on testing the tools on real data. Insights about the network’s properties and behaviour determine utility.

**Quantitative and Mathematical**

One of the key concepts borrowed from graph theory is the formalism to deal with graphs and to properly use its tools. The conceptual framework to seek universal organizing principles and the way to deal with randomness is derived from statistical physics.  
Theoretical formalism must be understood to efficiently master the tools, even if network analysis software has opened their use up to a broad community.

**Computational**

Many or most practical networks are big in size: Social networks include every human being and the neural networks consists of billions of neurons. The shear amount of auxiliary data regularly confronts network scientists to computational challenges. Algorithms, database management and data mining serve to tackle the strong computational character.

### 2.1.2 Networks at the Heart of Complex Systems

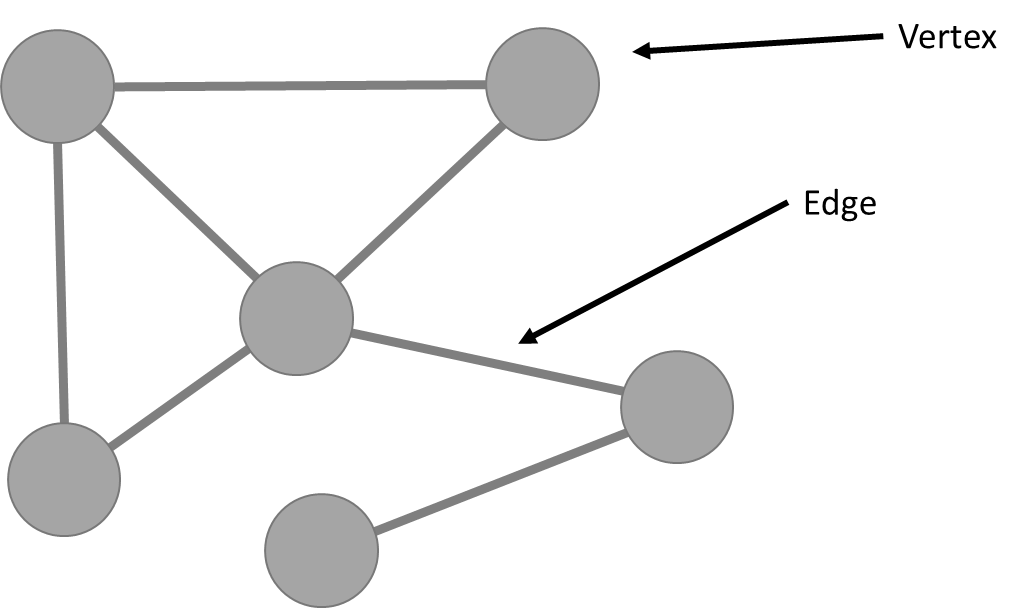
In the 21st century the world around us consists of complex systems. Global society is the cooperation between billions of individuals, communication infrastructure is integrated from cell phones and computers, which are interconnected and our brain requires the coherent activity of billions of neurons to let us comprehend our world. (Barabási and Pósfai 2014) These and many other cotemporally systems are called complex systems, whereas complex describes something that consists of many interconnected parts and is characterized by a very complicated or involved arrangements of parts (Dictionary.com, 22/11/2018). The emergence of network science is an approach to develop a deep understanding of complex systems, mathematically describe, predict and eventually control them.   
Behind all complex systems is a network in which the interactions between the components are hidden. The underlaying networks are driven by a common set of fundamental laws and organizing principles, disregarding the diversity of complex systems. Once we abstract away the exact nature of interactions and components, networks, no matter the size, nature, age or scope, are more similar than different. (Barabási and Pósfai 2014)

### 2.1.3 Two Forces Helped the Emergence of Network Science

Networks themselves are not new. Metabolic networks have a history of billions of years dating back to the origins of life, social networks are as old as humanity and mathematicians exploring graphs since 1736.

A detailed network map is necessary to describe the behaviour of a complex system. In the case of a social system this requires a list of your friends, your friend’s friends, and so on. Before the internet revolution we lacked the availability of data and the tools to map and track the huge amount of data behind complex systems. The technological advances allowed to create diverse maps from complex systems (e.g. Facebook and LinkedIn developed accurate depositories of friendships or professional ties). These diverse network maps allowed network scientists to overcome the differences of various networks we encounter in nature and science (e.g. nodes can be molecules in a metabolic network or individuals in social systems). A key discovery is the universal properties behind the different systems. The architecture of networks, disregarding the domain they’re emerging of, is similar to each other. This universality offers the foundation of the new discipline of network science. (Barabási 2014, pp. 2–3)

Figure 2 Visualization of an example social network.

Source: Own illustration

## 2.2 Tools

A graph is a mathematical object. Formally, Graph theory defines a network as the mathematical notation , where V is the vertex set (actors) and E is the edge set (relations or interactions) of the community. Networks are representations of real-world (complex) systems, where the context adds a specific meaning to nodes and links. Graph techniques can be used to analyse real-world networks, whereas all networks are graphs but not the other way around. (Mariani 2017, p. 13)

Social networks describe a group of people, in which each individual is acquainted with a subset of the others. The wiring structure of a network (who is connected to who) has important implications on the behaviour. For example, a variation in the average number of acquaintances that individuals have, the average degree of the network, influences the spread of information. By abstracting away details, graph theory helps social scientists to understand empirical data and describe their structure. A social network is represented as a set of points connected by lines. The points represent people which are connected by a line if they interact with each other (Newman, Barabási, & Watts, The structure and dynamics of networks, 2006, S. 221ff). The number of nodes is called the size of the network (N). Number of links (L) represents the total number of interactions between all nodes. High-degree vertices are often called hubs. In social network science we use the terms network, node, link which are interchangeably with the terms graph, vertex and edge from graph theory. (Barabási & Pósfai, 2016) In social network analysis (SNA) seeks to predict relationships shaping the interaction between entities with quantitative analysis (Butts 2008, p. 1). The specific nature of the entities depend on the object of study and can be persons (our case), groups or organizations (Butts 2008, p. 2). SNA has four underlaying key assumptions, defined by Freeman (2004): i) motivated by “structural intuition” based on the linking ties between actors; ii) based on empirical data which is systematically collected; iii) based on mathematics and/or computational models; iv) largely relies on graphics. SNA quantifies social structures, therefor it is also referred to as structural analysis (Wellman and Berkowitz 1988). Any social process or system that can be conceptualized as a set of units and a set of lines connecting pairs of units can be studied as a social network. It can be constructed for any kind of social community and are studied because of the increasing interest in patterns of human interactions and the implications for the spread of information. (Newman 2001) Examples of social structures that have been studied as networks are friendship among children in a school, family relations among members of a social elite, shared board members of corporations, trade relations between countries, and hyperlinks between websites (De Nooy 2010) The tools of social network analysis help to describe the roles of individuals and small groups within the network. SNA measures are divided into individual and group-level measures (Sauer and Kauffeld 2013) whereas the main individual / node-level measure is centrality. Centrality measures identify the most important vertices within a graph. Several centrality measures are distinguished: Degree centrality (number of acquaintances), closeness centrality (sum of its distances to all other nodes) and betweenness centrality (number of times a node lays along the shortest path between two other nodes) (Barabási & Pósfai, 2016). Subgroup measures show how a network can be partitioned into communities (Hawe, Shiell, and Riley 2004). Figure 3 explains the properties of communities, smaller but more densely connected groups within a network, have important influence on individual behaviour (Jenson 2007). Further terms used in the thesis are stated in Appendix A.

# 

# 3 Research State & gaps

This chapter is two folded. Part one describes the current research state focusing on email network analysis. The second, more exhaustive part is about connectivity patterns in networks.

## 3.1 Email network analysis

Many studies applying SNA measures to email networks study the Enron dataset **(Zehnalova and Horak 2015, p. 1). The Enron data-log became public during the investigations of SEC (Securities and Exchange Commission) and FERC (Federal Energy Regulatory Commission) for transparency, historical and academic research purposes. Diesner, Frantz, & Carley (2005) analysed the structural properties of a directed graph with weighted edges and identified key players in the crisis. Rowe, Creamer, Hershkop, & Stolfo (2007) reconstructed the social hierarchy based on a social score measure in an undirected graph while Gloor (2005) extracts evolution movies with temporal links and content analysis. Others studied the undirected network of Enron employees using network analytic measures (Chapanond, Krishnamoorthy, and Yener 2005) or considered networks constructed by changing the minimum number of mails exchanged that an edge exists (Shetty and Rey 2004) This thesis conducts a network analysis not on the Enron email corpus but on a new email dataset and overcomes the lack of large public email corpuses.**

## ****3.2 Connectivity patterns****

As mentioned in the part 2.1 Network Science, the discovery of recurring connectivity patterns stands at the very origin of network science. Ascertaining them is central to understand the underlaying microscopic (node-level) mechanisms of networks. Two macro-structural patterns emerge most prominent, nestedness and modularity. (Solé-Ribalta et al. 2018)

**Community detection with modularity**

Community detection is a clustering problem with the goal to assign each node in the dataset to a community in a consistent and meaningful way (Shah and Zaman n.d.). Whereas meaningful means, it helps to understand the daedal organization of complex networks (Fortunato 2010; Lancichinetti, Fortunato, and Radicchi 2008). Community detection is one of the most well-studied problems in the field of (social) network analysis. Modularity measures the strength of the division of a network into communities as a scalar value between -1 and 1. High modularity implies dense internal connectivity within a community, but sparse connectivity between nodes belonging to different communities. (Barabási et al. 2006). Modular patterns and their identification are a characteristic predominantly in SNA but used in other fields as well (Coleman 1964; Fortunato 2010).  
Exact modularity optimization is computationally intractable so many approximation algorithms have been developed to reasonably identify communities in the structure of networks (Barabási et al. 2006; Blondel et al. 2008; Newman and Girvan 2003). Without the intention of being exhaustive, the most widely used are: *Infomap*, which uses random walks; *Fastgreedy*, a greedy algorithm that optimises modularity score by pairing nodes for which it calculated the maximum improvement of modularity; *Edge betweenness* is getting communities by removing edges with high betweenness score; *walktrap*, a hierarchical clustering algorithm assuming short distance random walks to stay within the same community; *label-propagation* assigns nodes to the same community as the majority of its neighbours; *spinglass* is based on the Potts model to find edges which should connect nodes of the same community; *leading* *eigenvector*, optimizes the modularity by using eigenvalues and eigenvectors of the modularity matrix and *Louvain* (multilevel), also a greedy algorithm, assigns a different community to each node and then moving nodes to one of its neighbours with which it achieves the highest improvement in modularity. (Yang, Algesheimer, & Tessone, 2016, pp. 14f) Recent studies paying attention to the differences of community detection algorithms are the paper about community detection methods in social network analysis by Ahmad, Baharom, & Sapaat (2018) and the comparison between community detection algorithms on artificial networks by Yang, Algesheimer, & Tessone (2016)

**Nestedness**

Nestedness, first observed in biological networks, characterizes the geographical distribution of fauna in isolated, yet related landscapes (Patterson and Atmar 1986). A complex system shows nestedness if the interactions of any node is a subset of the interactions from a node with a larger degree (Solé-Ribalta et al. 2018). Beyond biological systems nested properties occur as well in several other, unipartite networks as inter-country trade relations (König, Tessone, and Zenou 2014), inter-bank loans (Soramäki et al. 2007) and in the management field within inter-firm knowledge growth networks or disparate industry sectors (Tomasello et al. 2016). BINMATNEST (Binary matrix nestedness temperature calculator) from Rodriguez-Girones & Santamaria (2006), a further development on NTC (nestedness temperature calculator) from Atmar & Patterson (1993), the nestedness metric based on overlap and decreasing filling (NODF) (Almeida-Neto et al. 2008) and FCM (fitness-complexity metric) from (Tacchella et al. 2012) are among the most widely used quantifying methods for nestedness. The omnipresence of nestedness, discovery in fields despite ecology and its implications for stability in them, implicated research about the relation to other common network structures (Jonhson, Domínguez-García, and Muñoz 2013; Lee 2016; Solé-Ribalta et al. 2018). Recent studies in this field also potter at new methods detecting nestedness in real-world networks, without the underlaying hypothesis that all nodes belong to a single nested component (Grimm and Tessone 2017).

**In-block nestedness**

An external file that holds a picture, illustration, etc.
Object name is zpq0220623880001.jpgA new debate among scholars started: Can modularity and nestedness co-exist? The two patterns appear incompatible, since the emerge from contractionary dynamics (cooperation and competition). A paper recently published by Solé-Ribalta et al. (2018) presents the concept of in-block nestedness. The concept and the corresponding method determine to what extent a network is composed of blocks (communities) whose internal connectivity exhibits nestedness. Different to other work they jointly consider both patterns and neither measures nestedness and modularity independently (Borge-Holthoefer et al. 2017; Fortuna et al. 2010; Olesen et al. 2007) nor operates sequentially (Flores, Valverde, and Weitz 2013; Lewinsohn et al. 2006). My goal is to test the recent approach of in-block nestedness on a new and rather large dataset, emerging from collaboration via email. I’m trying to give further evidence of this new model in real-world business networks.

Figure 3 Vertices in three communities. Within communities (shaded) there are many edges, between vertices of different groups only a few edges are present.

Source: Newman, 2006

# 4 M-Industry & Project ONE

The following paragraph outlines the organization from which the dataset is derived from. This helps to get the right setting for the analysis and the conclusion from the data.

## 4.1 Migros Group

Migros is the largest retailer and the largest private employer in Switzerland. The group comprises over 60 companies. Ten regional Cooperatives, each with an own management, are responsible for over one thousand stores. FMC (Federation of Migros Cooperatives), the central service provider of Migros represents the group externally and is owned by the cooperatives. The operating business of Migros is divided into six departments. Beside the industry which together with wholesale forms the department four; HR, communications, culture and leisure time are summed up in department one. Marketing is department two, logistics and informatics department three, finance department five and retailing forms department six. All six departments are part of the general direction, which together with the assembly of delegates, the administration and statutory auditors outline the legal bodies of FMC.

## 4.2 Migros Industry

As described in the previous section 4.1 Migros Group M-Industry is one out of six departments in the Migros Group. The M-Industry consists of 25 companies in Switzerland and eight companies abroad, mainly producing goods sold in the Migros supermarkets, but as well in other stores (Denner, Volg, etc.) and overseas (United Kingdom, Germany, China, United States, France, etc.). Migros Industry sells over twenty thousand different food and non-food products, has more than fourteen thousand employees and total net sales of 6.5 billion. (Migros-Genossenschafts-Bund 2017, 2018a; b)   
Within the M-Industry, more exactly among the 25 companies in Switzerland, which are all independent but affiliates of FME (Federation of Migros Cooperatives), efforts towards increase the competitive ability are rising to grasp the opportunities in foreign markets. The latest and biggest action to increase efficiency and effectiveness is called ONE smart solution.

## 4.3 ONE

ONE smart solution is a change project launched in September 2016 with the goal to roll out a uniform ERP system (SAP Hana) and restructure the underlying processes towards centralization to exploit better the market power. The core team, organized in thirteen subprojects, consists of 150 employees. Beside the core team almost 400 employees are part of the so-called site groups. Site group members are working part time in the project organization, beside remaining part of the original organisation in their “home” company. (Migros 2018)   
Due to the scope and impact of ONE change management is a core task of the project organisation, beside the work on the processes and the modelling of them in the new ERP system. The PMO (Project Management Office) includes a dedicated person for communication, supporting activities to increase change readiness within M-Industry. Within change management communication, inside the project team but as well as to the outside, plays a crucial role. This thesis is about the communication within the core team and the site groups members. Among the communication only the interactions constituted via mail are part of the analysis. To estimate how much of the total communication, we catch by analysing Microsoft outlook data, a survey among all project members is conducted.

# 5 Hypothesises

Based on the experience as an internal consultant in Migros Industry we have several hypothesises on what the result of the network analysis may look like.

The communication within a subproject, a group of people dedicated to work on a particular subfield of the project ONE, is expected to be dense. Typically, there is dense communication within a subproject, since members of a subproject are closely co-working on the same topic and originally come from similar business units in different firms and therefore speak a common language.

**Hypothesis 1:** A lot of communication within a subproject team.

Co-workers originating from different business units often show differences in culture. Further, if people work in different , they deal with distinct topics. Consequently, we expect, between team members of different subprojects, weak interaction.

**Hypothesis 2:** Weak communications between different subproject teams.

Assuming subproject membership defines the communities, these results would coincide with findings from community detection in network theory. As stated before, different clusters, in our case subprojects, are only joined by a few highly connected individuals, but show dense connectivity between nodes in the same community.

Weak communications between different subproject indicates a weak flow of information between two partial projects. This effect even is accentuated for site group members. Site group members are still engaged within the daily business of the company of origin. This results in spending less time for project one and little presence in the projects head office in Dietikon. We assume the time spent in the project organization of ONE is focused on the particular subproject they’re assigned to, resulting in a lack of information about other subprojects.

**Hypothesis 3:** Lack of information for site groups about other subprojects.

# 6 Concept & Method

The new dataset gathered from the business environment in Migros and the survey to estimate the part of project internal communication covered with email as well as the methods used on it, are described the following paragraph.

## 6.1 Data

The email log is the main data. To estimate the proportion of communication represented in it, I conducted a survey about communication tools used in the project organization.

### 6.1.1 Email log

If you think about communication in business context, the first thing that comes to your mind is email. Emails are systematically collected by the IT department of Migros (MITS) and stored in internal servers, regardless of this thesis. This provides a way to extract relevant information without hurting personal rights and privacy. (Nitz, 2013, p. 50)

The dataset contains emails sent or received between ONE project members. The member list contains all names involved in the project organization (reference date: May 31st 2018). The information base comprises every single mail sent or received between core team and site group members during the time period from November 2016 until the end of June 2018. The dataset includes 595 people and over two million mails.   
Each individual has the attributes subproject, role and company. The whole project team comprises 18 different subprojects and ten different roles– independent of each other. The 595 collaborators (originally) belong to 24 different companies.

The concerns regarding privacy are understandable and came up from ONE-managers, the IT and the legal department of Migros. In this work I address the question in two different ways. The dataset is fully anonymized by a third person so that only Migros possesses the key to verify who is the respective person behind the anonymous identity number. Nevertheless, the anonymization includes every critical aspect of an individuum such as subproject membership, company of origin and role. This assures the results still can be aggregated on the three attributes. Second, we don’t mine any content but only header information, which includes the sender, the receiver and time-stamp. The subject is used to distinguish between private mails, which are excluded from the analysis, and business mails.

**Restrictions:** The collection contains only the communication within the project organization (core team and site group members). The organizational structure of Migros, with a variety of independent legal bodies, requires approval from all involved companies to use to communication towards the outside as well. The approval process is time consuming and would exceed the time limit of this master thesis. Moreover, steering committee members and the accompanying mails send and received, are deleted from the dataset.

### 6.1.2 Questionnaire

Interpersonal communication is not only constituted via email. To estimate the proportion captured by analysing the email connections, I conduct a survey among the affected persons. The questionnaire (Appendix A) sent to all 595 project members tries to answer the question: “How much of the information flow happens via e-mail?”   
Even if the analysis of mail connections won’t represent the full truth, together with the questionnaire a good extrapolation becomes possible and we can see how much off the communication is included in the social network analysis.

## 6.2 Operationalization of the research question

To answer the question “How does network analysis help to map and understand the flow of communication in a project in real business environment?” node level measures – to understand the role of individuals – are used. Beside the average path length, the diameter and the degree distribution, I analyse transitivity, the probability that an adjacent of a friend is my adjacent as well. Density gives back the portion of existing connections compared to the number of connections possible. Different centrality measures such as degree centrality, betweenness centrality, closeness centrality are used to identify the most influential individuals and show if centralization is present: A graph is central if it is tightly organized around a central actor. Beside the above results I plot the whole network to get a visual impression of communication flow.

The hypothesises one to three are about community detection. Modularity measures exactly what the hypothesises are about. The measure is high if there are dense connections between nodes within a community (subproject) but spare connections between different modules. Based on the paper “A comparative analysis of community detection on artificial networks” from Yang et al. (2016) and because the in-block nestedness model from Solé-Ribalta et al. (2018) uses *Louvain*-method from Blondel et al. (2008), I use the *Louvain*- or also *called multilevel*-Method for community detection. Beside this modularity maximizing approach I also use *walktrap* and *label-propagation* to check and ensure Louvain has the highest modularity value. By comparing the Louvain-communities based on the empirical communication data with the administrative allocation of people into subprojects, the hypothesises can be tested. If the modularity maximized communities are congruent with the subprojects, hypothesis one and two are confirmed. If communities don’t match well with the *subproject.type* attribute of a node, I use the same method for the attributes *sp1.type* and *company.type* , which may map better.   
For hypothesis one and two I only use the core team members. Each core team node is assigned to a subproject (sp1.type in the node table). Hypothesis three is about the site-group members which all have sp1.type zero, because they are not officially assigned to a subproject by the project management office (PMO). To answer hypothesis three, I first have to assign the SG-nodes to a subproject. I do this in the most intuitive way by checking for every node into which subproject it has the most existing edges and then assign it to the corresponding *sp1.type* itself. After that, the same procedure as for hypothesis one and two is used to check weather site group member lack information about other subprojects. Hypothesis three predicts a high modularity value for site group members and congruent communities compared to the assigned subprojects.

The question about the understanding of connections between individuals is the question of the predominant connectivity pattern. The two predominant connectivity patterns, nestedness and modularity, can occur interlinked (Solé-Ribalta et al. 2018). To test if interpersonal connections in the ONE dataset follow the IBN patterns I use the method from the paper of Solé-Ribalta et al. (2018). As they state, this finding would challenge the understanding of the topology of social systems, calling for new models to explain the emergence of in-block nestedness.

To generate insights to improve information spreading I compare the SNA results with real events throughout the different project phases, trying to find meaningful commonalties. Comparing the matching patterns with the business experiences can lead to indications how improvements on communication can be done.

## 6.3 Execution

Centrality and most node level measures can be calculated for directed and weighted networks. The Louvain community detection algorithm demands networks to be undirected and the in-block nestedness algorithm from Solé-Ribalta et al. (2018) only runs on undirected and unweighted networks. To get comparable results along all analyses, all calculations are conducted on undirected and unweighted networks. Further I use undirected edges because the business’ focus of interest lies on who interacts with who, but not in the distinction between sending and receiving mails. The correlation between in- and out-degree is 0.9 and reciprocity of communication in the project ONE dataset is 0.76.   
The calculations of assortativity marks an exception and is calculated with the directed graph as well, due to different interpretation for undirected and directed versions of the network. Same for degree calculations of the nodes, which is conducted with the undirected weighted and the undirected unweighted graph because the directed graph represents mails and the undirected represents connections among people.

The analysis is split into three parts and except the *DynSnap* partition (Python) and the IBN calculation (Ubuntu) I used the program *RStudio* along with the *igraph* package for social network analysis.

### 6.3.1 Static network

To get an overview of the whole communication, time and date are ignored in this part. The network contains the whole dataset and shows who sends mails to who, respectively who interacts with who.

First, I plot the whole undirected and unweighted network as a graph and count the vertices and edges of which it consists. The basic *igraph* commands are used on the undirected and unweighted network to calculate basic measures of a networks: Transitivity, the average path length and the diameter, including the edges on it. The components function shows the number of strong and weak connected components together with the corresponding sizes. To catch the difference between strong and weak connected components the directed network is used, otherwise every existing connection would be strong. Assortativity for the weighted graph gives back if people sending or receiving a lot of mails mostly interact with people also communicating a lot via email. The assortativity of the unweighted graph tests if high degree nodes interact mostly with other high degree nodes, whereas high degree means interaction with many different people, not with many mails. Because of the different meaning of assortativity of simplified or not simplified networks, both are calculated.   
Degree, betweenness and closeness centralization indicate whether the graph on global level is tightly organized around its most central point or not. Different centrality measures provide insights for the most central – sometimes called most important – individuals. The histogram for each of them shows the distribution. A comparison with normalized mutual information measure indicates, if the most central node varies depending on the centrality used. I also compare the five most central vertices of each centrality measure manually.   
The centrality calculations on the interaction network is conducted with the unweighted network. Nevertheless, it’s interesting to know how many mails are sent and who sends the most, therefor I recalculate degree centrality with the un-simplified network as well.

The *igraph* package includes commands for multi-level community detection, community detection based on propagating labels and via short random walks. Modularity helps to rate for which community detection algorithm the partition is best.   
The communities from the method with the highest modularity, I then use to compare against the administrative clusters (subprojects, role, company of origin). We compare the empirical communities not only against the subproject membership but also against company membership and the role of the individuum in the project ONE. For hypothesis one and two only core team nodes (*r1.type* ≠ 1) are considered.   
Site group members have no subproject assigned to them. To test of hypothesis 3, they need to be assigned manually to a subproject. I do this by checking for every site group node into which subproject it has the most connections. Then the site group member gets assigned it to this subproject. The empirical communities based on communication are then compared to the assigned subproject, using the same method as for the core team members.

### 6.3.2 Dynamic network

The collaboration network is dynamic and we have the time stamp of every mail. To represent the timely evolution, dynamic networks normally are represented as a sequence of snapshots. Darst, Granell, Arenas, Gómez, & Saramäki (2016) elaborated an approach, which does not use constant but dynamic intervals. They propose a method able to detect evolutionary changes in the configuration. The code looks at similarity peaks between the event sets and it is publicly available (Richard K. Darst et al. 2016). I use their code to split the dataset into different snapshots to examine the network evolution in time. No input parameters or a priori assumptions that must be made. The input is simple, consisting of numeric event-ID (*From* / *To* columns) and time. The method finds an initial intrinsic scale to the data, where each interval represents roughly the same amount of change.

After defining the intervals, the analysis of the system turns into an investigation of the slices. All analyses described in *6.3.1 Static network* are conducted for each snapshot of the network*.* A time series for each measure shows the evolution over time and makes a comparison possible.

### 6.3.3 In-block nestedness

The last part of the analysis concerns the overall connectivity pattern in the collaboration network. I test for in-block nestedness with the model from Solé-Ribalta et al. (2018). The code for IBN detection runs on Ubuntu and uses the Louvain algorithm for community detection. As the network is dynamic, I use the *DynSnap* snapshots to analyse the variation over time of in-block nestedness. For each of the eleven networks of different points of time, I run the calculation fifty times and create the average out of it. So, we can see the trend on a time series plot.   
Even if the algorithm already subtracts the IBN that can be accounted to randomness, I do the same analysis for two random networks with the same properties as the ONE-network. The Erdös-Renyi network created with e*rdos.renyi.game* command in igraph, using the same amount of vertices and edges as in the original graph. The edges are chosen uniformly randomly from the set of all possible edges. The second random graph is one with a given degree sequence. The configuration model connects the out-stubs of the edges together. The IBN for the two reference models should be zero or near to zero since, as mentioned above, the algorithm accounts for the random part of IBN.   
To verify if the time series for the ONE-network is significantly different to the Erdös-Renyi-Networks and configuration model, I perform a paired t-test on a 99% confidence level.

# 7 Results

Subsequently the obtained results are unveiled. This section brings together the theoretical methodology and the real-world dataset.

## 7.1 Description of population & sample

The first part gives an overview about the data sample including the survey results to estimate the share of communication covered with the social network analysis used to create the base for understanding the further results.

### 7.1.1 Questionnaire

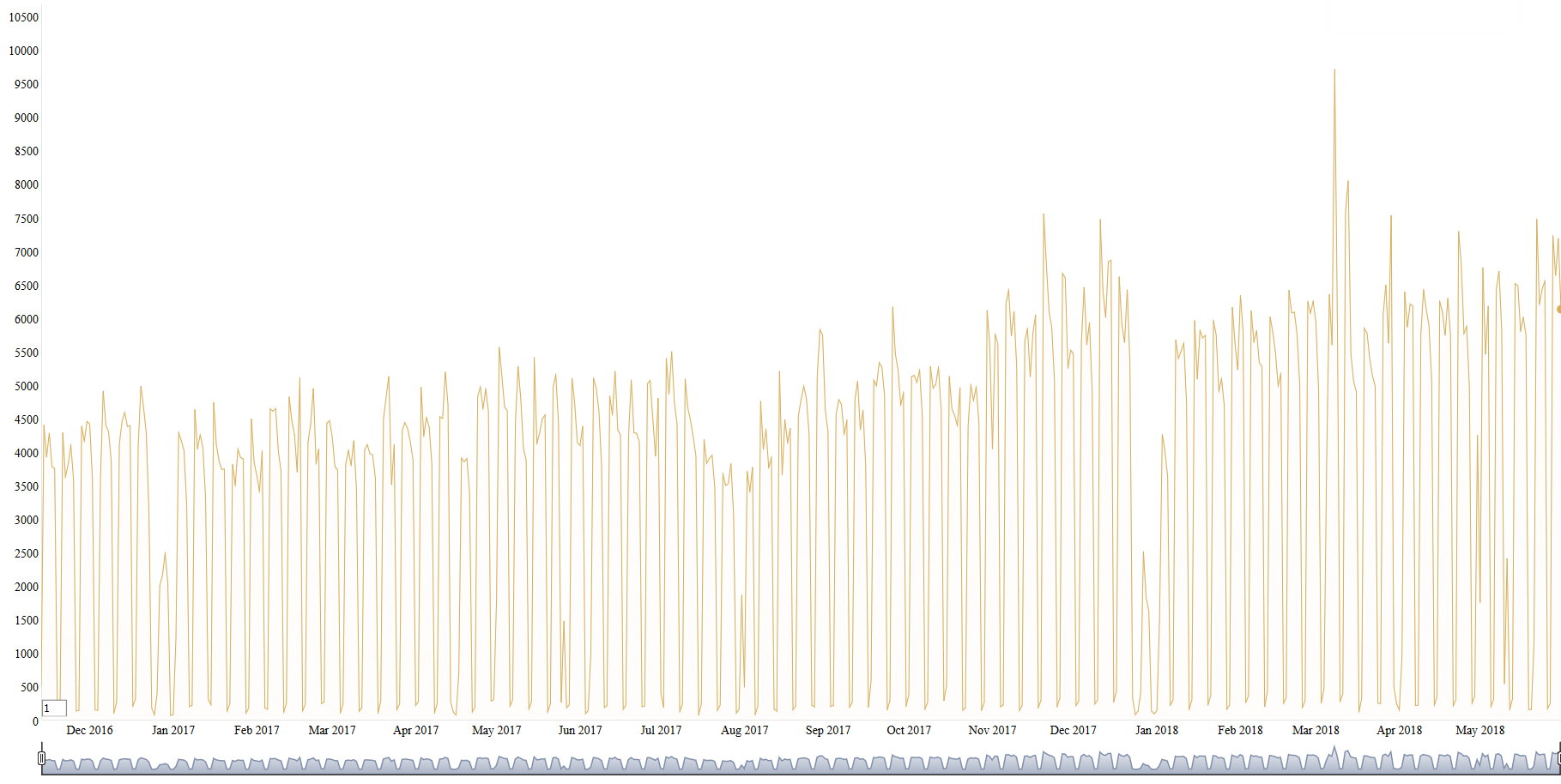
Completion rate of the survey is 30.6%. 216 out of a total number of visitors of 705 completed the whole survey. 478 visitors did not start the questionnaire. Out of the 227 visitor who begun the questionnaire only 9 didn’t finish. Appendix A summarizes the results from the 216 completed answers.

Question seven displays email is the most used communication tool (7.1/10), followed by personal interaction (5.7/10), phone (5.4/10), Skype calls and messages (4.9/10) and WhatsApp (3.4/10). Slack, Jira and Confluence are less used, even they are promoted by the project’s management office of ONE.   
Question eight to eleven of the questionnaire distinguish between communication within a subproject and the communication between different subprojects, as we also do in network analysis. To communicate within a subproject 38.6% use mails on a daily basis, while only 30.7% of the participants declare to have face-to-face conversations every day. Phone and Skype are used daily by 18.2% respectively 15.4% of the population. The other means of communication only play a submerged role in the daily interaction between members of the same subproject. By excluding direct face-to-face interactions, the results become even more clear: 60.4% of the tool-based communication is email traffic, 21% phone calls, 11.1% via Skype and 7.5% via other tools.   
The communication towards members of other subprojects shows similar patterns. Email is the tool. 33.4% use mails, 21.8% face-to-face interaction and 17.7% phone calls daily to communicate outside of the own subproject. Skype is used every day by 13% and other communication tools by 0.9%. Excluding direct interactions, 65% of communication is via email, 18% via phone calls, 7.8% via Skype and 9.2% via other communication channels.   
Even though the technology exists for several decades, email still is the most important communication tool in business. Among ONE team members it is used more often than every other tool and even more often than face-to-face communication.

### 7.1.2 Social network analysis

The whole dataset includes 2’003’867 mails sent between 595 ONE project members. The period under consideration includes 595 days or 85 weeks This amounts to an average of 3’367 mails per day (including weekends) or 23’575 mails per week. The temporal distribution in Figure 4 shows the exact time curve of the number of mails per day. The recurrent local minimums are the weekends and the drops at the end respectively beginning of each year show the holiday effects during Christmas time and new year. We may also observe the effect of summer holidays in the year 2017. Mail traffic decreases between end of July and beginning of August. Global maximum is on March 8th 2018 with 9722 mails sent on a single day. This event marks the start of the pilot process in Segment 3 of M-Industry. The pilot launched among the companies Aproz, Bina and Jowa was used for a first fit gap analysis.

Creating an undirected and unweighted network from the dataset, results in a network containing 19’037 edges. An edge stands for a unique interaction between two of the 595 project members. Each edge connects two nodes, resulting on an average degree of 64 (interactions to different colleagues).



Source: Own illustration

Figure 4 Number of emails per day during the period under consideration. The many sharp drops are the weekends on which less mails are sent.

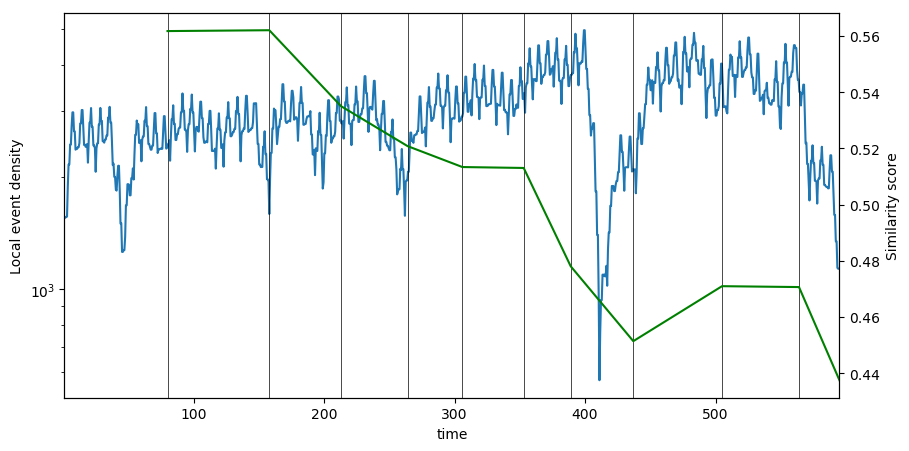
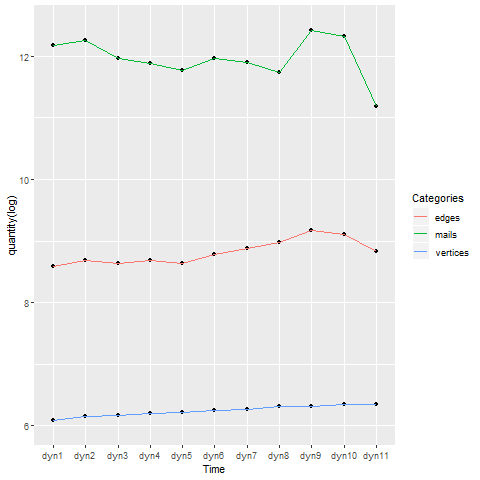
Source: Own illustration

Figure 5 Slicing the ONE dataset. Each event is an email communication within the ONE project team. We can see that internal length reacts to the rate of change of events. The similarity reflects the rate of change of events at other times.

The *DynSnap* model divides the dataset, following changes in event composition, into eleven slices (Figure 5). Time intervals of several weeks appear reasonable for changes in email communication patterns (Richard K Darst et al. 2016, p. 4). The timescales are:

Table 1 Overview DynSnap timescales. Includes start and end date, number of emails, edges and nodes.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Timespan** | **Start** | **End** | **# Emails** | **# Edges** | **# Nodes** |
| dyn1 | 13/11/2016 | 31/01/2017 | 194’308 | 5’365 | 442 |
| dyn2 | 31/01/2017 | 19/04/2017 | 211’823 | 5’868 | 468 |
| dyn3 | 19/04/2017 | 13/06/2017 | 156’345 | 5’648 | 478 |
| dyn4 | 13/06/2017 | 03/08/2017 | 145’105 | 5’872 | 494 |
| dyn5 | 03/08/2017 | 14/09/2017 | 129’574 | 5’628 | 504 |
| dyn6 | 14/09/2017 | 31/10/2017 | 157’322 | 6’530 | 515 |
| dyn7 | 31/10/2017 | 06/12/2017 | 146’142 | 7’177 | 523 |
| dyn8 | 06/12/2017 | 23/01/2018 | 125’398 | 7’959 | 548 |
| dyn9 | 23/01/2018 | 01/04/2018 | 246’908 | 9’613 | 553 |
| dyn10 | 01/04/2018 | 30/05/2018 | 224’658 | 8’987 | 574 |
| dyn11 | 30/05/2018 | 30/06/2018 | 71’585 | 6’788 | 667 |

Figure 6 shows the dynamic network evolution using the DynSnap timescales. According to Table 1, each slice comprises between 5’365 and 9’613 edges (interpersonal interactions) having its source in 71’585 to 224’658 emails during the separate time intervals.

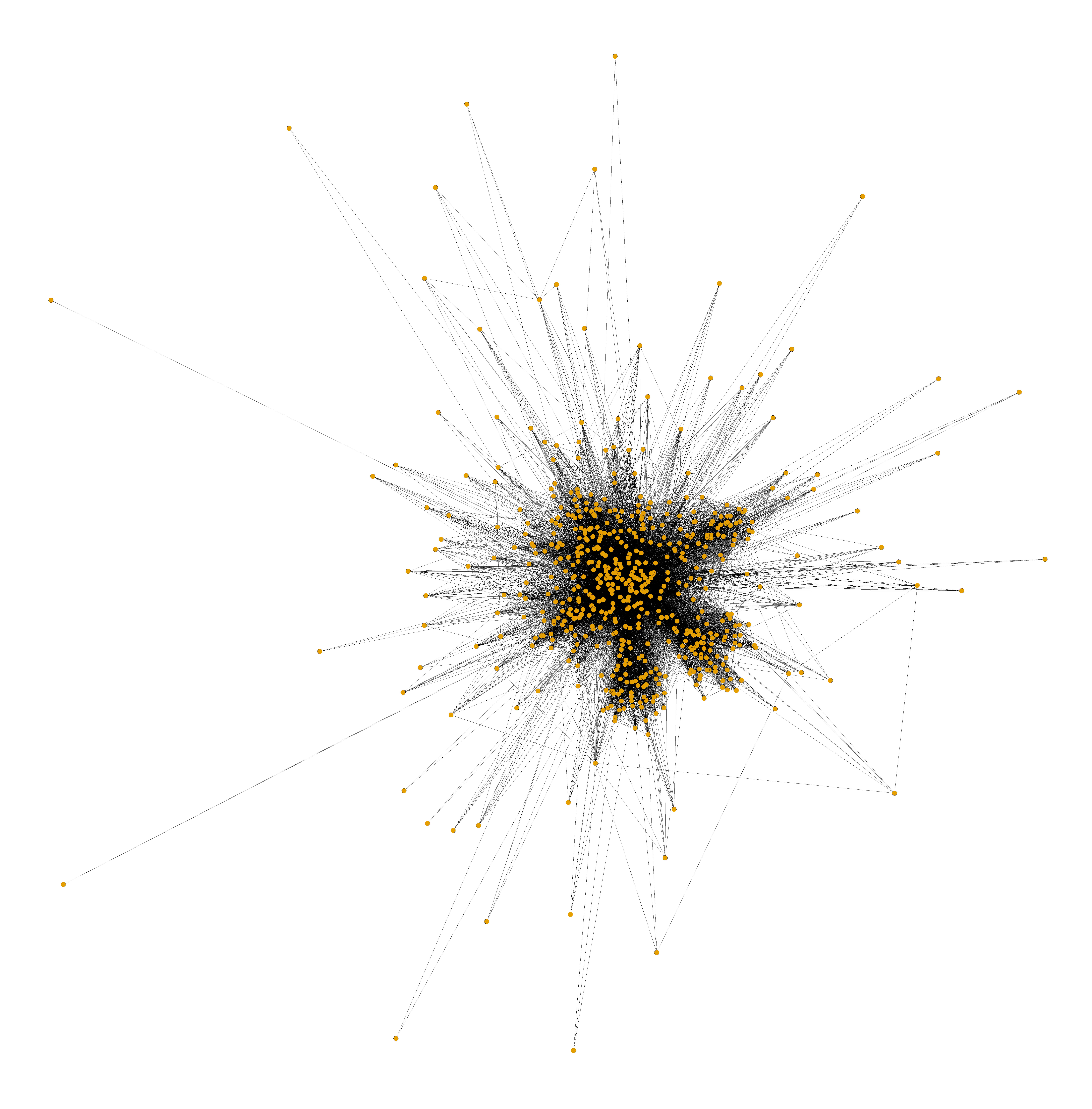
Source: Own illustration

Figure 6 Dynamic evolution of the number of mails, edges and vertices in the ONE network (along the DynSnap timescales).

## 7.2 Research questions & answers

Figure 7 shows the graph representation of the analysed network. It contains all 595 project members, connected by a line, conceding they interacted at least once via email in the time period from November 2016 until June 2018.

### 7.2.1 SNA helps to map and understand the flow of communication

Between 595 actors are 176’715 different (undirected) connections possible. The unweighted and undirected ONE network contains 19’037 edges, resulting in an overall network density of 0.1%. A more intuitive node level measure is average path length, also called mean distance. Within the project team the average path length is 1.9. On average every person can reach any other project team member (including site group members) in less than two steps. Short average path length in many real network lead to the concept of a small world, stating that every human being is connected by only six degrees of separation (Watts 1999). Among the individuals in the ONE network the mean distance is three times shorter, facilitating quick information transfer among the network. The longest path from an actor to another, called diameter, is four and measures the size of the graph from the ONE network.

Source: Own illustration

Figure 7 Simplified ONE network. Including 595 vertices and 19'037 edges. Lines represent interaction. Vertices placed using the force-directed layout algorithm by Reingold & Fruchterman (1991).

Transitivity, the probability that adjacent vertices are connected, is 0.38. The chance, that two persons I interacted with, have interacted directly with each other as well (or simpler: a friend of my friend is a friend of mine as well), is over one-third. This is significantly higher than the transitivity coefficient for a random network of the same size. Transitivity for random networks simply equals the density, 0.1% in our case. Assortativity, represents the tendency of similar vertices connect to each other. The values are positive for both, the simplified and the un-simplified network. An assortativity coefficient of 0.41 for the un-simplified networks indicate, that people who often communicate via email, tend to connect with people who also communicate a lot via email. The value for the undirected and unweighted networks is with 0.2 lower but still positive, indicating well-connected actors interacting more often with other well-connected people.

Figure 8 shows the evolution of the six measures, above described for the static network, for the dynamic timescales. Assortativity raises over time for both, the simplified and the un-simplified network. The longer the individuals are working together within the project organization, the more well-connected nodes tend to interact among each other. Density, transitivity, the diameter and the average path length seem to be quite constant. Except for the timescales dyn7 to dyn10 (October 31st 2017 to May 30th 2018), where they trend towards a more densely connected network with higher transitivity and lower path lengths than in the other time intervals.

Freemans measure of graph centralization is an expression of how tightly the graph is organized around its most central point. Figure 9 shows the variation of the three centralization measures over time. The curves for degree, betweenness and closeness centralization all show similar patterns. Constant during dyn1 to dyn6, raising from dyn7 until dyn9 and decreasing to the original level in the time intervals dyn10 and dyn11. For the time span under consideration closeness centralization is in every *DynSnap* network the highest, followed by degree centralization. Betweenness centralization invariably is on a lower level. Considering the static network with the data of the whole timespan, closeness centralization is 0.86, degree centralization 0.84 and betweenness centralization 0.24. The larger the centralization value, the more likely it is, that a single actor is quite central, with the remaining actors are considerably less central.

Source: Own illustration

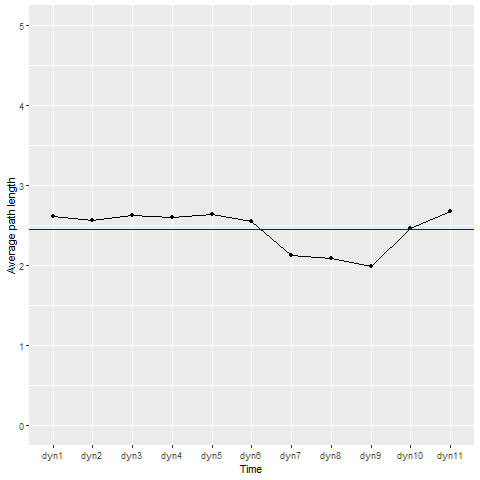
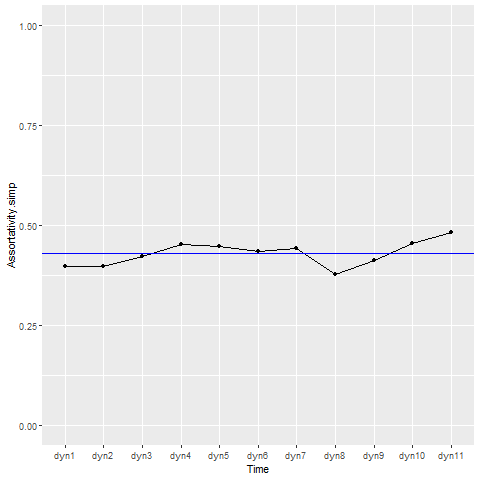
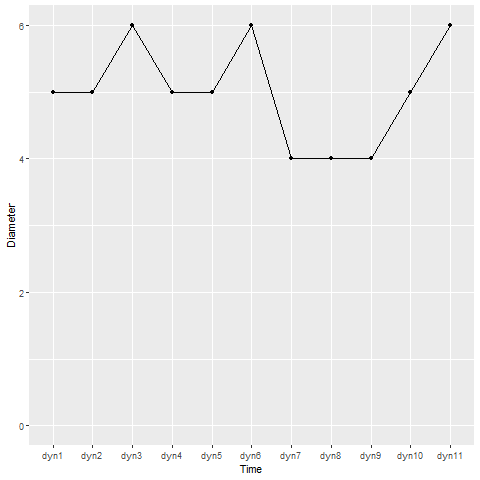
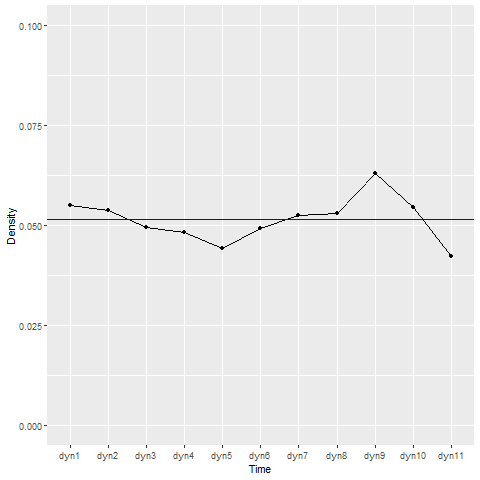
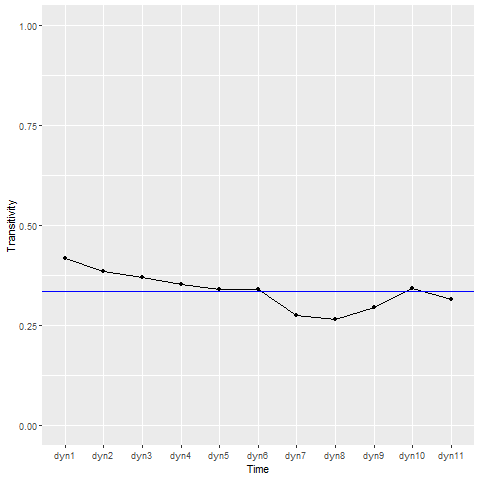
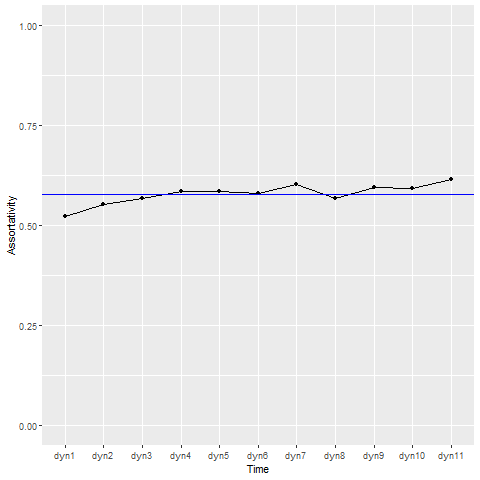
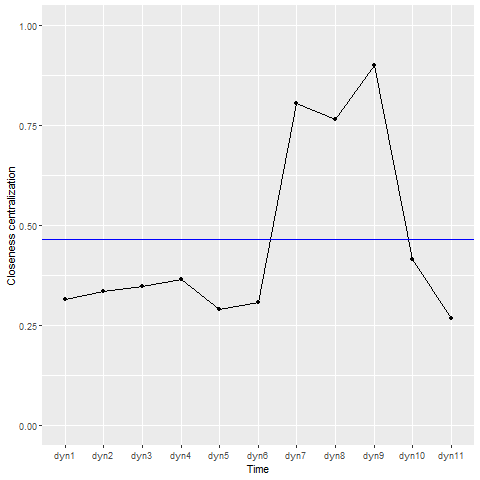
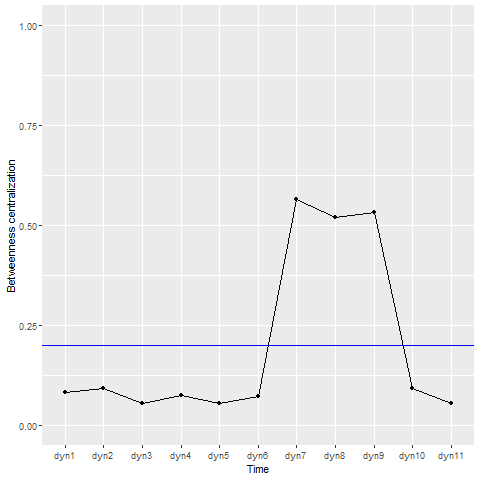
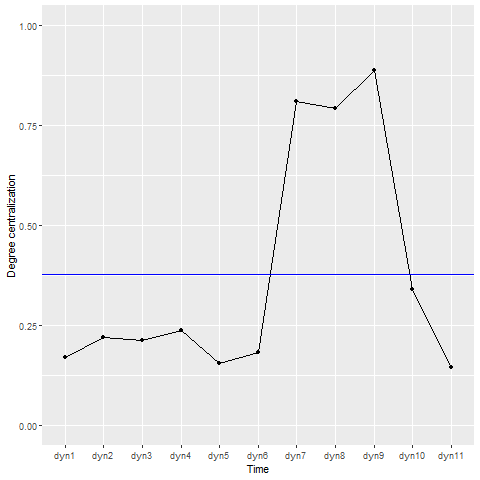


Figure 8 Node level SNA measures on the DynSnap timescales. Top row shows the evolution of assortativity in simplified (left) and un-simplified (right) network. Middle shows density and transitivity. Bottom row shows average path length and diameter.

Source: Own illustration

Figure 9 Centralization of DynSnap networks. Freeman's measures of centralization: degree (left), betweenness (middle) and closeness (right). Three types of graph centralization to capture the various aspects of centralization.



High degree centralization and closeness centralization score may indicate a node directly connected to most other nodes. Such a central actor would similarly have a high degree and a high closeness centrality (low average length of the shortest path to all other nodes). Figure 10 shows the distribution of degree centrality and closeness centrality. Both centrality measures show one node to have a far higher centrality value than the others. For both measures it’s the same node, ID 00.01.00.00.00.00.15.013, that has the highest centrality measure. The betweenness centralization of 0.24 suggests the highly central node is not to acting as a gate keeper between most other nodes. Nevertheless, the individual with ID 00.01.00.00.00.00.15.013 is as well the one with the highest betweenness centrality among all actors.

**Source: Own illustration**

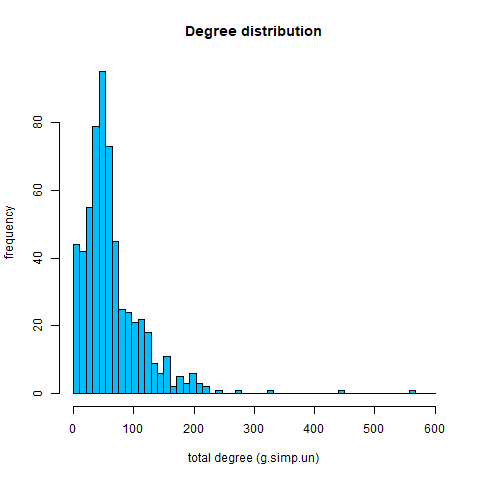
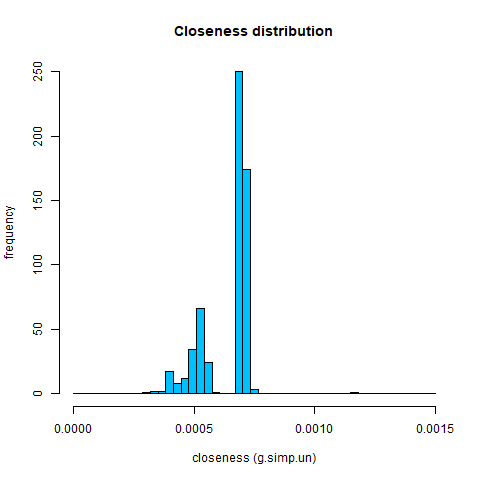
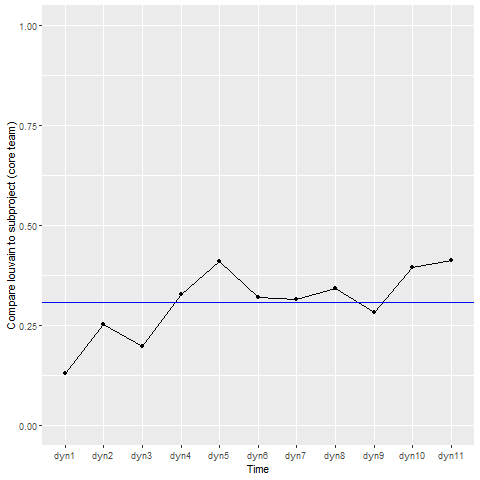
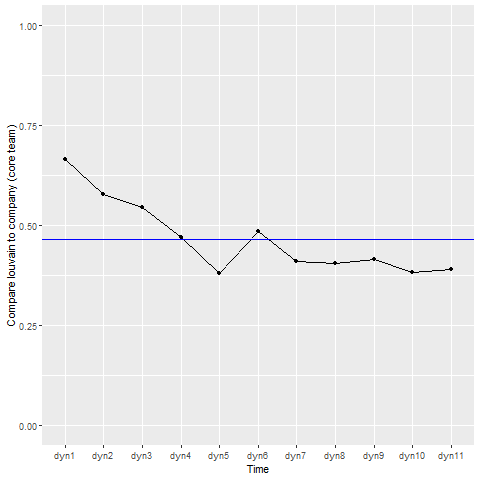


Figure 10 Centrality measures. Distribution of nodes. Degree centrality (left) and closeness centrality (right).

**The left plot in** Figure 10 **shows the degree distribution of the eleven ONE networks. The degree distribution of a network is an important topological property, which impacts the dynamic of diffusion. The distribution of the ONE network is, like most social networks, right skewed. Many nodes have a low degree centrality and just a few are connected to extraordinal many.**

**Source: Own illustration**

Figure 11 Mutual information measure of Louvain communities compared to company of origin (left) and assigned subproject (right) for core team members.



### 7.2.2 Hypothesis 1 & hypothesis 2

Louvain community detection algorithm has a higher modularity than the *walktrap* and *label-propagation* algorithm. The modularity value for the whole dataset is 0.70, indicating a good partition into communities with dense communication within and sparse communication between the nine detected communities. The comparison between the empirically detected communities from the Louvain method and the administrative subproject ascription of core team members, results in a normalized mutual information measure of 0.23. NMI between Louvain and the company of origin is 0.48 and 0.18 compared to the project role an actor got assigned to .   
Figure 11 shows the mutual information measure of the Louvain communities compared to the companies of origin and subproject belonging. The first nine snapshots of the network evolution the mutual information of Louvain communities and *company.type* is higher compared to *subproject.type*. The network for the last two DynSnap timescales are different. NMI comparing *subproject.type* with the empirical communities is higher than the one for comparing the companies. Over time. The curve representing mutual information measure between communities and companies is decreasing. In contrary the evolution of the NMI-line comparing with subprojects, reaches its maximum in dyn11.

### 7.2.3 Hypothesis 3

Using the same Louvain communities and comparing them for site group members, not the core team members, gives a different picture. Comparing Louvain communities to the company of origin and the manually assigned subprojects, gives a mutual information measure value of 0.86 for the companies and 0.22 for the subprojects. NMI of communities compared to roles is zero, because the role of each individual is “site-group member” (*role.type* = 1). As we can see in Figure 12 the curve for both comparisons are constant over time.

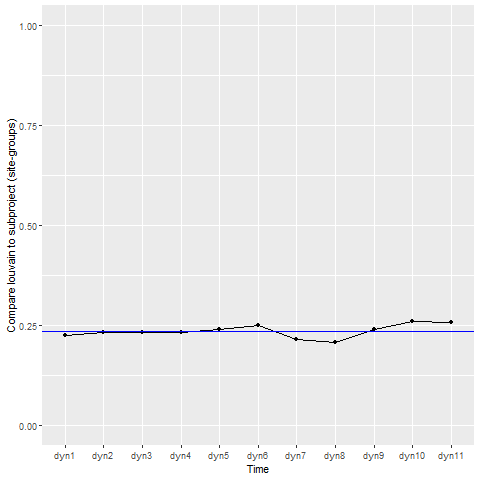
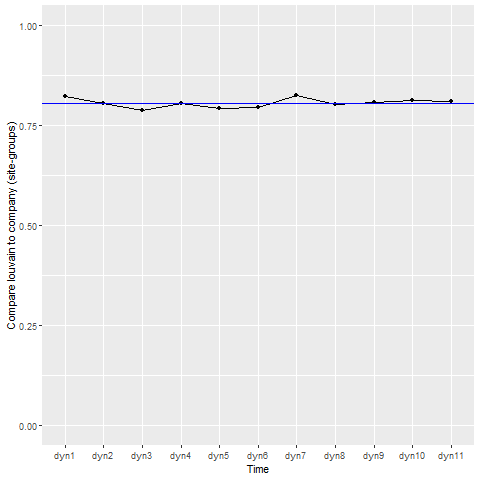


Figure 12 Mutual information measure of Louvain communities compared to company of origin (left) and manually assigned subproject (right) for site group members.

Source: Own illustration

### 7.2.4 SNA helps to understand the connections between individuals

To understand the connections within the informal organization, we need to identify the general connectivity pattern of a network needs to be identified. The static network, including the whole communication dataset of ONE, does not exhibit IBN structures, showed in a IBN value of 0.02. Accounting for the dynamic character of communication, we focus the analysis on the DynSnap networks. Figure 13 shows the in-block nestedness derived from the model of Solé-Ribalta et al. (2018). The observation of IBN structures decreases (from 0.28 to 0.16) between dyn1 and dyn4 and stays constant (0.16 – 0.18) for dyn4 and dyn6. Between dyn7 and dyn10 the IBN value alternate between 0.16 and 0.13 and ends in dyn10 and dyn11 with a value back at around 0.16. Figure 13 contains the IBN-lines for two random networks as well. The control networks are generated with the Erdös-Renyi model and the configuration model. Both lines are lower thatn the one for ONE network. Erdös-Renyi model exhibits an IBN value close to zero and the configuration model shows exactly zero IBN structures. The temporal line for each of the eleven ONE network is significantly different from zero. The paired t-test at the 99% confidence level gives back a p-value of 1.347e-08. Consequently, we can state the presence of in-block nestedness in the ONE email collaboration network at a confidence level of 99%. The observation of IBN structures denotes the informal structure of the ONE project team to consists of nested communities. The blocks, densely connected within and manifesting loose connection between, internally exhibit nestedness.

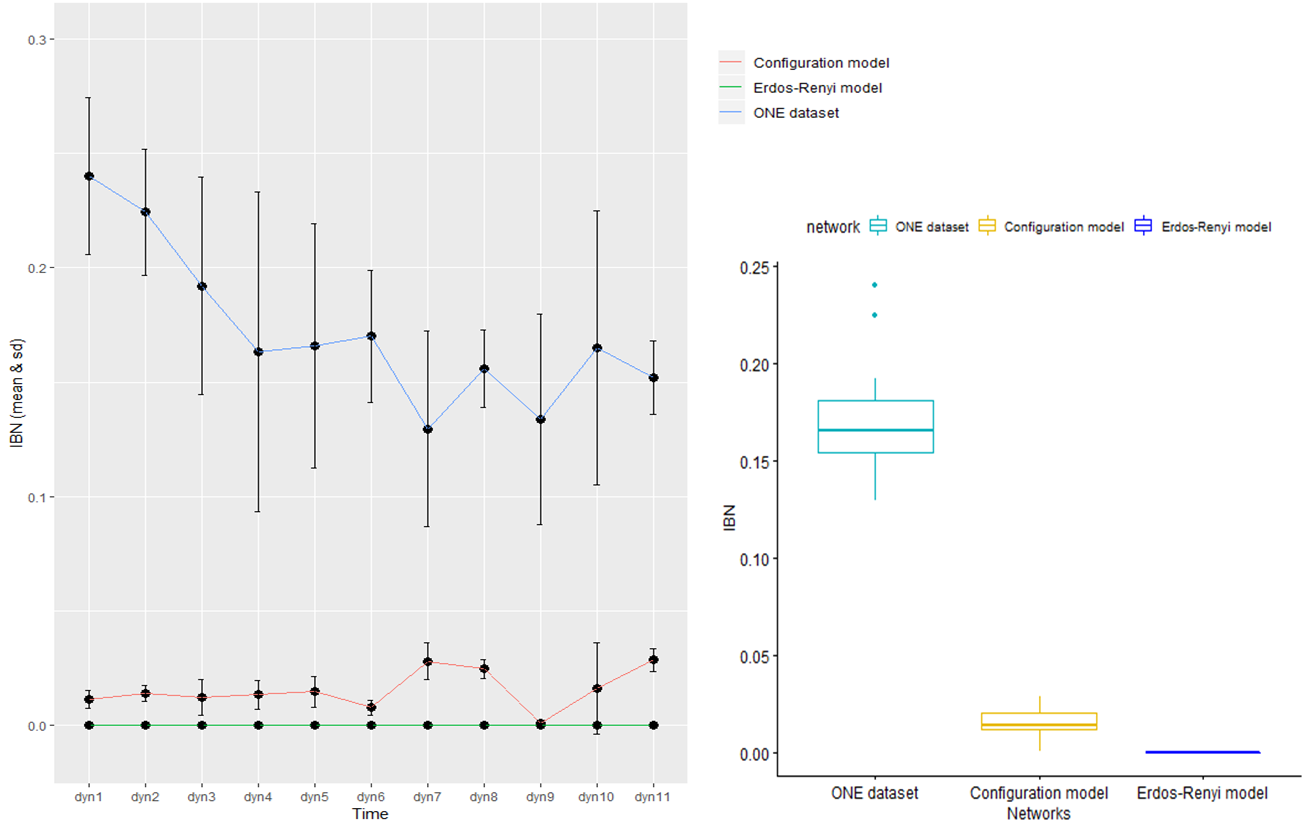
Source: Own illustration

Figure 13 Timeseries and boxplot of in-block nestedness on the DynSnap timescales for the ONE network, Erdös-Renyi model and Configuration model. Values for the ONE dataset are significantly different from the two random model indication in-block nestedness.

### 7.2.5 Insights to improve communication

To generate insights, that may help to improve the communication in further project phases or the next projects, the key results are compared with the project plan and accessible statistical data provided by the project ONE smart solution.

The considered time period in Figure 14 is from the beginning of the showed project plan until the red line (November 2016 until End of June 2018). By comparing the time intervals of the DynSnap algorithm with the relevant period of the project plan we can recognise matching events. End of the period dyn2 (Figure 5) is congruent with the first milestone in the project plan (end of initializing). The decrease of the local event density at the end of dyn2 is because of Easter holidays. The same effect (holidays) causes the density drop in the middle of dyn1 (Christmas and new year holidays) and at the end of dyn4 (summer holidays). The increasing local event density in dyn5 to dyn7 represents increasing involvement of site group members in the project organization. The absolute minimum of the local event density line marks the company holidays during Christmas and new year. The steep decrease in the similarity score after dyn6 is caused by the beginning of the preparations for roll-out of wave 1. The upcoming roll-out demands more interaction with the involved companies outside of the project organization. The end of design milestone comes with the partition between dyn9 and dyn10. The low event density and similarity score for the last timescale is caused by the data of the last month. For this month only, the mails sent by members of the project organization are part of the project team, but not the mails received.

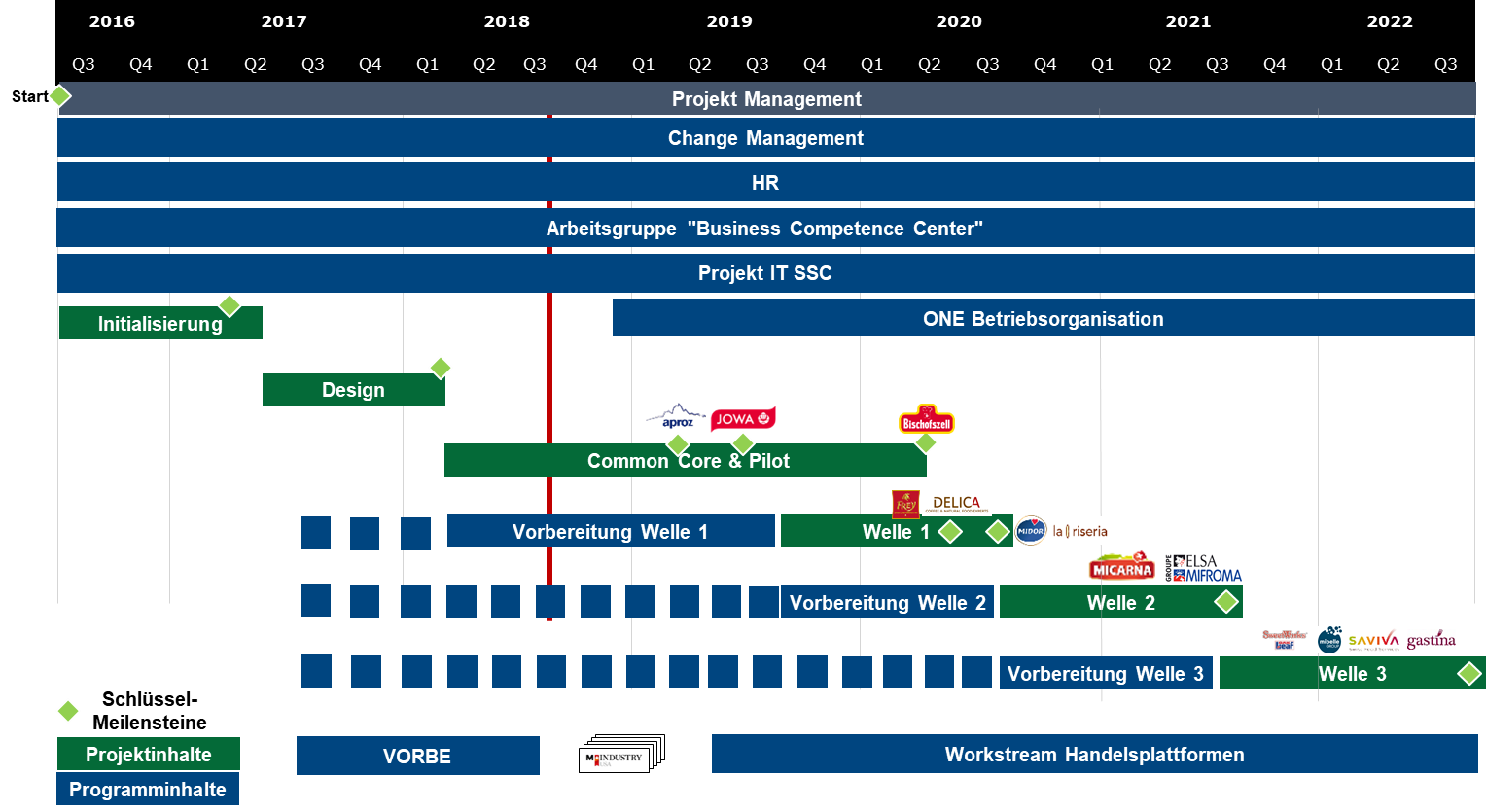
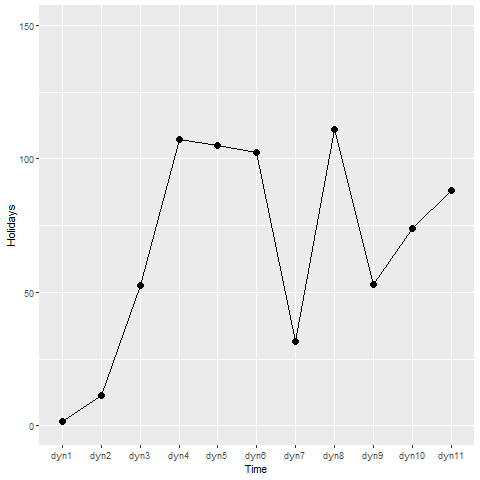
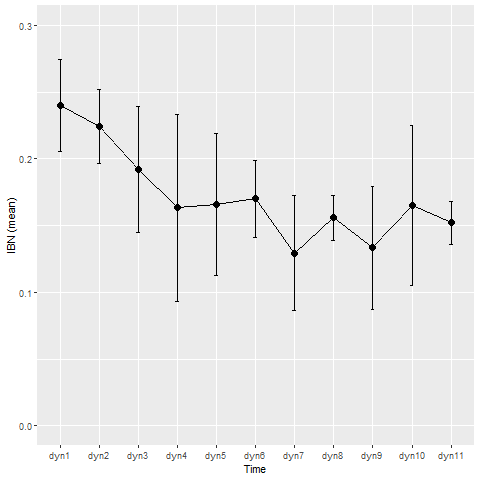
Source: Own illustration

Figure 14 ONE project plan. Including the changes and delays up to November 2018.

The DynSnap partition into timescales and the according event density show a rather high influence of holidays. Figure 15 shows the IBN scores and the number of holidays among ONE project members. The low initial value and the steep slope during the first four-time intervals are because not all 595 members were part of the project organization in the beginning and therefore didn’t have to communicate their holiday to the ONE PMO. The evolution of holidays shows striking similar patterns as to the IBN-curve. In-block nestedness and the number of holidays are on a constant level for dyn4, dyn5 and dyn6. Within the dyn7 timescale ONE exhibits less IBN structures. The number of holidays decrease as well. Both recovering in dyn8 and fall back below the average in dyn9.

Source: Own illustration

Figure 15 In-block nestedness (left) and number of holidays (right) trend. Drops in IBN at dyn7 and dyn9 correlate with drops in number of holidays within project team.



## 7.3 Summary of results

Well-connected individuals more often interact with other well-connected project team members (assortativity = 0.4, simplified network). Overall density of the network is relatively low, whereas on average individuals interact with around 10% of the possible communication partners within ONE. The transitivity score reveals, that the chances an actor communicates with a third member of the project are higher (over one-third) if one of his neighbours already interacts with this person. The dynamic analysis reveals lower transitivity, lower average path length and lower diameter but high centrality during the timescales dyn7 to dyn9 (October 31st 2017 to April 1st 2018).

By defining the slices according to the variability of the system through time with the DynSnap model, we see the defined dynamic intervals match with the evolution of the system. We followingly conclude, the configuration of the network to change over different phases of the project. The changes in configuration detected by the DynSnap algorithm match with events in the project’s evolution. Holidays impact the communication activity and milestones change the composition of communication.

Community detection algorithms reveal the presence of densely intra-connected, but loosely inter-connected blocks among the ONE core team. Hypothesis one and two, stating dense connected communities to correspond with the administrative allocation into subprojects, must be discarded. The communities detected by the Louvain algorithm contain higher mutual information compared to the company of origin, than compared to the subproject to which the ONE member belongs. The dynamic analysis with DynSnap networks shows NMI between communities and the company of origin decrease over time but, increase comparing the empirical communities to subprojects. During the last two time-slices the mutual information is higher for the subproject a core team member belonging to, than for the company of its origin. Nevertheless, the NMI values are relatively low, indicating that Hypothesis 1 and Hypothesis 2 are not supported and communities neither match with the subproject nor the company of a core team member.   
The community analysis for site group members displays that information doesn’t mainly flow within subproject boundaries. Whereas, the high NMI value comparing the Louvain communities and the company belonging, shows empirical evidence for the lack of communication between site group members originating from different companies within M-Industry. Accordingly, hypothesis 3, stating a lack of information for site-group members from different subprojects, cannot be confirmed as well.

The analysis of the general connectivity pattern shows the presence of IBN structures. Social network analysis on the ONE dataset proves the joint occurrence of modularity and nestedness. Connectivity patterns between project members support the recent structural property concept of Solé-Ribalta et al. (2018), generalizing the concept of nestedness and re-framing it to the mesoscopic level. This means, ONE organization consists of blocks, whose internal connectivity exhibits nestedness. By comparing the IBN structures of ONE with two random networks, the paired t-test clearly proves the presence of IBN structures can not be accounted to randomness.

# 8 Conclusion

The emergence of interaction patterns between co-workers is a consequence of the dynamics in the underlaying communication network. The ultimate goal is to understand how information flows within the informal organization and whether this aligns with the perception in the organization.   
The survey about communication tools used among ONE project members shows that email is the most widely used communication tool. Therefore, a social network analysis based on email log captures a major part of the collaborative interaction within a project team in the business world. Social network analysis serves to quantify collaboration patterns, reveals hidden information from the wiring structure of the interaction network and delivers insights to approach the informal organization.

SNA unveils the presence of communities within the ONE dataset. A low value for mutual information measure between subproject assignments compared to the empirical communities, counter-intuitively suggests the communities not to be based on the administrative partition of the project team. More than a common assigned working field, the common history of collaborators seem to be responsible for the interaction boundaries. Assuming co-workers originating from the same company know each other, their connection, possible to exist since several years, seems to stay alive within the new (project) organization. This would explain the mutual information measure between Louvain communities and *company.type* to be higher than NMI of the comparison between communities and *subproject.type*.   
Moderate NMI for both comparisons, suggest neither of assigned subproject-membership nor the company of origin represent the ground truth of empirical communities in the dataset. This finding indicates the underlaying ground truth of the empirical communities is not found. Further investigations searching for the ground truth attribute responsible for the shaped communities are necessary. Because the number of members in subproject is rather large, one can think about subprojects being split up into smaller groups.

Regardless of the absolute score, the NMI results comparing the company and community affiliation, decreases over time, while NMI between subproject and community increases. Initially the boundaries of communication are more associated with the origin from a company, but time spent in the project organization breaks up the “historical” boundaries. On the other hand, core team members communicate more within subprojects, the longer they are part of the project organization (increasing NMI value between *subproject.type* and communities). From a business point of view these findings seem to be natural. The more time someone spends within the project organization and the longer he / she works within the subproject, the more you interact with your new collaborators.   
Communities among site group members aren’t based on subproject affiliation neither. High mutual information between Louvain communities and the site group members company, makes clear corporate affiliation mainly shapes the interaction boundaries. Dense communication among members from the same company and sparse communication between members of different communities, is not surprising. Site group members only work part time for the project ONE and remain in their original company the other time.  
The organization’s perception of a lack of information about other subprojects cannot be proved, but seems to be plausible. The communication among employees from the same company is, with high chances, about the daily operational business. A lack of contact towards members of other companies can be interpreted as a lack of communication about topics concerning the project.

The analysis of the general connectivity pattern proves the presence of IBN structures. This finding aligns with the detection of IBN structures in other social networks (Solé-Ribalta et al., 2018). The prove of this recent method on a new dataset suggests, previous work analysing email datasets may have overlooked this important feature when discussing collaboration networks. Further investigation about the spread of IBN patterns, in collaboration and other social networks, need to be conducted. Assuming broad prove of the occurrence of such patterns, this calls for new models to explain the emergence of IBN structures .   
A nested structure on the mesoscopic network scale may indicate each community to have a hierarchical structure within itself. The team leader aggregates all relevant information for its employees and distributes them among the needs and the hierarchy within the team. This form of organization goes along with modern organizational theories, proclaiming flat and flexible organizational forms.

**Limitations**: This research was conducted using email communication data from a single organization. Hence any claims of generalizability are problematic. Field studies involving data from more organizations are needed before we can arrive at more definitive conclusions. Further research should also compare actual face-to-face communications, telephone communications and meetings along with electronic mail.

# 9 Recommendations for ONE

Existence of communities along with a high modularity (0.7 for Louvain) prove the presence of apart groups within the population, only connected bey a few highly linked individuals. Unfortunately, the Louvain communities do neither match with the assigned subproject nor with the company of origin from a core team member. Due to the lack of the underlaying ground truth of the empirical communities, recommendations are difficult.

Among the site-group members the underlaying ground truth of the communities detected by the Louvain algorithm seems to be clear, the company affiliation. The high score for normalized mutual information shows an omission of integrating them into the project organization. That explains why several site group information get the feeling of not knowing what’s going on within the project ONE. Further it would make clear why the M-Industry companies miss information from ONE. Site-group members are designated to carry information from the project to the companies.

Nonetheless the high modularity for the Louvain partition and the proved presence of in-block nestedness structures demonstrates the existence of loosely connected blocks. Assuming the individuals of an aforesaid group collaborate on closely related topics, the communities could match with the blocks one level below subproject teams. The IBN structures could arise from hierarchy within these subgroups of the subprojects. The most connected individual of a subgroup, probably the subgroup leader, interacts with (almost) all of his/her employees. The subgroup leader is probability also the one, who connects the community to other communities. He/She passes global project information and information about other subprojects, along the intra-group hierarchy, to his/her employees (reason for nestedness within the blocks). Not only it would explain the IBN structures, but it would as well explain the power law distribution of degrees. A few actors (subgroup leaders) have many connections while most actors (subgroup members) only have relatively few connections (probably only within a subgroup). If this is true, it would explain why some core team members feel a lack of information as well. Even if relevant information are distributed by the Project management office, subgroup leaders only transmit a subset of this information to their block and out of this subset only a few are passed down to the bottom of the intra-block hierarchy.

Following these argumentations, aligned with the business and declared as plausible by the project manager of ONE smart solution, I make two recommendations:

1. Ensure the PMO knows who are the subgroup leaders and make them better transmit the information towards their collaborators. This would ensure the subgroup employees see the big picture of the project.
2. Make sure site-group members get better involved in the project organization. One option to do this would be a higher presence time in the project’s office in Dietikon.

# Bibliography

Ahmad, Johanna, Salmi Baharom, and Myzatul Akmam Sapaat (2018), ‘Test case prioritization technique for event sequence test cases based on redundancy factor’, *Journal of Theoretical and Applied Information Technology*, 96 (18), 6041–52.

Almeida-Neto, Mário, Paulo Guimarães, Paulo R. Guimarães, Rafael D. Loyola, and Werner Ulrich (2008), ‘A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement’, *Oikos*, 117 (8), 1227–39.

Atmar, Wirt and Bruce D. Patterson (1993), ‘The measure of order and disorder in the distribution of species in fragmented habitat’, *Oecologia*, 96 (3), 373–82.

Barabási, Albert-László (2014), ‘Network Science Communities’, *K-Mode*, 26, 60.

———, Mark Newman, Duncan J. Watts, Albert-Laszlo Barabasi, and Duncan J. Watts (2006), *The Structure and Dynamics of Networks*, *Information ResearchAn International Electronic Journal*, New Jersey: Princeton University Press.

——— and Márton Pósfai (2014), *Network Science*, Cambridge: Cambridge University Press.

Blondel, Vincent D., Jean Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre (2008), ‘Fast unfolding of communities in large networks’, *Journal of Statistical Mechanics: Theory and Experiment*, 2008 (10), 1–12.

Borgatti, Stephen P. and Jose Luis Molina (2002), ‘What Can Social Network Analysis Do for Management Consulting?’, Boston College.

Borge-Holthoefer, Javier, Raquel A Baños, Carlos Gracia-Lázaro, and Yamir Moreno (2017), ‘Emergence of consensus as a modular-to-nested transition in communication dynamics.’, *Scientific reports*, 7, 41673.

Butts, Carter T. (2008), ‘Social network analysis: A methodological introduction’, *Asian Journal of Social Psychology*, 11 (1), 13–41.

Carroll, Bradley W. (2014), ‘Konigsberg Bridges Problem’, (accessed December 29, 2018), [available at http://physics.weber.edu/carroll/honors/konigsberg.htm].

Chapanond, Anurat, Mukkai S. Krishnamoorthy, and Bülent Yener (2005), ‘Graph Theoretic and Spectral Analysis of Enron Email Data’, *Computational and Mathematical Organization Theory*, 11 (3), 265–81.

Cheong, Christopher and Brian Corbitt (2009), ‘A social network analysis of the co-authorship network of the Australasian Conference of Information Systems from 1990 to 2006’, *ECIS 2009 Proceedings*.

Coleman, James S. (1964), *Introduction to mathematical sociology*, Free Press of Glencoe.

Contreras Arias, Carolina (2017), ‘The assets of the “poor”: rural indigenous livelihoods and their exemplar strategy to overcome scarcity’, *University of St. Gallen, Business Dissertations*, University of St. Gallen.

Csárdi, Gábor, Tamás Nepusz, G Nepusz, C Tamas, Gábor Csárdi, and Tamás Nepusz (2006), ‘The igraph software package for complex network research’, *InterJournal, Complex Systems*, Complex Sy (5), 1695.

Darst, Richard K., Clara Franell, Alex Arenas, Sergio Gómez, Jari Saramäki, and Santo Fortunato (2016), ‘GitHub - rkdarst/dynsnap: Dynamic time-slicing method’, (accessed December 20, 2018), [available at https://github.com/rkdarst/dynsnap].

Darst, Richard K, Clara Granell, Alex Arenas, Sergio Gómez, and Jari Saramäki (2016), ‘Detection of timescales in evolving complex systems’, *Nature Publishing Group*, 6 (July), 1–8.

Diesner, Jana, Terrill L. Frantz, and Kathleen M. Carley (2005), ‘Communication Networks from the Enron Email Corpus “It’s Always About the People. Enron is no Different”’, *Computational and Mathematical Organization Theory*, 11 (3), 201–28.

Flores, Cesar O, Sergi Valverde, and Joshua S Weitz (2013), ‘Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages.’, *The ISME journal*, 7 (3), 520–32.

Fortuna, Miguel A., Daniel B. Stouffer, Jens M. Olesen, Pedro Jordano, David Mouillot, Boris R. Krasnov, Robert Poulin, and Jordi Bascompte (2010), ‘Nestedness versus modularity in ecological networks: Two sides of the same coin?’, *Journal of Animal Ecology*, 79 (4), 811–17.

Fortunato, Santo (2010), ‘Community detection in graphs’, *Physics Reports*, 486 (3–5), 75–174.

Freeman, Linton C. (2004), *The development of social network analysis*, *Document Design*, North Charleston: BookSurge, LLC.

Garton, Laura, Caroline Haythornthwaite, and Barry Wellman (2006), ‘Studying Online Social Networks’, *Journal of Computer-Mediated Communication*, 3 (1), 0–0.

Ghali, Neveen, Mrutyunjaya Panda, Aboul Ella Hassanien, Ajith Abraham, and Vaclav Snasel (2012), ‘Social Network Analysis: Tools, Measures and Visualization’, in *Computational Social Networks: Mining and Visualization*, London: Springer-Verlag, 1–21.

Gloor, Peter A. (2005), ‘Capturing Team Dynamics through Temporal Social Surfaces’, in *Ninth International Conference on Information Visualisation (IV’05)*, IEEE, 939–44.

Golubic, Kruno (2013), ‘Discovering Value in Academic Social Networks: A Case Study in ResearchGate’, in *35th International Conference on INFORMATION TECHNOLOGY INTERFACES*, Zagreb: University Computing Centre - SRCE, 3–8.

Grimm, Alexander and Claudio J. Tessone (2017), ‘Analysing the sensitivity of nestedness detection methods’, *Applied Network Science*, 2 (1), 37.

Hargie, Owen. and Dennis. Tourish (2009), *Auditing organizational communication : a handbook of research, theory and practice*, New York: Routledge.

Hawe, Penelope, Alan Shiell, and Therese Riley (2004), ‘Complex interventions: how &quot;out of control&quot; can a randomised controlled trial be?’, *BMJ (Clinical research ed.)*, 328 (7455), 1561–63.

Himelboim, Itai (2017), ‘Social Network Analysis (Social Media)’, in *The International Encyclopedia of Communication Research Methods*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 1–15.

Jenson, Jeffrey M. (2007), ‘Neighborhood and Community Effects on Individual Behavior and Social Functioning’, *Social Work Research*, Oxford University Press.

Jonhson, Samuel, Virginia Domínguez-García, and Miguel A. Muñoz (2013), ‘Factors Determining Nestedness in Complex Networks’, *PLoS ONE*, (Y. Moreno, ed.), 8 (9), e74025.

King, Nigel. and Christine. Horrocks (2010), *Interviews in qualitative research*, SAGE.

Kirchhoff, Lars (2010), ‘Applying Social Network Analysis to Information Retrieval on the World Wide Web: A Case Study of Academic Publication Space.’, *University of St. Gallen, Business Dissertations*, University of St. Gallen.

König, Michael D., Claudio J. Tessone, and Yves Zenou (2014), ‘Nestedness in networks: A theoretical model and some applications’, *Theoretical Economics*, 9 (3), 695–752.

Kunc, Martin (n.d.), *Strategic analytics : integrating management science and strategy*.

Lancichinetti, Andrea, Santo Fortunato, and Filippo Radicchi (2008), ‘Benchmark graphs for testing community detection algorithms’.

Lee, Sang Hoon (2016), ‘Network nestedness as generalized core-periphery structures’, *Physical Review E*, 93 (2), 022306.

Lewinsohn, Thomas M., Paulo Inácio Prado, Pedro Jordano, Jordi Bascompte, and Jens M. Olesen (2006), ‘Structure in plant-animal interaction assemblages’, *Oikos*, 113 (1), 174–84.

Mariani, Manuel Sebastian (2017), ‘Introduction to Network Theory’, Zurich: University of Zurich.

Migros-Genossenschafts-Bund (2017), ‘Zahlen und Fakten - M-Industrie’, (accessed June 12, 2018), [available at http://www.mindustry.com/de/ueber-uns/uebersicht/zahlen-und-fakten].

——— (2018a), ‘Über uns - M-Industrie’, (accessed June 12, 2018), [available at http://www.mindustry.com/de/ueber-uns/uebersicht].

——— (2018b), ‘Organisation & Struktur | Migros Geschäftsbericht 2017’, (accessed June 12, 2018), [available at https://report.migros.ch/2017/governance/organisation-struktur/].

Migros (2018), ‘Intranet’, *Punkt M.*, (accessed December 12, 2018), [available at www.intranet.migros.ch].

National Research Council (U.S.) (2005), *Network science*, National Academies Press.

Newman, Mark E. J. (2001), ‘The structure of scientific collaboration networks’, *Proceedings of the National Academy of Sciences*, 98 (2), 404–9.

——— (2002), ‘Assortative Mixing in Networks’, *Physical Review Letters*, 89 (20), 1–4.

——— (2006), ‘Modularity and community structure in networks’, *Proceedings of the National Academy of Sciences of the United States of America*, 103 (23), 8577–82.

——— and M. Girvan (2003), ‘Finding and evaluating community structure in networks’, 1–16.

De Nooy, Wouter (2010), ‘Historical developments and theoretical approaches in sociology’, in *Social Network Analysis*, C. Crothers, ed., Amsterdam: Eolss Publishers, 378–461.

Olesen, J. M., J. Bascompte, Y. L. Dupont, and P. Jordano (2007), ‘The modularity of pollination networks’, *Proceedings of the National Academy of Sciences*, 104 (50), 19891–96.

Opsahl, Tore, Filip Agneessens, and John Skvoretz (2010), ‘Node centrality in weighted networks: Generalizing degree and shortest paths’, *Social Networks*, 32 (3), 245–51.

Otte, Evelien and Ronald Rousseau (2002), ‘Social network analysis: A powerful strategy, also for the information sciences’, *Journal of Information Science*, 28 (6), 441–53.

Ouyang, Fan and Christiane Reilly (n.d.), ‘Terminology - Social Network Analysis’, (accessed November 30, 2018), [available at https://sites.google.com/a/umn.edu/social-network-analysis/terminology].

Patterson, Bruce D. and Wirt Atmar (1986), ‘Nested subsets and the structure of insular mammalian faunas and archipelagos’, *Biological Journal of the Linnean Society*, 28 (1–2), 65–82.

Rodriguez-Girones, Miguel A. and Luis Santamaria (2006), ‘A new algorithm to calculate the nestedness temperature of presence-absence matrices’, *Journal of Biogeography*, 33 (5), 924–35.

Rowe, Ryan, German Creamer, Shlomo Hershkop, and Salvatore J Stolfo (2007), ‘Automated social hierarchy detection through email network analysis’, *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis - WebKDD/SNA-KDD ’07*, 109–17.

Sauer, Nils Christian and Simone Kauffeld (2013), ‘Meetings as Networks: Applying Social Network Analysis to Team Interaction’, *Communication Methods and Measures*, 7 (1), 26–47.

Shah, Devavrat and Tauhid Zaman (n.d.), ‘Community Detection in Networks : The Leader-Follower Algorithm’, 1–8.

Shetty, Jitesh and Marina Rey (2004), ‘The Enron email dataset database schema and brief statistical report’, *Information sciences institute technical report, University of Southern California*, 4 (1), 120–28.

Solé-Ribalta, Albert, Claudio J. Tessone, Manuel S. Mariani, and Javier Borge-Holthoefer (2018), ‘Revealing in-block nestedness: Detection and benchmarking’, *Physical Review E*, 97 (6).

Soramäki, Kimmo, Morten L. Bech, Jeffrey Arnold, Robert J. Glass, and Walter E. Beyeler (2007), ‘The topology of interbank payment flows’, *Physica A: Statistical Mechanics and its Applications*, 379 (1), 317–33.

Tacchella, Andrea, Matthieu Cristelli, Guido Caldarelli, Andrea Gabrielli, and Luciano Pietronero (2012), ‘A New Metrics for Countries’ Fitness and Products’ Complexity’, *Scientific Reports*, 2 (1), 723.

Tomasello, Mario V., Mauro Napoletano, Antonios Garas, and Frank Schweitzer (2016), ‘The rise and fall of R&amp;D networks’, *Industrial and Corporate Change*, 26 (4), dtw041.

Tran, Mark and Stefanie Khaw (2006), ‘Enron | Business | The Guardian’, *The Guardian*, (accessed November 8, 2018), [available at https://www.theguardian.com/business/2006/jul/06/corporatefraud.enron].

Watts, Duncan J (1999), ‘Networks , Dynamics , and the Small-World’, *American Journal of Sociology*, 105 (2), 493–527.

Welch, Mary and Paul R. Jackson (2007), ‘Rethinking internal communication: a stakeholder approach’, *Corporate Communications: An International Journal*, 12 (2), 177–98.

Wellman, Barry. and Stephen D. Berkowitz (1988), *Social structures: a network approach*, Greenwich: JAI Press.

Yang, Zhao, René Algesheimer, and Claudio J. Tessone (2016), ‘A comparative analysis of community detection algorithms on artificial networks’, *Scientific Reports*, 6 (March).

Zehnalova, Sarka and Zdenek Horak (2015), ‘Email Conversation Network Analysis : Work Groups and Teams in Organizations’, 1262–68.

Appendix

## A: Network science terms

Following we define indicators describing the structure of a network and the role of nodes. Many more are described in the literature, but we will restrict ourselves to these ones used in this work.

**Degree**

The degree of a node is the number of connections it has. In a social network based on mails, the degree of an actor is the number of mails sent and received. (Ouyang & Reilly)

**Undirected / directed**

In a social network a pair of actors (a dyad) either can be adjacent (connected) or not, if no relation is existent between them. A network is called undirected if existing relationships between actors are symmetric. Symmetric means no distinction between sender and recipient can be made or the distinction is not important. Collaboration networks are undirected because if one person works together with another the connection is reciprocal. If the direction is not inherently symmetric, in the sense that each relationship involves distinct ‘sender’ and ‘receiver’ roles, graphs are called directed graphs. This may be the case with email networks as it is not necessary that if someone writes an email to another person that the receiver also writes one back. (Butts 2008; Otte and Rousseau 2002)

**Loop**

An edge from a vertex to itself is called as a loop. Networks which have no loops and which are not multiplex (do not allow duplicate edges) are said to be simple.

**Path**

A path between two nodes exists if they are connected and consistently reachable by a sequence of node pairs. The number of connections between distinct node pairs defines the length of a path, called distance.

**(Geodesic) Distance**

The shortest path, connecting any pair of nodes, is the geodesic distance. In a directed network the path between two vertices can be different caused by the path direction.

**Average path length**

The average path length in a graph is calculated as the sum shortest paths between all pairs of **vertices divided by the number of all paths.**

**Diameter**

Diameter describes the maximum of shortest distances between any two nodes in a network. If the graph is not connected the diameter is infinite. (Golubic 2013)

**Transitivity or clustering coefficient**

Transitivity measures the probability that the adjacent vertices of a vertex are connected. This is also called the clustering coefficient.

**Assortativity**

The assortativity coefficient measures the level of homophily of the graph, based on some vertex labelling or values assigned to vertices. The coefficient is high, if connected vertices tend to have the same labels or similar assigned values. (Newman 2002)

**Density**

Density of a graph is defined as the comparison between the number of existing connections divided by the number of possible connections between all actors within the social network. Density is an indicator for the general level of connectedness. (Kirchhoff 2010).

**Degree centrality**

Counts how many connections with acquaintances a node has. The more mails, the higher the degree centrality (Opsahl, Agneessens, and Skvoretz 2010).

**Betweenness centrality**

**The number of shortest paths from all vertices to all others a node is laying on is called betweenness centrality. Ascribe a high centrality to nodes connecting distinct communities together.** (Golubic 2013)

**Closeness centrality**

**Closeness centrality measures the mean of all shortest paths to every existing node in the network. According to that its logic is different from the other two centrality measures, meaning nodes with low closeness centrality are more central and have a shorter way to reach all nodes.** (Golubic 2013)

**Communities**

**Detection of communities, one of the key properties of complex networks, is particular interesting. Community structures in networks, showed in** Figure 3**, describes groups of nodes that are more densely connected to each other than to nodes in other communities. These underlays the assumption that networks have natural divisions within it. Normally a node can only belong to one community, which is a simplification that not always holds in reality. Identifying communities in large networks is difficult and raised almost to a independent research field, filled with a lot of literature. (Ghali et al. 2012)**

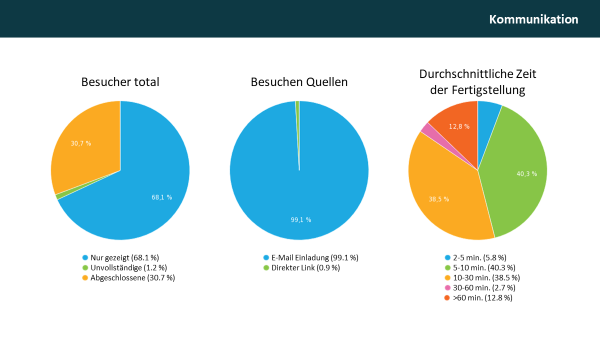
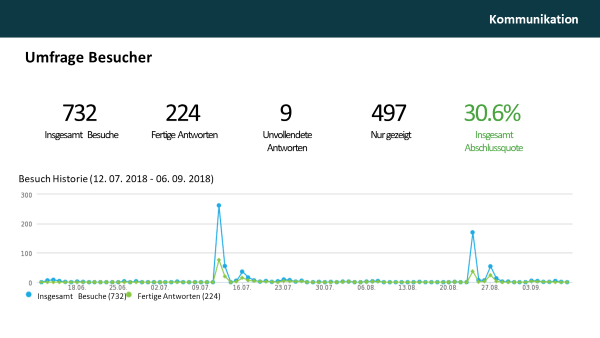
**Modularity**

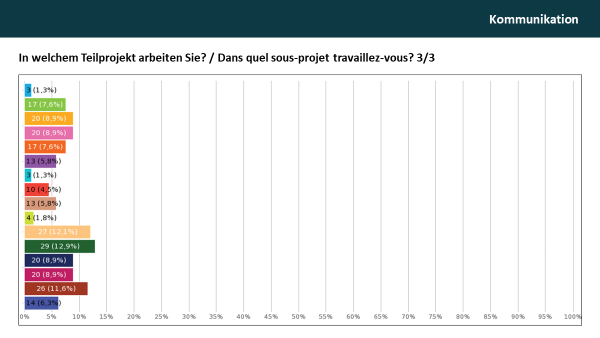
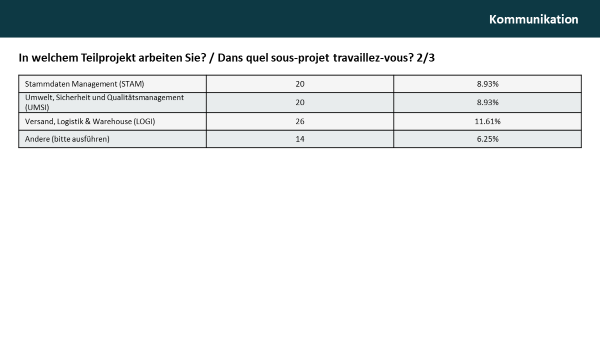
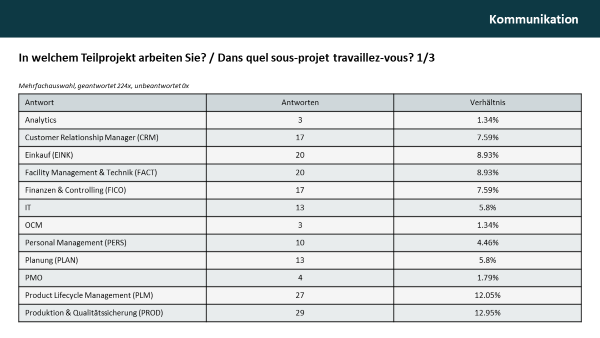
**The modularity of a graph measures how good the division into communities is, or how separated the different vertex types are from each other. It defined as:**

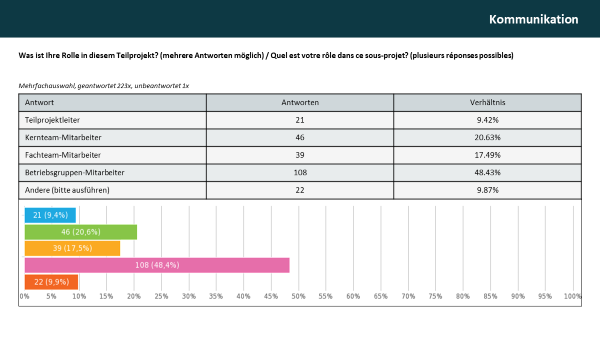
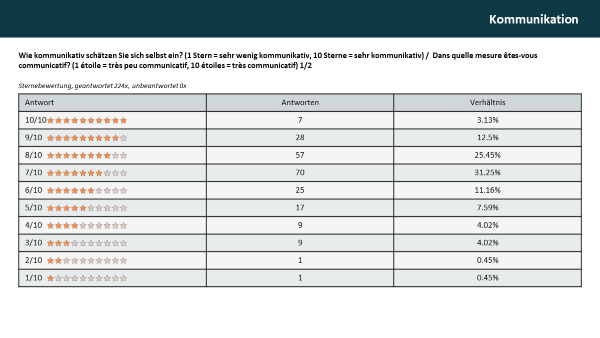
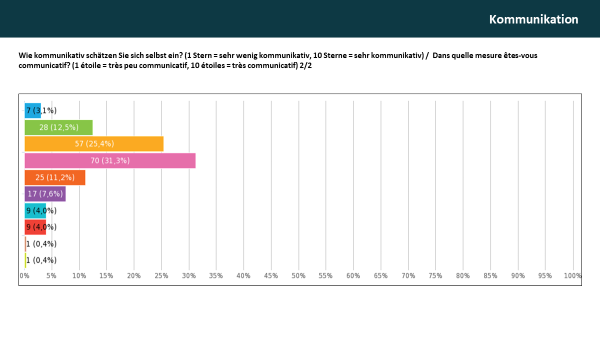
**m is the number of edges, Aij is the element of the A adjacency matrix in row i and column j, ki is the degree of i, kj is the degree of j, ci is the type (or component) of i, cj that of j, the sum goes over all i and j pairs of vertices, and delta(x,y) is 1 if x=y and 0 otherwise. (Csárdi et al. 2006)**

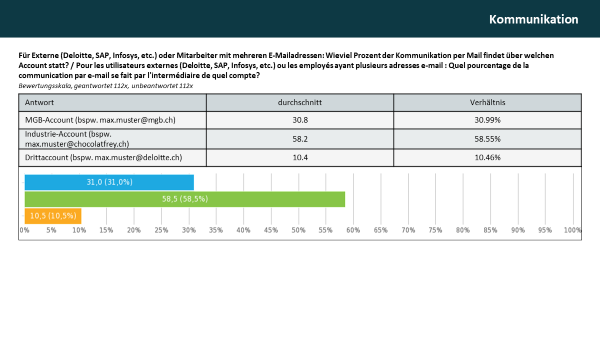
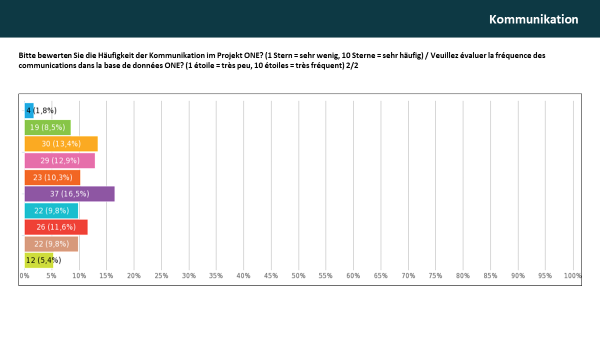
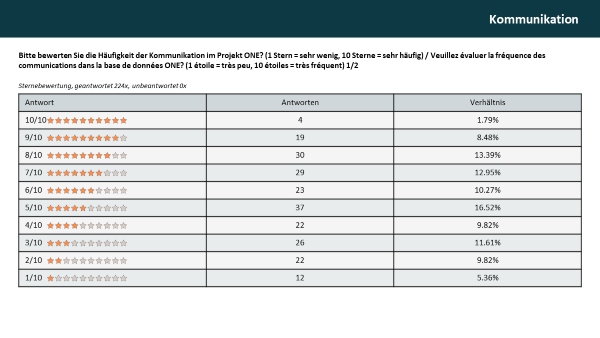
## B: Survey results

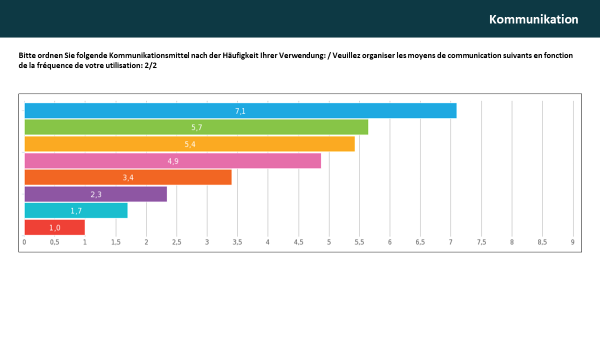
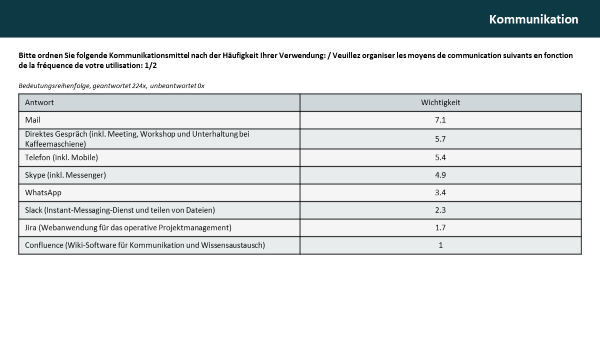
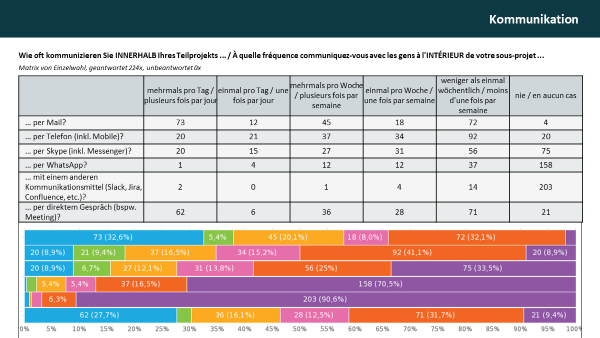
Below I show you the complete survey among the ONE project members. The main goal of the survey was to estimate the share of communication captured analysing emails.

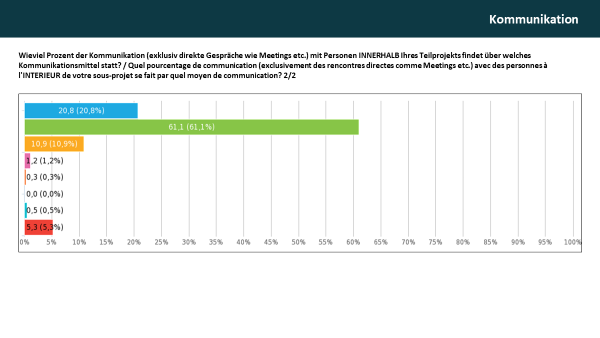
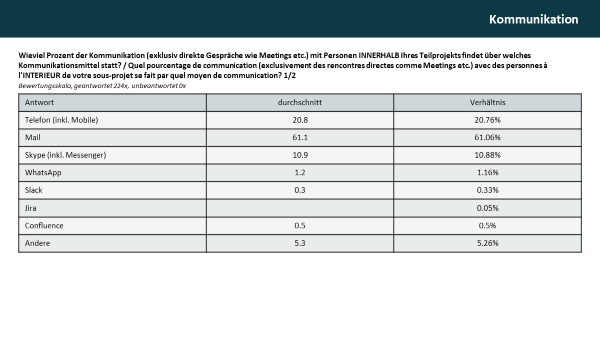
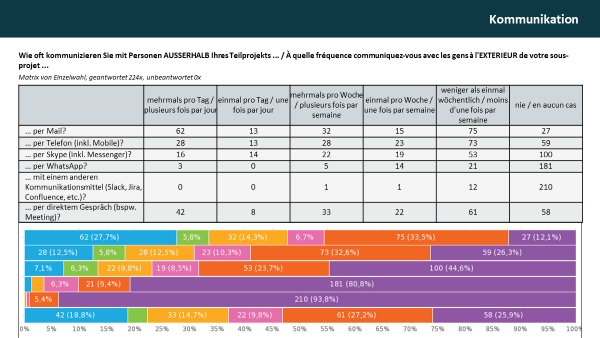
Survey summary:   


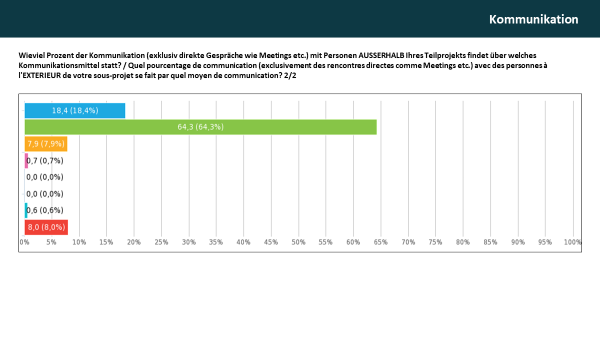
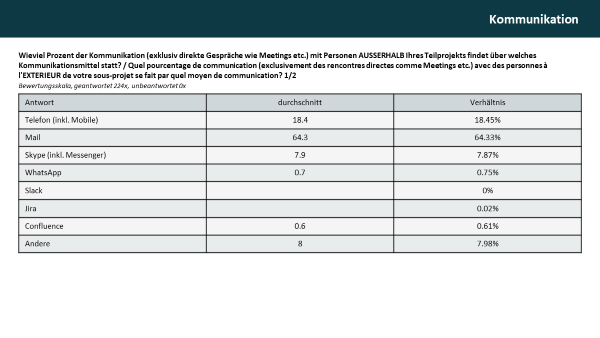
Question 1:   


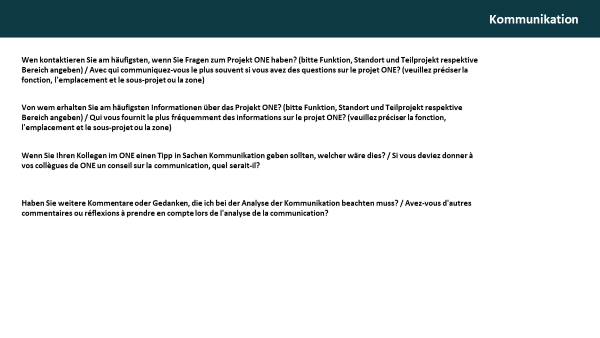
Question 2:   
Question 3:   
Question 4:   


Question 5:   
Question 6:   


Question 7:   
Question 8:   


Question 9:   
Question 10:   


Question 11:   


Open questions (questions 12-15):   


## C: SKRIPT R

…

**Eidesstattliche Erklärung**

Der/Die Verfasser/in erklärt an Eides statt, dass er/sie die vorliegende Arbeit selbständig, ohne fremde Hilfe und ohne Benutzung anderer als die angegebenen Hilfsmittel angefertigt hat. Die aus fremden Quellen (einschliesslich elektronischer Quellen) direkt oder indirekt übernommenen Gedanken sind ausnahmslos als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden.

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . Ort, Datum Unterschrift des Verfassers