Master of Science Thesis

Utrecht University
Faculty of Science
Department of Information and Computing Sciences
Business Informatics

Social Networking in the Workplace: Exploring the Structural Characteristics and Conversational Nature of Enterprise Social Networks

25.01.2016

Author:
Steffen Bjerkenås
s.bjerkenas@students.uu.nl

Universiteit Utrecht

Supervisors:
Prof. dr. ir. Remko W. Helms
Prof. Kai Riemer

[This page is intentionally left blank]

Additional Information

Author:	Steffen Bjerkenås
E-mail:	s.bjerkenas@students.uu.nl
Student ID:	3896161
Project supervisor:	Prof. dr. ir. Remko W. Helms
	r.w.helms@uu.nl
	Department of Information and Computing Sciences
	Utrecht University
Daily supervisor:	Prof. Kai Riemer
	kai.riemer@sydney.edu.au
	Discipline of Business Information Systems
	The University of Sydney
Second examiner:	Dr. Marco Spruit
	m.r.spruit@uu.nl
	Department of Information and Computing Sciences
	Utrecht University

Declaration of Authorship

I, Steffen Bjerkenås, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Title of thesis: "Social Networking in the Workplace: Exploring the Structural Characteristics and Conversational Nature of Enterprise Social Networks"

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;

2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;

3. Where I have consulted the published work of others, this is always clearly attributed;

4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;

5. I have acknowledged all main sources of help;

6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____
Steffen Bjerkenås

Place and date: Vienna, 25/-16

Acknowledgments

In the course of any labor-intensive endeavor, it is not uncommon to find yourself relying on the knowledge, expertise, and support of others in order to succeed. This project has been no exception. From the start to the very end, there have been several involved parties from whom I have received much appreciated (and needed) back-up, and for this I am truly grateful.

First off, I wish to express my gratitude towards my supervisors; Prof. dr. ir. Remko W. Helms and Prof. Kai Riemer. Through our collaboration on this project I have gotten the opportunity to get involved with the research community at the University of Sydney, which has been a much welcomed experience both personally and professionally. Their expertise and academic vigor have helped me overcome my own knowledge gaps many times during this project, and thanks to this I am left with a valuable learning experience. A special thanks to Prof. dr. ir. Helms, who has not only been heavily involved in this project, but who has provided me with challenges and opportunities throughout the entirety of my studies at Utrecht University.

Second, I wish to thank my parents and my brother. The unconditional support I have received, and continue to receive, has been the cornerstone of any achievement I dare to claim. Thank you for always being there.

Last -but definitely- not least, I am forever indebted to my wonderful girlfriend and life partner. Given the sheer amount of effort that goes into a thesis project, the day-to-day progress would not have been possible without her taking on the role as life-coach in times of need. Admittedly, these pages would certainly have suffered a different fate if it was not for her support.

Abstract

Enterprise Social Networks (ESNs) are recent and emergent sets of online platforms that allows for employees to collaborate and share information in the same manner as in more commonly known Social Networking Sites (SNSs) such as Facebook. The proven value-add of these platform to organizations have caused its popularity to soar during recent years, and as a result, causing academia's interest to rise accordingly. However, given the short time-span since ESNs started gaining interest among organizations, the study of how these platforms are being used by employees is still largely unexplored by the scientific community.

By combining methods from the fields of social network analysis (SNA) and content analysis, this research aims to answer the questions of how the conversational nature of an ESN network is related to its structural characteristics, and how the users' conversational nature is related to their structural characteristics. The exploration of these relations has been called for by earlier studies in the field, and is important in the sense that it illuminates how these new and exciting ways of collaborating actually manifests within an organization, and how employees' influence within these networks relates to their conversational topics.

The results obtained during this research show that the degree of clustering in an ESN is related to how much employees engages in discussions and conversations related to generating new ideas and brainstorming. These results provide grounds for postulating that these topics tend to generate more collective interest in participating in these communities as opposed to other topics such as plain information sharing and social talk. Furthermore, results show that employees with a high eigenvector centrality are the most influential in these networks, as they tend to occupy several other central positions simultaneously. Three distinct characteristics can be identified for these actors; they often engage in the exchange of personal opinions, they often provide links to resources of professional value for other employees to make use of, and they often provide updates about the current status of ongoing projects and alike.

As far as this research has been able to discover, this is the first attempt to draw a connection between conversational topics and structural characteristics within the context of ESNs. These results will hopefully assist in advancing the understanding of how ESNs are being adopted and used by employees, and how these informal networks can be analyzed in a meaningful manner.

Table of Contents

Chapter 1

Introduction

1.1 Background

Enterprise social networks (ESNs) are recent and emergent sets of online platforms that have been gaining increasing popularity among corporations and businesses during recent years. Richter, Riemer, and vom Brocke (2011) refers to ESNs as "*the phenomenon of social networking in an enterprise context*", and previous studies argue that the use of ESN within organizations has the potential of providing employees with new ways to collaborate, interact, and share knowledge by employing such technology (Berger, Klier, Klier, & Richter, 2014; Leonardi, Huysman, & Steinfield, 2013; K. Riemer, Richter, & Seltsikas, 2010; Zhang, Qu, Cody, & Wu, 2010). Furthermore, Gartner predicts that by 2016, 50 percent of large organizations will make use of internal social networking platforms, and that in 30 percent of these cases these platforms will reach the same level of importance for the organization as e-mail and telephones (Mann, Austin, Drakos, Rozwell, & Walls, 2012).

Several recent studies have pointed out the limited amount of existing research aimed at the use of ESNs within organizations (Berger et al., 2014; K. Riemer & Richter, 2012; M. Smith, Hansen, & Gleave, 2009). As the use of other social networking services (SNSs) such as Facebook and Twitter have proliferated and gained immense popularity among the general public since their inception, so has the attention from academia and alike with respect to studying the phenomenon of social interaction in these online communities. Studies aimed at online interaction among employees using ESN platforms however, have only recently began to emerge and create interest with respect to analyzing the structure and communication content residing within these ESN communities. As have been the case with SNSs such as Facebook, the study of ESN communities has the potential of providing valuable insights into the connections among its participants, which, in turn, allows one to illuminate the informal hierarchy in which the participants are organized online and the types of communication that is present within the community.

In the pursuit of describing and constructing inferences from social network data, a wide array of metrics and methods have been developed (e.g., Freeman, 1977; Hanneman & Riddle, 2005; Wasserman & Faust, 1994), which is commonly referred to as *social network analysis* (SNA). These methods can, as they have in previous studies, be applied to describe the structure of online communities, the actors within them, and the ties that connect these actors. As ESNs, as with SNSs, provides a digital repository of the exchange of messages between actors that are part of the network, the aforementioned methods and metrics can be exploited in order to describe and make inferences about the nature of ESNs that exists within an organization. Furthermore, through studying the types of communication, i.e. messages, that is exchanged between the users,

a researcher is able to quantify and categorize the semantic characteristics of the conversations going on (e.g., Dougherty, 2005; Weber, 1990), which is often referred to as *genre analysis*. By combining these two methods, it is possible to acquire new and interesting insights into the emerging phenomenon of enterprise social networking, and how the informal structure and communication of an organization is reflected through the use of this technology.

1.2 Problem statement

The formal study of ESNs is still in its infancy. However, there has been a surge in interest among researchers in recent years with respect to studying this emergent phenomenon. The majority of these studies have either approached the use of ESNs within companies from a purely content-minded perspective (e.g., Riemer, Diederich, Richter, & Scifleet, 2011; Riemer, Scifleet, & Reddig, 2012; Zhang et al., 2010), or from a SNA perspective with the aim of localizing experts within the organization or testing the applicability of social capital theories in ESNs (e.g., Škerlavaj, Dimovski, & Desouza, 2010; Smith et al., 2009). However, few studies have aimed to merge the analysis of content within ESNs with the analysis of their structural characteristics, which results in only one aspect of ESNs being focused on. That is, previous research tells the story of the structural characteristics that can be found in ESNs or the type of communication that is being proliferated in these networks, but does not attempt to make the connection between how or if the two different aspects of ESNs are *actually* inter-related. A user might be deemed important or central in an ESN from a SNA perspective, but what is he/she actually talking about? Might there be a link between the conversational nature in an ESN community and the level of engagement from users? Individually, SNA or content analysis cannot answer these questions, but they require a merging of the two different approaches in order to do so. Moreover, the aforementioned studies of ESNs tend to make use of ESN data taken at a specific point in time. This method in terms of data collection leaves little room for exploring the evolution of these networks over time. As Panzarasa, Opsahl, and Carley (2009) writes in the introduction of their paper; "*The extent to which individuals' choices at the local level dynamically affect the global network structure is largely an empirical matter that can only be investigated by using a longitudinal network dataset (…)*". This also applies to ESNs, as it would be a naïve assumption to make that the structure and conversational nature of social networks within organizations remain static over time. As the dynamics of SNSs such as Facebook have been widely studied during recent years (Archambault & Grudin, 2012; Brandes, Lerner, & Snijders, 2009; Panzarasa et al., 2009), longitudinal analysis of ESNs remains largely unexplored.

Exploring the structural characteristics along with the conversational nature of ESNs is a non-trivial problem, as it draws on theory from both SNA and genre analysis. Moreover, the modeling of the dynamic interaction among users and the evolution of a social network as a whole requires the dissemination of literature which can be said to occupy a domain of its own within SNA (Snijders, 2005; Toivonen et al., 2009). Combining these different approaches in order to adequately describe how the conversational nature of an ESN relates to the structural characteristics of the network also requires the adoption of statistical methods that are not directly transferrable from conventional methods applied in social research (Hanneman &

Riddle, 2005, p. 286). However, given that the study of ESNs and its users has previously been given credence both from a SNA- and a communication content-perspective each on their own (Berger et al., 2014; A. Richter & Riemer, 2013; K. Riemer et al., 2010), it follows that combining these approaches could potentially provide new insights into the conversational nature and structural characteristics of ESNs.

1.3 Research questions

Two main research questions were formulated based on the unexplored territory identified within ESN-related research. RQ1 aims to analyze the conversational nature and the network structure of an ESN, and then proceed to explore the relation between these two constructs. The motivation behind this approach is the previously mentioned lack of research in this area, and the potentially new insights that can be gained from combining the two approaches. The term "conversational nature" is here used with respect to the message topics that are discussed within these networks, and the term "network structure" refers to the SNA metrics that exist for describing how the structural characteristics of a social network.

RQ1: What, if any, is the relation between the conversational nature in an ESN and its network structure?

RQ2 puts focus on "central actors" in ESNs and aims to answer the question of what the key characteristics of these actors are. To measure to what extent actors in a social network can be deemed central will rely on their structural positioning within the network, as determined by different SNA metrics. As with RQ1, the "conversational nature" refers to the message topics that exist, but then on a user-level. Their level of participation across multiple sub-groups and central positions within the network is also addressed in this thesis, which in RQ2 is referred to as "redundancy".

RQ2: What are the key characteristics of central actors in an ESN with respect to their conversational nature and their redundancy across different central positions within and between groups?

1.4 Research aim and objectives

This research project aims to advance the understanding of how the communication content of ESN online communities within the context of ESN relates to their structural characteristics. In addition, this research aims to provide further insight into how central users and their communication content relates to their roles within the different communities. In order to adequately describe the structural properties of these communities, there is also a need to investigate what the available methods consists of in that respect. This requires an evaluation of currently available methods to describe such networks, and how they can be applied within the context of this research. Moreover, considering that the study of ESN communities can be said to still be in its infancy, there exists a need to critically evaluate the current state-of-the-art with

specific focus on applying genre analysis and social network analysis methods to describe these networks. The Theoretical Background-chapter will aim to satisfy these requirements and provide insights into the framework in which this research is being carried out. The technicalities and intricacies of graph theory and statistical analysis of network data will be further elaborated in the Research Methods-chapter, and will provide the mathematical and statistical foundation on which the analysis of the collected network data has been based.

The specific objectives that have been set for this research include;

1. *Identify* the current state of the art within genre- and social network analysis of enterprise social networks
2. *Evaluate* the existing metrics used for describing the structure of online social networks
3. *Evaluate* the existing metrics used for describing actors in online social networks
4. *Evaluate* the existing methods applied for analyzing the conversational nature of online social networks
5. *Explore* the relation between the centrality of users and their conversational nature
6. *Explore* the redundancy of central users across multiple sub-groups within the network, and across different central positions
7. *Explore* the relation between the conversational nature of an ESN online community and its network structure

The aim of objective 1 will mainly be covered in the Theoretical Background-chapter, which will provide an overview of previous studies that have been aimed towards analysis of social networks and the exchange of content between their users, with particular focus on enterprise social networks. This objective has a generic and conceptual view of the methods, theories, and previous findings within social network- and genre analysis, hence the keyword *identify*. Objective 2, 3 and 4 will also be satisfied in the Theoretical Background-chapter by applying a more in-depth *evaluation* of current frameworks relevant for this study, while objective 5, 6 and 7 will be met through the results, discussion, and conclusion presented in chapter 4, 5, and 6, respectively.

1.5 Relevance

The importance and relevance of studying ESNs from both a SNA- and a content-perspective, individually, have each been emphasized in previous studies (Berger et al., 2014; Heidemann, Klier, & Probst, 2010; A. Richter & Riemer, 2013; K. Riemer, Scifleet, et al., 2012). There is, however, a lack of studies that employ both methods in order to study these networks. The potential value of using such a combined approach has previously been demonstrated in non-ESN related research. One example is De Laat, Lally, Lipponen, & Simons (2007), that used both content analysis and SNA to research both structural and communicative patterns in a Computer-Supported Collaborative Learning (CSCL) environment. They argued that it is useful to combine the findings from SNA "*[...] with the outcomes of content analysis to interpret whether central participants, as determined by SNA, are also central to the learning and teaching activity within this group.*"(De

Laat et al., 2007). Although this approach has not yet been employed in an ESN context, the need for doing so has previously been expressed by Stieglitz, Riemer, and Meske (2014), who states that "*A future study could conduct content analysis in conjunction with quantitative analysis to better understand the dialogues and communication behavior of the employees in deriving influence from the network*." As will be shown in this thesis, influence can be inferred from a number of SNA metrics, and communication behavior can be described through content analysis. Hence, it follows that combining these two approaches would be a step further in understanding the potential relationship between the two within an ESN context.

Given the rapid adoption of ESN technology in organizations and companies world-wide (Mann et al., 2012), the aim of further illuminating how such technology is being used could provide valuable insight into how implementing ESNs could affect the nature of online collaboration and communication within an organization. Previous studies have also shown that ESNs are dominantly used for professional, as opposed to social, purposes among employees (Berger et al., 2014; K. Riemer et al., 2010). These results provide further ground for postulating that ESNs can be a productive and valuable addition to an organization's communication- and collaboration-infrastructure, and that such technology is welcomed by its users. Moreover, previous results imply that the hierarchy found in informal networks such as ESNs do not necessarily reflect the formal structure that exists within an organization (Stieglitz et al., 2014). As Stieglitz et al. (2014) points out; "*(...) prolific knowledge workers on all hierarchical levels might benefit, as they are able to draw on the network for contributions, not having to rely on information flows along the organizational hierarchy.*" Hence, it is reasonable to assume, as the proliferation of ESNs can be expected to continue rising in the future, that decision makers within companies will need to be aware of these new ways of sharing information in order to keep track of the intra-organizational collaboration networks that exist.

By studying the role of central users in these networks, organizations that use ESN platforms can get a more refined picture of how single employees might play a different role within a company compared to what might be assumed from his/hers formal position. It would be a logical deduction that as the importance of ESNs within companies continues to rise in the future (Mann et al., 2012), so does the importance of users who play a central role in these ESNs. Since little is known about the characteristics of these users with respect to their conversational behavior and their structural position within the network, it would be of interest to acquire this information in order to identify such users. For this to be possible there is also a need for specifying what metrics and methods that can be applied in this pursuit. As previous studies have only applied parts of either SNA or content analysis to accomplish this goal, this project aims at applying a more comprehensive approach in order to provide more insight into how central users communicate within ESNs. The centrality of users can also be measured using different metrics within SNA, i.e. deducting the importance of users relies on how a researcher defines "importance". Different users talk about different topics, and in the same way they can be central to the network depending on how their centrality is measured. Thus, by examining how the conversational nature of users relate to their relative importance, one might get a more refined

picture of how different types of "importance" is related to different conversational characteristics.

From a network perspective, this project is relevant in the sense that it tries to make the connection between conversation topics and structural characteristics on the level of the network. In the same way as with users, it is possible to make inferences about how the message content within a network relates to its structure from a "birds-eye" perspective. For managers and companies it will be important in the future, as with the identification of central users, to identify and characterize what their informal networks look like and what the conversational topics are. Perhaps there are certain types of networks with respect to conversation topics that induce closer collaboration than others, i.e. clustering. Organizations can make use of this information in order to guide the implementation and maintenance of these informal networks, and manage expectations regarding how these networks will be used by employees. By comparing how different types of conversations are related to different types of structural characteristics it would also be possible to say something about how information proliferates through the informal networks, and how this relates to the formal hierarchy within the organization.

As previously mentioned, ESNs, as with other social networks, cannot be assumed to be static entities but rather dynamic networks that evolve over time. Riemer, Overfeld, Scifleet, and Richter (2012) have previously shown an example of how the conversational nature in an ESN evolves and changes as time passes. However, their results only tell one side of the story as the structure of an ESN can also be assumed to change as time passes. By investigating and evaluating how both the message content and structure in an ESN co-evolves, companies can get a better understanding of how ESNs might be adopted and maintained by its users. By only analyzing a network at a certain point time, the emerging trends within a network are ignored and hence, will not provide this insight to managers, users, and decision-makers. This project aims to fill this unexplored gap within ESN research.

1.6 Outline

The rest of this paper will be presented as follows. Chapter 2 will present the results from the literature research that was conducted with respect to methods and theory relevant to this research. Chapter 3 will outline the specific research method employed in order to answer the stated research questions, and chapter 4 will present the results from following these methods. Chapter 5 will discuss the results presented, before presenting final conclusions and notes on future work in chapter 6.

Chapter 2

Theoretical Background

This chapter will present the results of the literature review that was undertaken in order to identify and evaluate existing research and methods relevant for this research project. In the first sub-chapter, an overview will be provided describing the theoretical fundament for describing and analyzing social networks. A general introduction to social networks is given first, followed by an evaluation of SNA metrics available to describe the structural characteristics of social networks and their actors. A sub-chapter is also dedicated to giving an overview of existing models available for describing and analyzing longitudinal networks, and to the additional constraints and considerations that apply to the statistical analysis of social network data. The second sub-chapter aims to give an overview of existing research related to ESNs, both with respect to genre analysis and SNA. Some parts of this project, such as the collection and analysis of longitudinal social network data, intersect with a field of science for which the body of research is extensive and for which an exhaustive literature review would be unfeasible. However, in such cases the areas of research that have been deemed most interesting within the scope of this project are emphasized, and in instances where this is not the case, references to more elaborate, in-depth evaluations of such research are presented.

2.1 Social network analysis

2.1.1 Definition and application

Social networks are defined as a set of nodes that are tied together by a set of relations (Wasserman & Faust, 1994), and can be represented as a graph formally defined as

$$G = (V, E) \tag{2.1}$$

where G represents the whole network, V represents the set of vertices (nodes), and E represents the set of edges (ties) between the nodes. These two sets of information can further be organized in an adjacency matrix, which is a two-dimensional matrix where the rows and columns correspond to the nodes in the network, and the adjacent cells between the rows and columns represent the ties between the nodes. Visually, any network can be represented by drawing the nodes in any geometrical shape, e.g. a circle, and connecting these shapes by lines. Figure 2.1 shows a visual representation of a simple network consisting of three nodes and three ties.

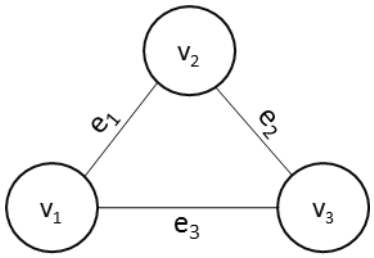

Figure 2.1 - Example of a network

In Figure 2.1, the graph *G* consists of a set of nodes *V* and a set of ties *E* where

$$V = \{v_1, v_2, v_3\} \tag{2.2}$$

$$E = \{e_1, e_2, e_3\} \tag{2.3}$$

The nodes in such networks do not necessarily refer to individuals, but can refer to any entity or a combination of different types of entities, e.g. organizations, scientific papers, or departments in a company. The set of ties between the nodes can also be constituted based on different types of relations, e.g. economic trade-relationships between countries, e-mail exchange between employees, or through user-defined friendship-ties in social networking sites such as Facebook.

The network depicted in Figure 2.1 is referred to as an *undirected* or *symmetric* network, i.e. a network where the ties between the nodes share a "*bonded tie*" (Hanneman & Riddle, 2005, p. 55). In symmetric networks, no attention is given to the direction of the relation between the nodes. For example, in the case of the ties representing friendship relationships in such a network, it would not be possible to distinguish whether the sense of friendship between two arbitrary users is reciprocal or not. In *directed (asymmetric)* networks on the other hand, it is taken into consideration whether a tie between two nodes is reciprocated or not. Using the previously mentioned example, it would be possible to distinguish between friendship ties that are reciprocated and those who are not.

The formal methods used to study and describe social networks is commonly referred to as *social network analysis* (SNA) (e.g., Hanneman & Riddle, 2005; Scott, 2000; Wasserman & Faust, 1994). Early examples of SNA methods being applied to social network data can be found in e.g. Almack (1922) and Milgram (1967), both of which contributed to the discovery of well-known phenomena within social networks; *homophily* and *the small-world effect*, respectively. More recently, the methods embodied within SNA have been applied in a number of research areas and different domains within science. Especially after Internet became commonplace among the public, with multiple online platforms for people to interact and share information, SNA has gained a vast repository of information in which the methods can be applied. Examples of such research can be found with relation to knowledge sharing among individuals using online platforms (e.g.,

Adamic, Zhang, Bakshy, Ackerman, & Arbor, 2008; Berger, Klier, Klier, & Richter, 2014), social capital in social networking sites (SNSs) (e.g., Huffaker, 2010; Mandarano, 2009), and in identifying important actors in online communities (e.g., Ehrlich, Lin, & Griffiths-Fisher, 2007; Zhang, Ackerman, & Adamic, 2007). In each of the aforementioned studies, different methods and metrics found within SNA were applied in order to describe the structure of the networks and the actors contained within them. Based on the results from applying these metrics, a researcher is able to answer questions such as; *who is the most influential individual in network X?*, *how is the network structure in a Twitter-network associated with the proliferation of information among users?*, and *what individuals or groups should be targeted in order to maximize the effect of a specific marketing strategy?*. By applying SNA metrics such as node centrality, degree of clustering, tie strength, and reciprocity, a number of characteristics of a network can be quantified and measured in order to answer questions such as those previously stated.

2.1.2 Describing the structural characteristics of social networks

This section will address the different metrics available within SNA in order to describe a set of nodes and the ties between them on the level of the network, i.e. the metrics applied when looking at the network as a whole as the unit of analysis. The metrics presented should not be considered as an exhaustive list, but a selection of metrics that are relevant for the research at hand in accordance with the stated aims and objectives in Chapter 1.

There are a number of factors that impact the choice of metrics in describing a social network. As stated in the previous sub-chapter, one such factor is whether a network is to be considered directed or undirected. Often is the case that these considerations are constrained by the network data collected, as some datasets might not allow a researcher to distinguish between these two types of networks. Another factor that is to be included is whether the ties between actors are allowed to carry a *weight*, i.e. that a tie between two actors can be assigned a value that is not binary (Scott, 2000, p. 19). The introduction of weights in a network, as opposed to defining ties as either existent or non-existent, has implications for how the network-level metrics are computed. The following sections will give a general overview of the relevant metrics for this research. The formal definition of these centrality metrics can be found in Appendix A.

Network density

The density of a network is defined as the ratio between the number of present ties in the network and the number of possible ties (Hanneman & Riddle, 2005, p. 98). Depending on whether the graph is directed or undirected, this metric can take on different values as the number of possible ties in a directed network is twice as many as in an undirected network. In a *complete network*, i.e. a network where all possible ties are present, the density will be equal to 1, and equal to 0 if none of the possible ties are present.

Reciprocity

Reciprocity within SNA describes the tendency of ties between actors in a directed network to be reciprocated (Hanneman & Riddle, 2005, p. 121). Measuring the reciprocity in social networks only makes sense if the network is directed, as the ties between nodes in undirected networks can only take on two states; existent or non-existent. As an example, if the tie between node A and node B only consists of a directed tie from node A to node B, but there is not a tie present from node B to node A, then the tie between the dyad consisting of node A and node B is not reciprocated.

Reciprocity in directed networks comes in different versions. Borgatti, Everett, and Freeman, (2002) implemented three different versions of this metric; dyad-based reciprocity, arc-based reciprocity, and hybrid, where the latter is a combination of the first two. They defined dyad-based reciprocity as "*the number of reciprocated dyads divided by the number of adjacent dyads*" and arc-based reciprocity as "*the number of reciprocated arcs divided by the total number of arcs*". In other words, the first version will give the ratio between the number of adjacent nodes that are connected by a tie and the number of adjacent nodes that are connected by a reciprocated tie. The second version will give the ratio between the number of ties in the network that are reciprocated ties and the total number of ties.

Average path length

The path length between nodes in a network is a key concept in SNA, and provides the fundament on which a number of node centrality metrics such as betweenness centrality and closeness centrality are calculated (see next sub-chapter). The shortest path between two nodes in a network can be defined as the geodesic distance between the same two nodes in a network (Hanneman & Riddle, 2005, p. 46). The average path length in a network can then be expressed as the mean of all shortest paths in a network (Hanneman & Riddle, 2005, p. 135). For a shortest path to be calculated between two nodes in the network, it is necessary for a path to actually exist between the nodes. Node B is said to be *reachable* from node A, and vice versa, if there exists such a path, or a sequence of paths through intermediaries (Freeman, 1978).

Finding the shortest paths in an unweighted network is straightforward by following the previously mentioned definition. In such cases where a tie between nodes are either existent or non-existent, one can simply count the number of intermediary ties between all pairs of nodes in the network. The most commonly implemented algorithm for finding the shortest path between nodes in such cases in the breadth-first search (Moore, 1959), which sequentially traverses through the different levels of neighbors in a graph structure in order to find the shortest path between any two arbitrary nodes. This algorithm handles both directed and undirected networks, but does not account for the weights of intermediary ties along the shortest paths.

In weighted networks, the shortest path cannot be calculated solely based on the number of intermediary ties between two arbitrary nodes in the network. To cope with the influence of tie

weights, several algorithms have been brought forth. Bellman (1958) proposed an algorithm that took into account the weight of ties by using the analogy of cities being connected by roads where each of the roads would require a different travel time. The algorithm would then aim to calculate the minimal amount of travel time, i.e. shortest path, between any two cities. As opposed to several other algorithms, Bellman (1958) is also capable of handling negative edge-weights and can also handle asymmetric, i.e. directed, networks. Dijkstra (1959) proposed an algorithm which later came to serve as the underlying principle for several other algorithms designed to find the shortest path between nodes. The algorithm works by sequentially calculating tentative distances to all other nodes relative to a node X in the network, and selecting the path of least resistance to all other nodes relative to node X. The algorithm resolves when there are no nodes left unvisited. This way of calculating the shortest path, or path of least resistance, has later been applied within different areas of application such as global positioning systems (GPS), in routing protocols used in computer networks, and SNA. One algorithm which used parts of the method described by Dijkstra (1959) was Johnson (1977), which designed an algorithm that mainly focused on improving the computation time required for weighted, sparse networks. Newman (2001) also proposed an algorithm that was based on the algorithm designed by Dijkstra (1959), where the shortest path between any two pairs of arbitrary nodes were equal to 1 divided by the tie weight between two adjacent nodes along the shortest path. Ties with low weights would therefore be assigned a higher value of their weighted version of the shortest path, and ties with larger weights would be assigned a smaller value. In this specific paper, the weights of the ties between the nodes were based on the count of times an author (node) had collaborated with other authors in writing scientific papers. The algorithm itself however, can be generalized to facilitate any weight-function that determines the weight of the ties between nodes in a network. In more recent times, the problem of finding shortest paths in weighted networks has continued to receive attention in the scientific community (e.g., Opsahl, Agneessens, & Skvoretz, 2010; Yang & Knoke, 2001). The matter of calculating shortest path in weighted is a non-trivial problem which is evident by the diversity of algorithms available in different software packages, and due to the seminal role of shortest path-values in calculating a range of centrality metrics for nodes in a social network.

Clustering coefficient

The clustering coefficient of a social network describes the degree to which nodes in the network tend to cluster together (Hanneman & Riddle, 2005, p. 124). A distinction is made in literature between the local clustering coefficient and the global clustering coefficient, where the former relates to the degree of clustering in the local neighborhood of a given node (Watts & Strogatz, 1998), and the latter refers to the tendency of nodes to cluster together on the level of the network (Luce & Perry, 1949). In other words, the local clustering coefficient gives a value of the density in a local neighborhood relative to a given node in the network, and the global clustering coefficient is calculated by counting the number of *closed triplets* in the network, i.e. the number of triplets of nodes that are complete graphs, and dividing this value by the total number of open triplets in the network (triplets of nodes that are connected by two ties). Another version of the clustering coefficient is commonly referred to as the *average clustering coefficient* which can be

measured by taking the mean local clustering coefficient over all nodes in the network (Watts & Strogatz, 1998).

The previously mentioned metrics for the degree of clustering in a network, does not take into account the weights of the ties that exist between nodes either in a local neighborhood (local clustering coefficient), or between the ties that exist between open or closed triplets of nodes (global clustering coefficient). Opsahl & Panzarasa (2009) proposed a generalized version of the global clustering coefficient for weighted as well as directed networks, which consisted of four methods for determining the influence of weights in the global clustering coefficient; arithmetic mean, geometric mean, minimum value of the weights that make up a triplet of nodes, and the minimum value of these weights. This method of calculating the weighted global clustering coefficient in a network was also implemented in a software package for the programming language R (Opsahl, 2009). Another solution for calculating the local clustering coefficient in weighted networks was proposed by Barrat, Barthélemy, Pastor-Satorras, and Vespignani (2004). In their paper, they propose a method that not only measures the local density of a local neighborhood relative to a given node, but that measures the strength of ties that exist within the local neighborhood. The method proposed has been implemented in the *tnet* package by Opsahl (2009), where the different methods for determining the influence of local weights in a neighborhood are the same as can be found in the method for calculating the global clustering coefficient.

2.1.3 Describing the centrality of networked actors

As the previous section gives an overview of different methods used to describe the structural characteristics on the level of the network, this section will give an overview over methods available for describing the centrality of nodes. Different centrality metrics can be used to measure in what respect a node can be said to be important. The following sub-sections will describe the different centrality metrics that have been applied in this research and what inferences about a node's importance can be made based on the value of each centrality metric.

Weighted degree centrality

Freeman (1978) noted in his seminal paper on node centrality that even though there was a wide consensus on the claim that "*(...) centrality is an important structural attribute of social networks*", there was a need to clarify the concept of centrality and get a better idea of how it can be used to understand human groups. One such centrality metric is referred to as *degree centrality*, which is a measure of the number of ties to other nodes any given node has in a network (Freeman, 1978). The absolute value of the number of ties a node has can be applied in undirected networks to get an understanding of the connectedness of a node, but in directed network the degree centrality can be distinguished further by introducing two different metrics for degree; in-degree and out-degree (Hanneman & Riddle, 2005, p. 97). In-degree can then be measured by counting the number of incoming ties, and out-degree by counting the number of outgoing ties for any given node in the network. By examining the in-degree, one can get an impression of the popularity of a given actor (Hanneman & Riddle, 2005, p. 97). As with social structures elsewhere in society, it

is intuitive to assert that individuals with a large number of incoming connections can be regarded as popular, and by applying this metric for actors in online social networks it is possible to make the same conclusions. The out-degree can help in identifying influencers in the network (Hanneman & Riddle, 2005, p. 97). Actors who score high with respect to their out-degree can often be viewed as active, or gregarious, individuals (Opsahl et al., 2010) who have a large network of directly connected actors and therefore have access and the possibility to influence others.

The method presented in the previous paragraph for determining the degree of networked actors can only be applied in unweighted networks. By only counting the absolute number of incoming or outgoing connections from a given node, one does only get an impression of the count of connections and not the weight of the relation to or from adjacent nodes. Barrat et al. (2004) generalized the concept of node degree to weighted networks by multiplying the number of connections of a given node with the sum of their weights. Their generalized version of weighted degree centrality in can easily be applied in directed networks by counting the number of incoming and outgoing connections to and from a node and multiplying this number with their respective weights.

Betweenness centrality

The betweenness centrality of a node depicts how often a given node appears on the shortest path between pairs of nodes in the network (Freeman, 1977). Nodes carrying a relatively high value for betweenness centrality in a network are assessed to have power and influence in a network in the sense that these users can withhold information and can to some extent influence the ability of other users to communicate with each other (Freeman, 1977; Hanneman & Riddle, 2005, p. 163). In other words, such users possess the power to "make things happen" and act as a *broker* in the network through their central position.

Brandes (2001) proposed an algorithm that not only aimed to decrease the computing time needed for calculating betweenness centrality of nodes, but that also was able to integrate tie weights along the shortest path between nodes. For finding the shortest path between nodes, the algorithm relied on Dijkstra's (1959) algorithm for finding the path of least resistance by taking into account the weights along the shortest paths. It is worth noting that the algorithm presented in Brandes (2001) only takes into account the cumulative sum of tie weights between nodes, and not the absolute number of ties when calculating the betweenness centrality. In an attempt to account for both the weights along the shortest paths between two nodes and the number of ties these weights were distributed along, Opsahl et al. (2010) used their generalized version of shortest paths for weighted networks in order to calculate the betweenness centrality. In their paper they argue that the original features of betweenness centrality as proposed by Freeman (1977) were ignored by not taking into account the number of ties in calculating betweenness centrality, and hence the relative importance of tie weights and number of ties could be skewed. To balance the influence of number of ties and tie weights, Opsahl et al. (2010) introduced a

tuning parameter that allowed for a user-defined value which specified the relative importance assigned between each of the two parameters.

Closeness centrality

The closeness centrality of a node is defined as the inverse of the sum of distances, i.e. shortest paths, from the node to all other nodes in the network (Sabidussi, 1966). As the degree centrality is only able to describe the adjacent connections of a node, it does not reflect how close a node is to all other nodes in the network outside its own immediate neighborhood (Hanneman & Riddle, 2005, p. 153). For this purpose, betweenness centrality is a more appropriate metric. Nodes who carry a relatively large score according to closeness centrality can be viewed as central actors in the sense that they are closer to all other actors in the network, and are therefore able to spread information faster compared to other nodes with a lower closeness centrality score (Berger et al., 2014).

As with the previously mentioned centrality metrics for nodes, the original methods for calculating closeness centrality do not consider tie weights. As was done by Brandes (2001) with betweenness centrality, Newman (2001) generalized closeness centrality to weighted networks by summarizing the weights of the ties in calculating the closeness centrality of a node. Similar to other generalized version of centrality metrics for weighted networks, their algorithm for finding the shortest paths between nodes were based on the original work by Dijkstra (1959). Opsahl et al. (2010), as they did with Brandes' (2001) generalized version of betweenness centrality for weighted networks, expanded upon Newman's (2001) work by implementing an algorithm that allowed for a pre-defined tuning parameter which can be used to set the relative importance between number of ties and tie weights in calculating shortest paths.

Eigenvector centrality

The three previously mentioned centrality metrics all use the structural position of a given node in order to calculate the centrality of the same node, while eigenvector centrality determines the centrality of a node based to the centrality of its neighboring nodes (Bonacich, 1972). In social networks, a proper analogy to describing the concept of eigenvector centrality would be that being connected to popular individuals increases one's own popularity or status (Bonacich & Lloyd, 2001). Hence, such actors can be regarded as central in their own right by being affiliated with other central users, and have the ability to exert power by influencing its neighbors who themselves have power. Other similar centrality metrics have been implemented for use in web search engines, perhaps the most known example being the "PageRank"-algorithm applied in the Google search engine (Brin & Page, 1998).

Newman (2004) generalized eigenvector centrality to weighted networks by taking into account the tie weights in the adjacency matrices between nodes used to calculate eigenvector centrality. Moreover, Bonacich and Lloyd (2001) proposed a generalized version of eigenvector centrality which they termed *alpha centrality*. They showed in their paper that

the alpha centrality would be a more appropriate measure in certain cases, and that the conventional way of calculating eigenvector could give peculiar and strange results if applied incorrectly.

2.1.4 Longitudinal social network data

Social network data is often collected in the form of a "snapshot", i.e. the data captures the static structure of the social network at a certain point in time and not the dynamic evolution of the network prior to when the snapshot was taken. Social networks are, however, dynamic in nature. For example, in SNSs such as Facebook and Twitter the intensity of communication between actors are likely to vary over time, and the actors can become part of or disappear from each other's ego-networks as time goes on. That is to say, it is reasonable to expect that both the characteristics on the level of the network as well as the users, as discussed in the previous section, will change depending on which point in time the network is being observed. Studying such dynamic networks requires a different approach both in terms of data collection and analytics, for which a large number of previous scientific inquiries have aimed at describing. An exhaustive in-depth analysis of all available models and frameworks available for longitudinal network data, given the sheer magnitude of existing research on this topic, is out of the scope of this project. More detailed overviews can be found in a variety of sources, e.g. Doreian and Stokman (2013), Snijders (2005), and Toivonen et al. (2009). However, the next paragraphs will present an excerpt of existing models and frameworks that were evaluated for use in this project, and their applications within social network analysis.

One method for modeling dynamic social network data that has received a great deal of attention and interest is the Stochastic actor-oriented model proposed by Snijders (1996). In his paper he presented a method based on continuous Markov chain models (Holland & Leinhardt, 1977), where historical data of a network along discrete points in time creates the basis for statistically analyzing the driving force(s) behind its evolution. As the title of the paper suggests, the actions of the users (nodes) in the network were assumed to be driving force behind the evolution of the network. Furthermore, it assumed that a preference function can be defined for each actor in the network, i.e. that not only the current state of an actor is known, but also what the actor's preference is with respect to the relation with other actors in the network. The model assumes that when actors in the network have the opportunity to impact their outgoing ties towards other actors, their choices will aim to optimize this preference function. Such a preference function can, for example, be defined based on social theory as was done in the case study presented in the same study (Snijders, 1996), where the model was applied to study the mechanisms behind the forming of friendship-ties in a fraternity. One of the strengths of the method proposed by Snijders (1996) is its ability to incorporate both intended change in the network based on the actions of the actors, and random change due to external or unknown factors. The model can also facilitate a wide spectrum of actor-bound attributes, both dynamically changing attributes such as opinions, and static attributes such as gender or race. One limitation however, is the assumption that a network is continuously observed along several discrete points in time, which in many cases of social network data is an assumption which is not

met. Moreover, the model requires that the set of nodes observed over time is a constant entity, i.e. the model does not incorporate the dynamics related to actors joining or leaving the network over time. As with the example of Facebook and Twitter given in the previous paragraph, it would in several similar instances be naïve to assume that the set of actors in a network remain unchanged, especially if the network is being observed over a longer period of time.

A commonly applied method to measure the evolution of social networks is by extracting descriptive measures of a network at different points in time, and observing how these measures change. This method can perhaps best be described as collecting an array of consecutive "snapshots" of the same network, where snapshot 1 is taken at time t_1, snapshot 2 at time t_2, and so forth. This method of observing the evolution of a network has been applied in a number of studies related to SNA. Newman (2001a) applied this method to measure the phenomenon of preferential attachment in scientific collaboration networks. By collecting data about which scientists had collaborated together during a certain time interval, and then collecting the same data at a later time interval, Newman (2001a) found that the probability of two scientists working together is a function of the number of previous collaborations and previous collaborators. Panzarasa, Opsahl, and Carley (2009) conducted a longitudinal study of an online community at the University of California, Irvine, by observing the communication activity between students over a time period of seven months. By partitioning the data in to time intervals they were able to observe the change in network metrics such as the average path length, degree distribution, and reciprocity over the given time period. In a similar fashion, Halatchliyski and Cress (2014) collected snapshots linked articles on Wikipedia in an attempt to model the evolution of the knowledge-networks over time. Using SNA metrics such as betweenness centrality and eigenvector centrality they aimed to describe the significance of pivotal articles with respect to the future evolution of the knowledge-network. A number of other studies also attest to the applicability of this method, e.g. Ahn, Han, Kwak, Moon, and Jeong (2007), Kossinets and Watts (2006), and Angeletou, Rowe, and Alani (2011). The main advantage of using consecutive snapshots in studying the evolution of a network lies in its simplicity. As each snapshot produces a complete network structure and hence, can be analyzed independently, it allows a researcher to make use of common SNA metrics in order to describe the network in each snapshot. This longitudinal array of network metrics can then be further used to statistically analyze the evolution of the network and the actors within it. However, by collecting snapshots, perhaps with a large time interval in between (e.g., Halatchliyski & Cress, 2014), one runs the risk of ignoring intermediary events which might otherwise have altered the outcome of the statistical analysis.

A third method available to model longitudinal social networks is *multi-agent systems*. The principle behind multi-agent systems is rather straight-forward; a network can be represented by inter-connected nodes who have the capacity to autonomously act on their own behalf and make decisions about their own behavior within the network (Wooldridge, 2002, p. 3). The nodes in a multi-agent system can interact with each other through any form of communication depending on what system is being represented, which can be anything from robots in a factory to online social networks. The idea is that the input towards the nodes comes from the environment and

the output from the nodes affects the environment, and that this cycle carries on causing dynamicity within the network (Wooldridge, 2002, p. 16). With respect to social networks, individuals can be defined as agents who are operating within an environment consisting of other individuals, and where the individuals and the environment have the capacity to co-evolve (Carley, 2003). By representing the individuals and the environment according to a set of dynamically changing attributes, the multi-agent model can be used to observe how and in what way they co-evolve, enabling statistical analysis of the network over time. An example of an application of multi-agent systems within SNA include Sabater and Sierra (2002), who applied SNA metrics in a multi-agent model of a network in order to calculate the reputation of its actors. Lospinoso, McCulloh, and Carley (2009) conducted a longitudinal study of collaboration among cadets in a military environment by modeling the cadets as agents in a multi-agent system, and studying their collaboration networks using SNA. An advantage of applying multi-agent systems in studying social networks is the complexity in which the agents and the environment can be modeled, which allows for the dynamics of social networks to be put under severe, statistical scrutiny. However, one limitation in applying this method is related to the same reason, namely that human behavior is not trivially easy to model. Hence, the method puts strict requirements on the theoretical foundation that needs to exist in order to apply it in a scientifically valid manner (Wooldridge, 2002, p. 11).

2.1.5 Statistical analysis of social networks and networked actors

Analyzing social network data often differs from standard statistical analysis in the way that the observations within the dataset cannot be deemed as independent samples from a larger population (Hanneman & Riddle, 2005, p. 287). As an example, if one has a small network consisting of the nodes A, B, and C, and the nodes A and C are only connected through B, then it would be unreasonable to assume that the two relations existent in the network are independent of each other since they both have node B in common. One could in such cases make use of standard statistical tools to explore the relationship between the two relations, but in order to make statistical *inferences*, i.e. be able to generalize the results to some larger population, a different approach must be adopted (Hanneman & Riddle, 2005, p. 286). A number of appropriate methods for conducting statistical analysis of social network data can be found in Hanneman and Riddle (2005, p. 285-319).

A common denominator for conducting statistical tests of social network analysis is the adoption of permutation tests in order to calculate standard errors and confidence intervals. The essence of permutation tests is to randomly shuffle the observations that are included in the statistical analysis in order to calculate the sampling distribution, and then comparing the observed statistical result with the permuted result in order to draw conclusions from the data (Hanneman & Riddle, 2005, p. 287). In other words, since the distribution among the population of a sample taken from a social network is often unknown, the sample itself becomes the population through permuting the sampled observations. The purpose behind doing this, as opposed to just applying more common statistical methods when analyzing social network data, is to avoid making too

optimistic conclusions based on the results from the analysis, i.e. type 1 errors or false positives (Hanneman & Riddle, 2005, p. 286).

2.2 Enterprise Social Networks (ESNs)

2.2.1 History and definition

The use of social media among the general public has become wide-spread over the last decade, and it would not be far-fetched to submit that the majority of people now have either a relationship with, or at least knowledge of, the concept of social media. Since the launch of Facebook in 2004, the number of monthly active users on the platform is as of today approaching 1.5 billion users (Statista, 2015), with other SNSs such as Twitter and Instagram also experiencing rapid growth during recent years. This development did not go unnoticed among organizations and companies, and several studies have previously documented how different types of social media were used in different work-related aspects. Efimova and Grudin (2007) conducted a study among weblogs written by employees in Microsoft over a five-week period in 2005, and found that the employees used weblogs as a way of communicating with external and internal parties, organize their work, and to show the "*human side*" of the company. DiMicco and Millen (2007) studied the use of Facebook among IBM employees in order to create a better understanding of how social media was being used in a professional setting. Their results showed that Facebook was becoming part of the daily routine among young hires, and that the users of Facebook in IBM mainly fell into one of three categories; those who considered Facebook to only occupy their personal domains and not to be considered part of their work-life, those who used Facebook mainly for self-presentation and online identity, and those who used Facebook to maintain social connections within the company. In the conclusion of their paper they stated that; "*As social networking sites become more popular in general, these sites are likely to become an integral part of the workplace.*"

The previously mentioned studies mainly dealt with commercial social media platform which could both be used for internal as well as external communication. Furthermore, the use of these technologies were mainly driven by the employees' own initiative and not directly sanctioned or encouraged from a managerial level. However, a rising need for creating better systems that could facilitate the storage and retrieval of internal company knowledge gave rise to a number of in-house platforms that enabled employees to connect and share information (Fulk & Yuan, 2013). These platforms came in a number of different shapes, but one of the most commonly implemented solutions were in the form of internal wikis (Leonardi et al., 2013), that enabled employees to create, store, and retrieve information from a central database. It should be noted that Leonardi et al. (2013) categorized internal wikis under their definition of *Enterprise Social Media (ESM)*, which they deem to be a broader term than ESN. The use of these internal wikis were studied by Majchrzak, Wagner, and Yates (2006), who through a survey among 168 users found that the use of wikis were perceived to have three main benefits; enhanced reputation, work made easier, and improved company processes.

During recent years there has been a rising influx of commercial platforms that offer more integrated solutions for social networking in the workplace, with features that resemble more that of other publicly available SNSs such as Facebook and Twitter. Examples of such platforms include Jive Software's Jive[1], Microsoft's Yammer[2], and Tibco Software's Tibbr[3]. The popularity of adopting social network technology in the workplace has also gained a rising interest among corporations in recent years. Gartner, Inc. predicts that by 2016, 50 percent of large organizations will make use of internal social networking platforms, and that in 30 percent of these cases these platforms will reach the same level of importance for the organization as e-mail and telephones (Mann et al., 2012). Recent research has also found evidence that the use of social media for internal communication and collaboration has been welcomed by the employees, and that this new way of exchanging information through informal networks has several perceived benefits both from a user- and a business-perspective (Fulk & Yuan, 2013; K. Riemer et al., 2010; Zhang et al., 2010). From a user perspective, some of the benefits observed by users include increasing the visibility of what others are working on within the company (Zhang et al., 2010), building social capital (Risius, 2014), and to localize experts more effectively (Shami, Ehrlich, Gay, & Hancock, 2009). From a business perspective, previous studies have shown that the internal use of social media in the workplace can potentially increase the effectiveness of knowledge sharing on an organizational level (Fulk & Yuan, 2013) and enable more effective collaboration across silos within the company (Zhao, Rosson, Matthews, & Moran, 2011). However, as Fulk and Yuan (2013) points out in their paper, social media technology is only a medium for enabling collaboration and information sharing, and does not automatically ensure any immediate benefits by simply being implemented. Furthermore, the perceived usefulness might deviate depending on in what context such platforms are implemented, as the potential benefits might be different across different industries and work environments (Riemer, Overfeld, Scifleet, & Richter, 2012).

As with the concept of social networking technology in general, the use of social media within organizations and companies has taken on several names and abbreviations. Leonardi et al. (2013) uses the term *enterprise social media (ESM)* to refer to "*Web-based platforms that allow workers to (1) communicate messages with specific coworkers or broadcast messages to everyone in the organization; (2) explicitly indicate or implicitly reveal particular coworkers as communication partners; (3) post, edit, and sort text and files linked to themselves or others; and (4) view the messages, connections, text, and files communicated, posted, edited and sorted by anyone else in the organization at any time of their choosing.*" Kügler and Smolnik (2014) uses the term *enterprise social software platforms (ESSP)* to refer to "*intra-organizational social software tools*", and Richter, Riemer, and vom Brocke (2011) lends the description "*the phenomenon of social networking in an enterprise context*" to the term *enterprise social networking (ESN)*. These different terms do undoubtedly have overlapping features, and by investigating recent studies that deal with the phenomenon of social media in the workplace it is apparent that these different terms and definitions are sometimes used interchangeably to describe the same concept. Henceforth, the

[1] www.jivesoftware.com
[2] www.yammer.com
[3] www.tibbr.com/

term and definition provided by Richter et al. (2011) will be adopted to refer to the use of social media in an enterprise context.

2.2.2 Genre analysis of ESNs

A substantial interest has been invested during recent years in studying the conversational nature in ESNs, i.e. what employees talk about using this medium. One of the most common methods applied for studying the conversational nature in online communities is known as *content analysis* (Pfeil & Zaphiris, 2009), which is defined as "*a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use*" (Krippendorff, 2004, p. 18). However, the term *content analysis* is on occasion used interchangeably with the term *genre analysis*, where the latter is the most frequently adopted in recent literature related to the study of ESNs (e.g., Riemer, Scifleet, & Reddig, 2012; Riemer, Diederich, Richter, & Scifleet, 2011). Another term that is applied when solely studying the textual characteristics of conversations is *text analysis*, which can be seen as a subset of, or a less broader term than, content analysis (Berger et al., 2014). For simplicity and clarity, given the different terms applied when referring to the study of conversations between online community users, the term used during this project is *genre analysis.* According to Swales (1990), a genre is defined as "*a recognizable communicative event characterized by a set of communicative purposes identified and mutually understood by the members of the professional or academic community in which it regularly occurs.(...)*" As can be read from this definition, the genres have to be identified within a framework made up by a set of *purposes*, which in turn implies that the construction of such a framework should be guided by what the research at hand aims to infer from the genre analysis. In other words, the genre framework, henceforth referred to as the genre repertoire, can vary across studies depending on what typology makes sense relative to the research question. This is also the case with respect to the study of conversations within ESNs, as will be further illustrated in the following paragraphs. Moreover, the manner in which a genre repertoire is conceived can be partitioned into two main methods; inductive or deductive coding (Mayring, 2000). Inductive coding is the method where a researcher uses the communication data as basis for creating the genre repertoire, i.e. the genres emerge as the data is being analyzed (Mayring, 2000). Deductive coding refers to the method of defining a genre repertoire *a priori* according to a theoretical foundation, i.e. the genre repertoire is developed before the analysis and classification of the messages takes place (Mayring, 2000). This method is relevant if a researcher aims to make inferences from the conversational data within a framework of e.g. social capital theory, where there exists sufficient theoretical basis for linking certain types of communication to, say, behavioral characteristics of users within the online community. The following paragraphs will give an overview of previously conducted genre analysis within the context of ESNs, which will culminate into a comparison of the different methods applied in, and characteristics of, these studies (Table 2.1).

DiMicco et al. (2008) conducted what appears to be the first study of ESNs with respect to the conversational nature in these communities. They used a combination of genre analysis and

interviews in order to study the use of Beehive[4], an ESN created by IBM for internal use. In their paper they collected data from both interviews and the platform itself, where the data collected from the platform consisted of user profiles and the communication between the users. One of the research questions put forth in their paper was related to the types of content that were being exchanged between users in order to explore the use of the (then) recently adopted ESN platform. To accomplish this, they categorized a sample of content from 100 users into three categories; caring, climbing, and campaigning, each of which reflected the perceived purpose behind the content being posted. These categories were developed based on feedback from employees through interviews about their individual motivations behind using the ESN. It should be noted that in their study the content analyzed was not solely textual, but also included photos. Their findings implied that users within the company used Beehive for purposes related to either social stimulation, career advancement, or the campaigning of projects, and that the different types of content (e.g., photos, status messages) were being used in different ways depending on the ambitions of the user.

In the same fashion as with DiMicco et al. (2008), Zhang, Qu, Cody, & Wu (2010) conducted interviews as well as data collection from an ESN platform in order to provide new insights into the use and adoption of ESN. Their data was collected from a Fortune 500-company's use of the ESN Yammer[5] over a period of ~4 months, which comprised 458 users and 3391 messages. Of these, 300 messages were sampled and classified in accordance with an *a priori* genre classification scheme consisting of five distinct genres; "Me", "Conversation seeking", "Share news or new found", "About Yammer", and "Others". Each of these categories was further partitioned into sub-categories, resulting in a total of 13 categories included in the classification scheme. The dominant top-genre in their case study was found to be "Share news or new found", in which the dominant sub-category was "Internal news", meaning news that revolved around internal matters within the company. Their findings implied that the conversational nature of ESNs differed from what had previously been observed in other SNSs such as Twitter, and that the vast majority of messages sent on the platform were of a professional nature as opposed to social, or not work-related, messages. Moreover, only 16% of the messages were classified as "Me", implying that the ESN was mainly being used to engage with other co-workers as opposed to users using the platform to promote themselves and their persona.

Zhao, Rosson, Matthews, and Moran (2011) collected data from a large IT company's use of the Yammer ESN platform with the purpose of exploring what employees talked about in these networks. Their data consisted of the accumulated messages over a five week period among 40 employees spread across 19 different groups, with a total of ~900 messages being transmitted between the users during this time period. Each message was categorized according to a genre classification scheme consisting of seven different categories; "Project task status", "Info/idea sharing", "Other work status", "Question", "Social personal", "Availability, and "Other". Their results implied that, in their case study, the conversational nature within the company was heavily dominated by "Project task status"-messages, followed by "Info/idea sharing"- and "Other

[4] No longer an active ESN platform
[5] http://www.yammer.com

work status"-messages. An interesting observation made in their paper was that only 9% of messages were not work-related, whereas the remaining 91% of messages were aimed at productive purposes within the company.

Riemer, Diederich, Richter, and Scifleet (2011) conducted a genre analysis of the ESN communication activities between employees is Capgemini, a large consultancy company. The study, as with the two previously mentioned studies, was based on the company's use of Yammer. In total they classified ~1000 messages according to a classification scheme which was conceived inductively, i.e. the genres emerged as the messages were coded. In total, they defined five main categories with additional sub-categories, where the main categories consisted of "Opinion and clarification", "Problem solving and support", "Information sharing", "Updates and notifications", and "Others". Their findings suggested that the use of Yammer in Capgemini was largely dominated by "Opinion and clarification", which in their paper included all messages that were part of interactive discussions where personal opinions were voiced and exchanged. In addition, they conducted a second analysis where they used the threads as the unit of analysis instead of the messages. Seven additional genre categories where defined in order to categorize the threads, and the results showed that the majority of conversation threads were related to either "Joint problem solving" or "Aligning of activities". They concluded, based on the findings, that employees in the company mainly used Yammer as a conversational medium to exchange opinions and engage in discussions, as opposed to self-promotion as had been previously observed in other forms of social media such as Twitter.

A similar study as was done by Riemer et al. (2011) for Capgemini was conducted for the consultancy company Deloitte by Riemer, Scifleet, and Reddig (2012). By analyzing 1985 messages over a two-week period they aimed at creating a better understanding of how ESNs were applied in a knowledge-intensive environment. The manner in which they developed a genre classification scheme resembled much of the approach found in Riemer et al. (2011), where each message was classified according the perceived purpose of the message from the perspective of the community. The analysis resulted in a scheme comprised of seven categories; "Discussion", "Updates", "Idea generation", "Information sharing", "Problem solving and advice", "Social and Praise", and "Other". Each of these top-level categories were each further refined to several sub-categories. The most dominant genre in their case study was "Discussion", which accounted for 38% of all messages analyzed. The four genres "Updates", "Problem solving and advice", "Social and Praise", and "Information sharing" were found to be somewhat equally represented in the dataset, ranging between 12%-15% each. A notable observation made by the authors was that the results from their study implied that the use of Yammer in the context of Deloitte deviated from what had previously been observed in other similar knowledge-intensive contexts. More specifically, they did not find evidence of Yammer being used for work coordination in any significant way, but rather that it was mainly being used by employees for conversational purposes. The authors inferred for this that Yammer in this particular case represented a "*personalization approach*" to ESN, meaning that the ESN was not being used to store and retrieve knowledge, but rather as a medium for employees to exchange ideas and engage in discussions with each other.

Riemer and Richter (2012) compared five case studies of the use of ESN in organizations in order to develop a framework of use cases commonly observed in the previous studies. The message data collected in the five studies were from two different ESN platforms, namely *Communote*[6] and *Yammer*. Combined, a total of 7437 messages had been analyzed across the five previous case studies. Riemer and Richter (2012) argued that "*All of these studies applied the same general research design and analysis methods (…)*", and therefore allowed for the compilation of these results in order to develop a common framework that can be applied to study the conversational nature of ESNs. Their analysis resulted in a framework consisting of twelve distinct genres; "Problem solving", "Idea generation", "Input generation", "Information storage", "Work coordination", "Meeting coordination", "Meeting coordination", "Status updates", "Event notification", "Discussion and opinion", "Informal talk", and "Other". These genres were then further grouped in order to develop a set of six use cases that represented the benefits offered to companies by adopting ESNs for internal use; "Socialising", "Organising", "Crowdsourcing", "Information sharing", "Awareness creation", and "Learning and Linkages". Moreover, they put emphasis on that the use of ESNs varied depending on the context in which they were being applied. To distinguish between the different contexts they made a distinction between the use of ESN in a large enterprise-, teamwork-, and project coordination-context. In their conclusion they point out that the framework can be used by decision-makers who want to get a better understanding of how the use of ESN can benefit their organization, depending on in what environment and context they are operating.

As previous studies have mainly focused on collecting and analyzing data on the level of the whole network within an ESN, Riemer and Tavakoli (2013) conducted a study that aimed to analyze the conversational nature on the level of ESN groups in Yammer. Furthermore, as previous studies have mainly applied genre analysis with individual messages as the unit of analysis, Riemer and Tavakoli (2013) classified the conversations on the level of the thread. Their method for classifying the threads involved sampling a total of 7,409 messages spread across 2,978 threads, where the conversational nature of the thread in its entirety was used as basis for the analysis. The genre repertoire was developed using inductive coding, which resulted in nine top-level genres being identified. These top-level genres were further partitioned into a total of 30 sub-genres. The sample itself was taken from a total of 13 groups, in addition to the "All-network stream", which are messages that are visible to all users within the Yammer network as opposed to group messages which are only visible to members of the groups. An interesting observation made in their paper was that the results of the genre analysis using threads as the unit of analysis was markedly different compared to a previous study using the same dataset but then with messages as the unit of analysis. Furthermore, they conducted a cluster analysis of the 13 sampled groups which resulted in the emergence of 4 different group types based on their conversational nature; conversational groups, solution-oriented groups, people-centered crowdsourcing groups, and information-sharing groups. In the conclusion of their paper, they emphasize that their results imply the vital importance of information sharing as a central practice in their particular

[6] http://www.communote.com

case study, and that this type of communication serves as a common characteristic of the groups studied.

Kügler and Smolnik (2014) proposed in their paper a framework for classifying the different types of use cases that are common with respect to the use of social software within organizations, which they named ESSUF (Enterprise social software use frameworks). To develop this framework they conducted feedback-sessions from employees in the form of workshops in addition to analyzing messages retrieved from an online discussion event where the topic revolved around organizational social software and business impact. The information retrieved through the workshops and the online discussion forum, along with existing literature, were then further analyzed in order to develop the set of use cases in their framework. As a result, four distinct use cases were identified; "Consumptive use", "Contributive use", "Hedonic use", and "Social use". The amount of activity in the online discussion forum used to (partly) develop the framework was stated in their paper as ›2,100 messages spread across ›2,700 users, but the amount of messages actually classified were not disclosed. The exact relative distribution across the genres in the online discussion was also not explicitly mentioned. Furthermore, as opposed to the previously mentioned studies, the message content in this case was not directly retrieved from a company's internal use of an ESN but rather from a forum talking *about* ESN. Their genre analysis can therefore be seen as a sort of meta-analysis describing employees' perception of social software within the organization.

Berger et al. (2014) set out to "*investigate the structural characteristics of value adding users in ESN (...)*" in order to better understand the behavior of such users within the context of ESN. One of the methods they applied was genre analysis of an undisclosed company's internal use of Yammer over a period of ~4 years. A total of 8,142 messages were coded according to their paper, where the genre repertoire used for the coding consisted of two categories; "Professional" and "Non-professional". The messages coded only consisted of messages that had received at least one bookmark or at least one "like" by Yammer-users. The results showed that among both bookmarked and "liked" messages, the vast majority (81% and 94%, respectively) served a professional purpose, and that there was a large overlap (51%) between the top 1% users with respect to the number of bookmarked and "liked" messages within the network.

Risius (2014) studied the relationship between social capital and conversational practices by analyzing the content of an undisclosed company's enterprise microblogging (EMB) system. He conducted an initial content analysis of 6,306 messages, which was used to conduct a cluster analysis which resulted in the users being partitioned into two separate "*communication types*"; Meformers and Informers. The coding scheme was developed based on previous theory related to user behavior which resulted in a genre repertoire consisting of four different genres; "Factual information", "Self-statement", "Relationship-indicator", and "Appeal". The users in both of the two groups were further studied and different characteristics of the users such as group memberships and betweenness centrality were used to measure their social capital. Social capital in this case was an abstract construct consisting of three sub-constructs, which also were constituted by a total of eight variables which were used to measure individual indicators of

social capital. The results of the study indicated that the communicative behavior of the users had impact on two out of the three sub-constructs of social capital, namely structural and relationship capital, while the sub-construct cognitive capital did not appear to be affected.

The studies mentioned and described in this section shows how genre analysis can be applied within the context of ESN in different ways, and how it can be used to provide profound insight into the conversational nature that exists between employees in an informal network. The different studies differs in terms of the genres applied in order to classify each message depending on the research question, the ESN platform used in the study, and what the author(s) wish to measure. As an example, in cases such as DiMicco et al. (2008) where the authors wish to investigate the ambition of ESN users by analyzing their messages, one can expect to find a different genre repertoire compared to Riemer et al. (2011) where the genre analysis aims to investigate the nature of the internal communication itself. In order to compare and visualize the similarities and differences between previous studies mentioned in this section, a series of characteristics about each study have been compiled and structured. The result of the comparison can be found in Table 2.1.

Publication	DiMicco et al. (2008)	Zhang et al. (2010)	Zhao et al. (2011)	Riemer et al. (2011)	Riemer et al. (2012)	Riemer and Richter (2012)	Riemer and Tavakoli (2013)	Kügler and Smolnik (2014)	Berger et al. (2014)	Risius (2014)
Basis for genre repertoire	Feedback from employees through interviews	Literature (deductive coding)	Message data (inductive coding)	Message data (inductive coding)	Combination of message data and literature	Compilation of previous case studies	Message data (inductive coding)	Employee feedback, literature, and message data	Literature (deductive coding)	Literature (deductive coding)
Unit(s) of analysis	Messages and photos	Messages	Messages	Messages	Messages	Messages	Threads	Messages	Messages	Messages
Number of top-genres identified	3	5	7	5	7	12	9	4	2	4
Top-level genres identified	"Caring" "Climbing" "Campaigning"	"Me" "Conversation seeking" "Share news or new found" "About Yammer" "Others"	"Project task status" "Info/idea sharing" "Other work status" "Question" "Social personal" "Availability" "Other"	"Opinion and clarification" "Problem solving and support" "Updates and notifications" "Information sharing" "Others"	"Discussion" "Updates" "Idea generation" "Information sharing" "Problem solving and advice" "Social and praise" "Other"	"Discussion and opinion" "Status update" "Event notification" "Problem solving" "Idea generation" "Provide input" "Work coordination" "Meeting coordination" "Information storage" "Social praise" "Informal talk" "Other"	"Information sharing" "Discussion" "Problem solving and advice" "Idea generation" "People" "Social and Praise" "Update" "Coordination" "Other"	"Consumptive use" "Contributive use" "Hedonic use" "Social use"	"Professional" "Non-professional"	"Factual information" "Self-statement" "Relationship-indicator" "Appeal"
Number of messages analyzed	100	300	886	1,196	1,809	N/A (results from previous case studies applied)	7,409 (across 2,978 threads)	(Undisclosed)	8,142	6,306
Organization	IBM	(Undisclosed)	(Undisclosed)	Capgemini	Deloitte	(Multiple)	Deloitte	N/A	(Undisclosed)	(Undisclosed)
ESN platform	Beehive	Yammer	Yammer	Yammer	Yammer	(Multiple)	Yammer	Online discussion forum	Yammer	(Undisclosed)

Table 2.1 – Overview of previous genre analysis studies in ESN contexts

2.2.3 Social network analysis of ESNs

Compared to the body of research that exists for genre analysis of ESNs as described in the previous section, there seems to have been markedly less academic interest with respect to applying SNA to analyze and investigate ESNs. The use of SNA metrics and methods can however give profound insights into the structure of such social networks (as described in sub-chapter 2.1), which can enable a researcher to make inferences about the characteristics of informal networks that exist within an organization. It should be noted that the literature research undertaken to explore SNA within the context of ESNs in this project has been guided by certain restrictions; not all literature that applies SNA in an organizational context has been subjected to further investigation, but only literature that deals specifically with applying SNA on ESN platforms. Research that has derived and analyzed social network data from non-ESN platforms such as workflow management systems (e.g., Kazienko, Michalski, & Palus, 2011; van der Aalst, Song, & Aalst, 2004) has not been included in this literature review. The reasoning behind this decision lies in the nature of the research questions stated in this project. In order to make inferences about the social structures that exist within informal networks in organizations, it is an important presumption that the activity in these networks are initiated and intended by the users themselves, and that the activity resides within a system that has as its purpose to facilitate such social interactions and collaboration (with emphasis on the term "social"). Hence, extracting and analyzing data from e.g. e-mail correspondence, CRM systems, ERP systems, or any other form of platform that does not fit into the previously stated definition of an ESN, has been excluded.

Lin, Ehrlich, Griffiths-Fisher, and Desforges (2008) proposed an expert-locator systems called SmallBlue, which was designed to help employees discover their social network within a company in order to locate communities or individuals who might be of professional value to a given user. By analyzing different types of communication data from the computers of employees, SmallBlue applied a combination of data mining and social network analysis in order to provide search mechanisms available to users. One such example was through the SmallBlue Find-tool, where employees were able to specify a set of search terms that resulted in a list of relevant connections being returned to the user. In addition to a picture and some general information about the connections returned, SmallBlue also provided information about the shortest paths that existed between the seeker and the relevant connection, and which intermediary connections were present in the case that the seeker and the relevant connection was not directly connected. The software also offered to visualize the social network of a particular group or a set of connections based on the search result, which allowed a user to visually identify and explore his/her informal network within the organization. SmallBlue itself does not fit the given definition of an ESN, as it did not provide the communicative tools that can be found within typical social media platforms aimed at organizations such as Yammer. Neither was it an analysis tool specifically aimed at analyzing ESN platform, as the main data collection was done using e-mail communication. However, it does provide an example of how SNA can be applied in an organizational context with the purpose of localizing and visualizing the informal knowledge networks of employees. A survey conducted by Ehrlich et al. (2007) indicated a high level of user

satisfaction among the employees who had tried the system over a period of time, which provides further testimony to the usefulness of SNA in an organizational context.

Smith, Hansen, and Gleave (2009) put focus on the increased adoption of social media platforms among businesses and corporations, and how this development opened a new and interesting data repository for studying the informal networks that existed within a professional environment. This paper is especially relevant for the research at hand as they compiled a set of social roles within a social network and specifically related certain SNA metrics such as in-degree and out-degree to these social roles. In their paper the set of social roles comprised six categories; answer person, question person, discussion or comment person, originator, and influencer. A combination of 14 social metrics was then used to describe the different social roles, ten of which were commonly applied SNA metrics such as eigenvector centrality and betweenness centrality. The remaining four metrics were based on other characteristics such as the number of days active and the number of messages started. They also categorized eight different types of social networks based on three separate characteristics related to how the relationships between the actors are defined. These three characteristics related to whether the network was direct or indirect (implicit or explicit ties), symmetric or asymmetric (directed or undirected ties), and weighted or unweighted (dichotomized or weighted ties). Their paper did not however provide further characterizations of networks with respect to SNA metrics such as density or clustering coefficient, as was done with the six different categories of users. In their conclusion they stress that although social network data can be a valuable source of information, the interactions among users using a social media tool does not necessarily translate directly into actual social relationships. They do however point out that the analysis of enterprise social media platforms can provide value and insight to decision-makers when it comes to mapping the internal communication patterns among employees, and can be used as an asset to measure organizational health and for human resource management.

Cao, Gao, Li, and Friedman (2013) studied six months of activity data from the ESN platform Jive used by a large, undisclosed, international corporation. The main objective stated in their paper related to the association between the hierarchical (formal) positioning and the geographical location of employees and their interaction patterns. By collecting a series of meta-data about each employee in the network from the Jive ESN platform, and the interactions between them, they were able to map the informal structure of the corporation. The geographical location and other human resource-related data about each employee were retrieved from the internal system of the corporation. The interaction graph, i.e. the social structure based on the communication between the users, was constructed in the form of a directed, weighted graph, where the weights were calculated based on the accumulated sum of interactions between the employees. To study the association between the users' connectedness and their hierarchical position, they conducted a correlation analysis between the users' degree centrality and their hierarchical position. The results implied that the hierarchical position (measured by the number of hops between a given user and the top level of the corporation) had a significant positive correlation with the users' in-degree centrality, but not with the users' out-degree centrality. Furthermore, in order to measure the effect of geographical location and hierarchical position on the interaction patterns of the

users, they defined a logistic regression model using exponential random graph models. Based on the results, the study concluded that both the hierarchical position and the geographical location of a pair of users can be used to predict the probability of the same two users interacting with each other. That is, the study showed that if a pair of users is from the same country, they are significantly more likely to interact with each other than two users from different countries. The same conclusion was made about a pair of users that are closer to each other with respect to hierarchical distance. The study by Cao et al. (2013) is, as far as the literature review in this project has been able to discover, the only study that has applied statistical methods appropriate for social network data in order to investigate the relationship between user attributes and the structural characteristics of users in ESNs. However, their study only used a selection of SNA metrics (degree centrality and betweenness centrality) to describe what they called "*influential users*", although other centrality metrics are also available and valid for this purpose.

Berger et al. (2014) aimed to describe the structural characteristics of what they called "*value-adding key users*", which was referred to as users who could be characterized as important for the network in some respect, depending on the metric applied. In their particular study they defined key users as users who had received the most "likes" and bookmarks in their messages. Their dataset consisted of ~2 years of recorded interaction activity in a Yammer dataset, provided by an undisclosed organization. In addition to a qualitative text analysis (the same study has been discussed in section 2.2.2), Berger et al. (2014) calculated a set of different centrality metrics in order to describe the structural characteristics of these key users, which included; betweenness centrality, closeness centrality, in-degree centrality, and eigenvector centrality. Furthermore, the social networks constructed in their paper consisted of two distinct graphs; the social graph and the activity graph, whereof the former was based on the following-relationship between users, and the latter was based on the communication activity between users. To investigate the structural characteristics of key users, they performed an overlap analysis between key users and central users in both the social graph and the activity graph. Their results showed a generally large overlap between key users and central users in both the social graph and the activity graph. E.g., the overlap between the top 1% key users and the top 5% users according to in-degree centrality in the social graph was 93%, i.e. 93% of the top 1% key users were also present among the top 5% of users according to in-degree centrality. In the activity graph, the largest overlap could be found between the top 1% of key users and the top 5% of users according to betweenness centrality. The study conducted by Berger et al. (2014) is interesting for the research at hand in the respect that it specifically applies a range of centrality measures in order to investigate the structural characteristics of users in an ESN. It is also, as far as the literature review in this project has been able to identify, the only study that has conducted a genre analysis (see reference to same study in section 2.2.2) and a social network analysis of an ESN. One limitation from the perspective of this project however, is that Berger et al. (2014) does not make a connection between the genre analysis and the social network analysis. Furthermore, it is not clearly communicated how the centrality metrics in the study have been calculated. As the different centrality metrics (as shown in section 2.1.3) are calculated differently depending on

whether tie weights are taken into account or not, it was unclear to what extent this was accounted for in their study.

In this section, a summary of existing research on the topic of applying SNA in the context of ESNs has been presented. Surely, the overview provided should not be regarded as an exhaustive list of research relating to SNA in ESNs, but rather as an excerpt of how SNA methods and metrics can be applied in order to gain insights into the informal networks of an organization through ESN platforms. Table 2.2 shows an overview of the literature research conducted in this sub-chapter.

Publication	Lin et al. (2008)	Smith et al. (2009)	Cao et al. (2013)	Berger et al. (2014)
Objective(s) of the study	Locating experts within large enterprises Enhance users' situational awareness in the enterprise	Create a typology framework for characterizing different types of networks based on enterprise social media data Show how ESN data can be used to describe users' social roles within the network	Explore the relationship between users' social position and their organizational position	Investigate the structural characteristics of "Key users" (users who have received the most "Likes" and Bookmarks in the network)
Type of data used for creating network	E-mail Instant messaging	N/A (the typology presented allows for multiple types of data to be used depending on what data is available)	Document sharing Discussions Blog posts Micro blogs	Messages "Following" data (a Yammer functionality that allows users to subscribe to and follow the activity of other users)
Unit(s) of analysis	Users	Users Network	Users	Users
SNA metrics applied in the study (what they measure)	Degree centrality (well-connected users) Betweenness centrality (bridges in the network)	In-/out-degree centrality Betweenness centrality Closeness centrality Eigenvector centrality Clustering coefficient (several different social roles are linked to each of these metrics)	In-degree centrality (popular users) Out-degree centrality (active users) Betweenness centrality (influential users)	In-degree centrality Betweenness centrality Closeness centrality Eigenvector centrality (all above-mentioned metrics applied to detect central users)
Size of network	>150,000 users	N/A (no case study presented)	7,400 users	10,434
Organization	IBM	N/A (no case study presented)	(Undisclosed)	(Undisclosed)
ESN platform/Application	SmallBlue	N/A (no case study presented)	Jive	Yammer

Table 2.2 - Overview of previous SNA studies in ESN contexts

Chapter 3

Research method

This chapter will describe the specific steps conducted during this research project as outlined in Figure 3.1. The first sub-chapter will describe how the data used in this research was collected, the structure of the data, the filtering that was conducted prior to proceeding with further analysis, and the additional attributes calculated for the dataset. The second cub-chapter will give a brief introduction to the tools that have been applied in order to evaluate and analyze the data during different stages of the project. In the third sub-chapter, a detailed overview of the method applied in conducting the genre analysis of the data will be given, including the sampling method used, the selection of a genre repertoire, and the inter-rater reliability analysis conducted in order to evaluate the level of agreement between the separate coders involved in the genre analysis. The fourth sub-chapter will move on to the SNA part of the analysis, where an overview will be given of the SNA metrics applied in describing the networks and their actors, and the statistical evaluation undertaken in order to decide on a network creation algorithm for use in this project. The fifth and final sub-chapter will give an overview of the statistical methods applied in inter-correlating the results from the SNA and the genre analysis.

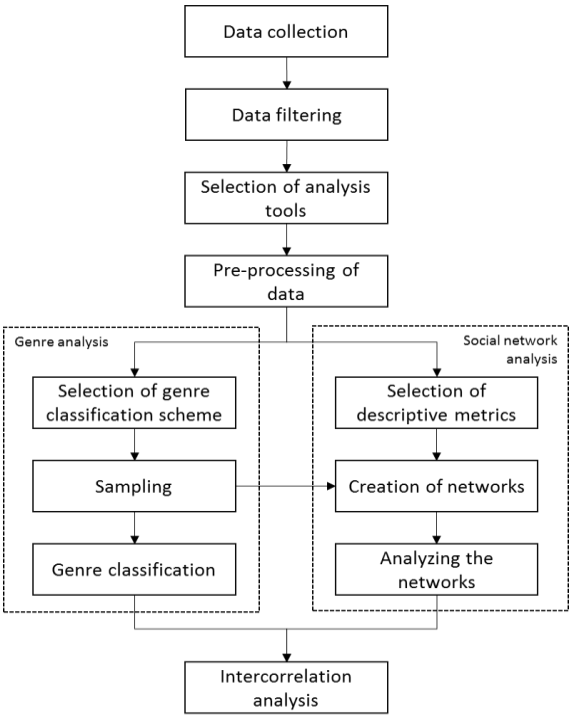

Figure 3.1 - Research method outline

3.1 Data collection

3.1.1 Method

In order to explore the characteristics of ESNs with regards to the structure of the networks over time, the users within the networks, and the message content exchanged between these users, an ESN dataset was collected from a large consultancy firm's (>100,000 employees) use of the ESN Yammer. The company has presence in over 100 countries, and offers consultancy services within finance, tax, risk management, and information technology.

Due to the raw ESN data being regarded as company sensitive information, the use of the data was subject to a non-disclosure agreement (NDA). Thus, in order to protect the anonymity of the company as specified in the agreement, the company will henceforth be referred to as "ABX". Whenever necessitated, meta-data about the

objects contained within the dataset, such as the names of groups, have also been altered accordingly in order to avoid that the identity of the company can be deducted from these pieces of metadata.

3.1.2 About the dataset

Descriptive statistics

The data was extracted from ABX's internal use of the Yammer ESN service over a time period ranging from 15th October 2010 to 30th October 2014. In total, the dataset contained 10,868 messages exchanged between 775 different users within 7 different groups. Table 3.1 shows a summary of descriptive statistics related to the dataset.

Metric	Count
Total number of messages	10868
Number of groups	7
Number of messages in most active group	3377
Number of messages in least active group	254
Total number of unique users	775
Number of users with >1 sent messages	545
Number of users with >10 sent messages	167
Number of users with >100 sent messages	17
Most posts by one user	760
Average number of posts per user	14.02
Number of threads	2597
Most posts in one thread	164
Average number of messages per thread	4.19

Table 3.1 - Descriptive statistics

As can be read from Table 3.1, there is a notably large deviation between the most and least active group in terms of message count, where the difference between the two is 3123 messages. The same applies for the difference between the users in terms of message count, where only 17 out of 775 users have sent more than 100 messages, accounting for 2.19% of the total user base. Furthermore, only 545 out of 775 users have sent more than 1 message, leaving 29.68% of the user base as "idle users" without any post entries. Further details on the distribution of message activity between the users are depicted in Figure 3.2.

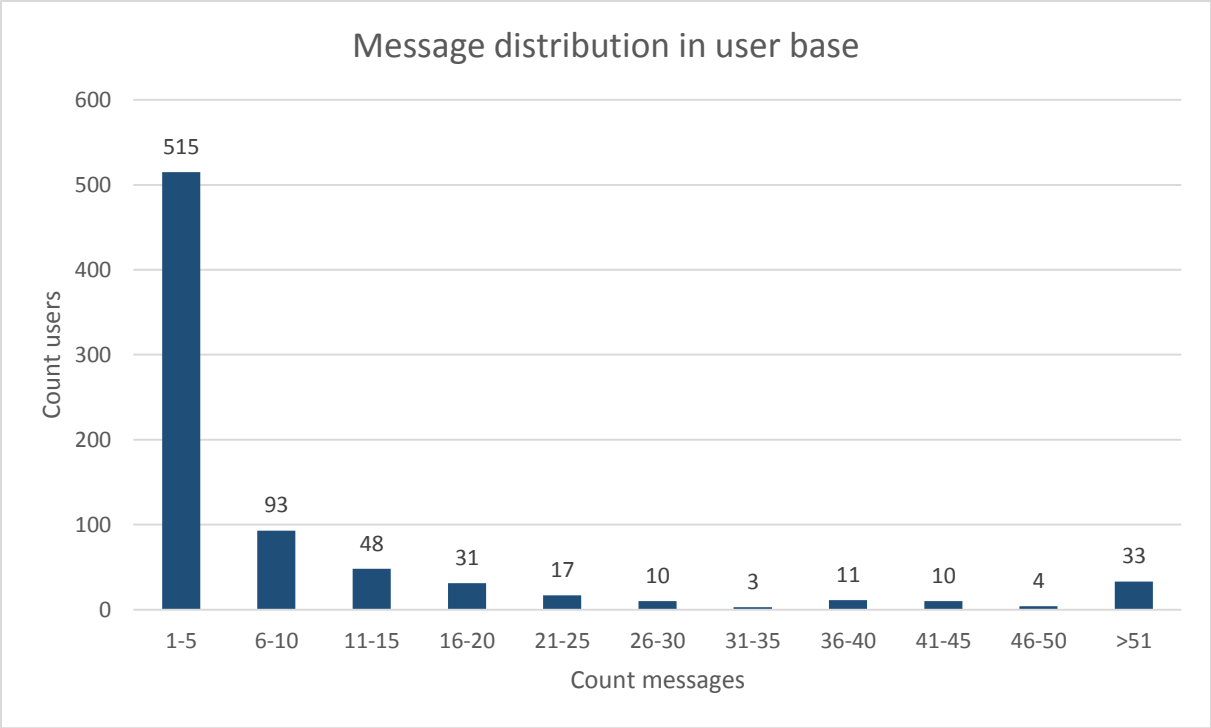

Figure 3.2 – Count of users grouped by message count

Further inspection of the data showed that the top 78 users with regards to message count accounted for a total of 7720 messages, corresponding to 10% of the most active users accounting for 71.03% of the content in the dataset. However, these numbers are not rear within online communities. Previous studies have shown that a minority of users in such communities normally contribute a disproportionately large amount of content, and that the majority of users remain inactive to a large extent. Nielsen (2013) showed that 90 percent of users in most online communities are "*lurkers who never contribute*", while the majority of the activity is sustained by just 1 percent of the total user base. However, the proportions between lurkers and active contributors have been shown to vary greatly depending on the type of community at hand and what type of information is being exchanged within the community (Nonnecke & Preece, 2000).

The content within the dataset was partitioned into a total of 7 user groups. Table 3.2 depicts the differences within each group with regards to message count, duration, and the average activity per user and per time period. It should be noted that, since users in Yammer can join and leave groups over time which entails that the number of users in a group is not a static value, the user count in Table 3.2 represents the total number of users that have been active within the group at any point during its lifecycle. Furthermore, the Yammer dataset received from ABX did not offer the data necessary to detect users who join a group but who remains idle, i.e. does not post any messages in the group. Hence, the method applied in order to calculate the number of users relied solely on which users were visible within the group by either posting or receiving a message, as this data was available by investigating the metadata for each message. However, as idle users do not actively contribute content within a group, which is a prerequisite for investigating the very

characteristics of users and the relationships between them, these (unaccounted for) users were not deemed to be a main concern in this research project.

Group name (anonymized)	Created at (date)	Message count	User count	Duration (days)	Average count of messages per day	Average count of messages per user
ABX1	2011-09-05	2421	178	1151	2,10	13,60
ABX2	2012-06-20	1177	126	862	1,37	9,34
ABX3	2010-10-15	1467	256	1477	0,99	5,73
ABX4	2011-04-13	1804	199	1296	1,39	9,07
ABX5	2012-10-07	3377	156	753	4,48	21,65
ABX6	2011-12-14	368	53	1052	0,35	6,94
ABX7	2014-02-28	254	85	245	1,04	2,99

Table 3.2 - Descriptive statistics groups

As the data depicted in Table 3.2 represents the accumulated values over time, it does not provide information about how the content in each group is distributed over time. With specific attention to the longitudinal analysis, this information was needed in order to make a qualified decision on whether the content was sufficient for such an analysis. The result is visualized in Figure 3.3.

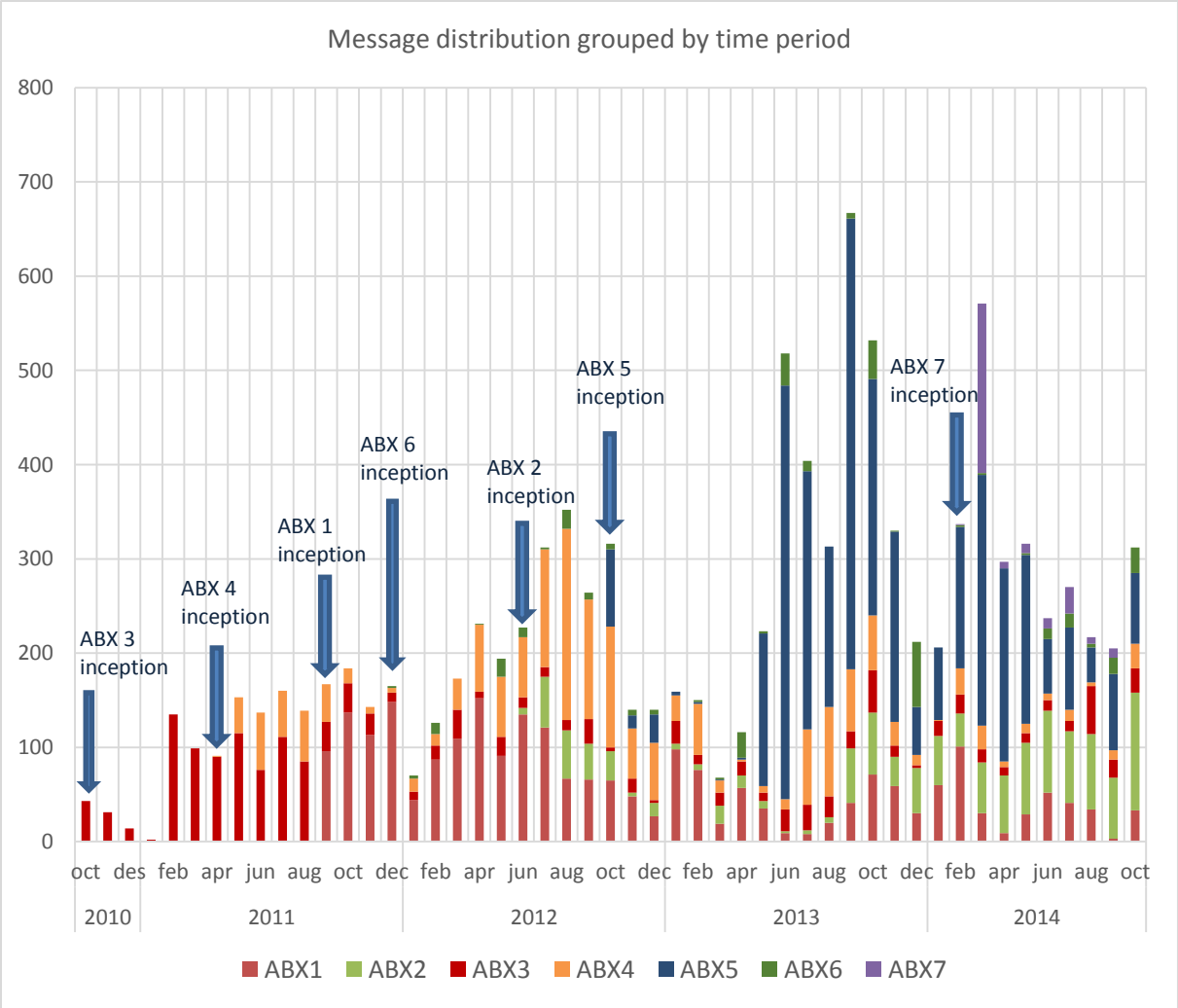

Figure 3.3 - Message count over time per group

It can be observed from Figure 3.3 how the activity level measured by the count of messages fluctuates throughout each year. The relative bottom can be found during January 2011, where only 2 messages were sent during the whole duration of the month. The relative peak of activity could be found during September 2013, where the total amount of activity amounted to 668 messages across the six different groups active at this point in time. The most active group, ABX5, contained a minimum of one message in March 2013, and a maximum of 478 messages in September 2013. The message activity in the least active group, ABX7, varies between one message during February 2014 and 180 messages during March 2014. As for the average activity per day factoring in how long the group has been in existence, ABX6 has the lowest score with an average of 0.35 messages per day since its inception in December 2011. On the other end of the spectrum, ABX5 has the highest average with 4.48 messages per day. Furthermore, while six of the seven groups do not have monthly time periods where the activity amounts to zero messages, ABX6 has a total of five monthly time periods where there is an absence of any activity; March 2012, January 2013, August 2013, January 2014, and April 2014.

By further studying the graph, it also became apparent that a trend of low activity could be observed in the months of December and January each year. A plausible explanation for this phenomenon is that this is an assigned summer holiday period in the company, therefore causing a higher number of employees being absent as opposed to other periods during the year. Another noticeable feature is that the intra-group activity level fluctuates from month to month, causing the overall activity to remain volatile throughout each year. As with other types of online social networks, fluctuations in the activity can have multiple reasons; 1) external factors such as company happenings, news, or alike that stimulates a higher level of discussion for a period of time, 2) influx of new users who contribute to the activity within a group, 3) work-related happenings such as new projects that increases the need for online collaboration between employees, 4) creation of new groups that contribute to the overall activity within the ESN. As data on these contributing factors were missing in this project, it was not possible to further investigate the reasons for the volatile activity within this ESN. Furthermore, while it would be interesting to map what sort of stimulus that is most contributory with respect to the activity level within the ESN, this was not within the scope of the research and was also not relevant in order to meet the stated objectives.

Yammer and its data structure

Yammer is an ESN that proclaims to be a "*private social network that helps employees collaborate across departments, locations, and business apps.*" (Yammer, 2014c). For users to register on Yammer they need a valid work e-mail address, meaning that the ESN is not accessible for users who are only in possession of a commercial e-mail address such as anyone@hotmail.com. Since Yammer was launched in 2008 it has grown to >8 million registered users across >200,000 companies worldwide (Yammer, 2014a), a growth further fueled by Yammer being acquired by Microsoft in 2012 for USD 1.2 billion (Lietdke, 2012).

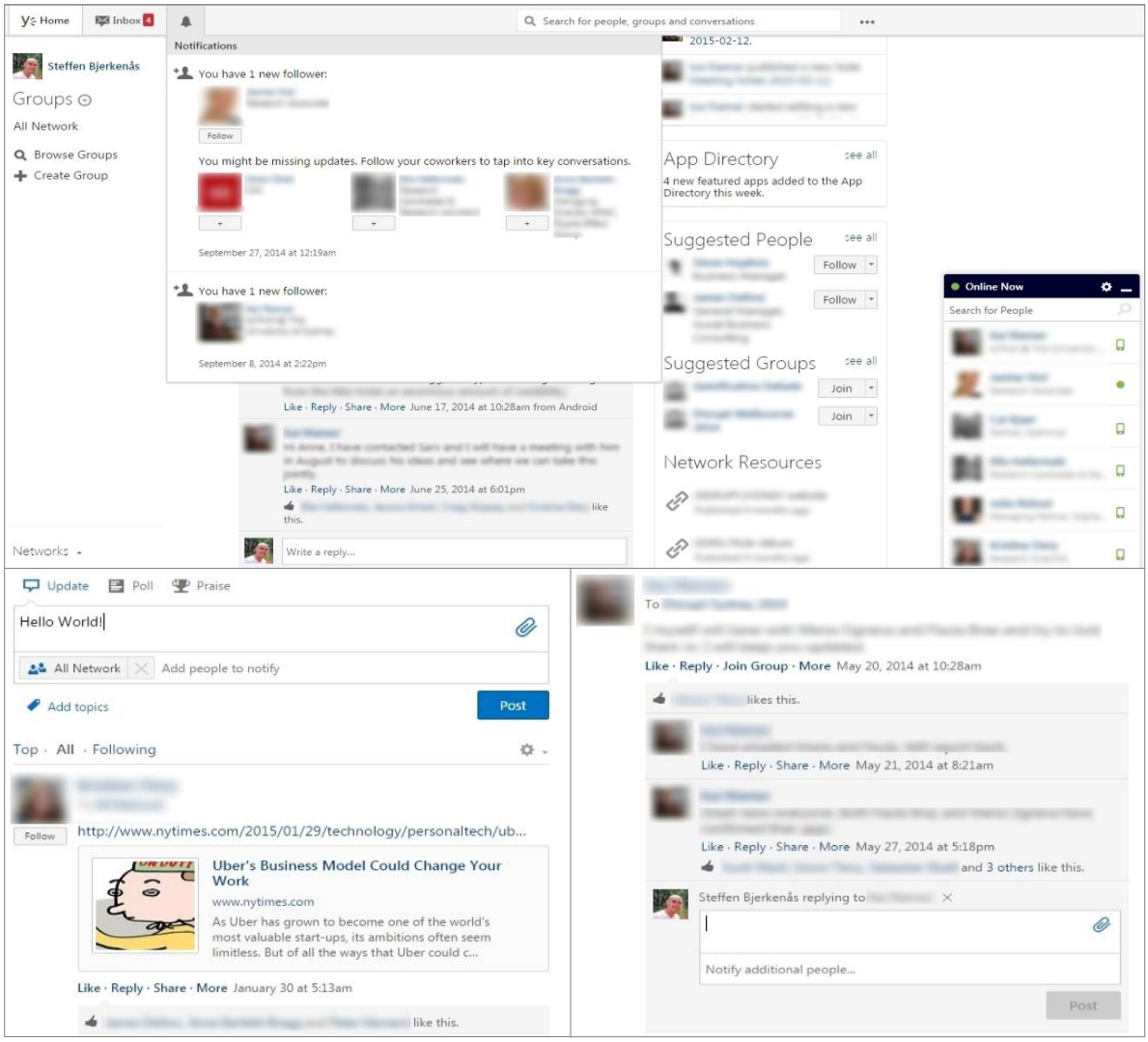

Figure 3.4 - Screenshot from Yammer. Top: The "Home" page in Yammer. Bottom left: Posting a message on the “home wall”. Bottom right: Posting a message within a thread

The user interface of Yammer bears similar graphical characteristics as can be found in other commonly used social networks such as Facebook, Twitter, and Google+. Yammer-users have the ability to create their own profiles with a profile picture, job title, location, and other meta-data relevant to their individual profiles. Other features in the Yammer platform include creating and joining groups, share documents and files between employees, initiate thread-based conversations, search for other users, and the opportunity to “like” and “follow” other users in the same fashion as can be found in Facebook and Twitter.

The Yammer platform allows an administrator to export the data contained within the platform in the form of .csv files (Yammer, 2014b) by running a script towards the Yammer application program interface (API). The separate .csv files contain details about users, groups, files, and messages, depending on how much information is exported. For the purposes of this research project, there were three separate sets of data that were needed, namely 1) details about the users such as ID, what time they joined and (if applicable) left the network, 2) details about the groups such as time of creation and the number of messages, and 3) details about the messages

themselves such as the ID of sender and receiver. Accordingly, in coordination with ABX, the data was then exported in three separate files containing these details. Table 3.3 depicts an exhaustive list of the attributes contained within each of the exported files as specified by the Yammer website for developers[7].

Dataset	Attribute	Description
Users	id	Numeric identifier for user
	job_title	User-defined job title
	location	User-defined location
	department	User-defined department
	api_url	API URL address to user
	deleted_by_id	If user deleted, it specifies the ID of which user performed the deletion
	deleted_by_type	Type of object responsible for the deletion, e.g. “User”
	joined_at	Date and time for when the user joined the network
	deleted_at	Date and time for when the user was deleted (if applicable)
	suspended_by_id	If user suspended, it specifies the ID of which user performed the suspension
	suspended_by_type	Type of object responsible for the suspension, e.g. “User”
	suspended_at	Date and time for when the user was suspended (if applicable)
	guid	Globally unique identifier for user
	state	State of user, e.g. “Active”, “soft-delete”
Messages	id	Numeric identifier for message
	replied_to_id	If message is a reply, to which message ID is it a reply to
	thread_id	Numeric identifier for which thread the message is posted in
	conversation_id	Numeric identifier for conversation (if applicable)
	group_id	Numeric identifier for which group the message belongs to
	group_name	Name of which group the message belongs to
	participants	Participants in conversation (if applicable)
	in_private_group	TRUE or FALSE value depending on if the message is posted in a private or public group
	in_private_conversation	TRUE or FALSE value depending on if the message is posted in a private or public conversation
	sender_id	Numeric identifier for which user sent the message
	sender_type	Type of sender (if applicable)
	sender_name	Name of sender
	sender_email	E-mail of sender
	body	The content of the message
	api_url	API URL address to message
	attachments	Attached files (if applicable)
	deleted_by_id	If message deleted, it specifies the ID of which user performed the deletion
	deleted_by_type	Type of object responsible for the deletion, e.g. “User”
	created_at	Date and time for when the message was posted
Groups	id	Numeric identifier for the group
	name	Name of the group
	description	Description of the group (if applicable)
	private	TRUE or FALSE value depending on if the group is public or private
	moderated	TRUE or FALSE value depending on if the content posted is being moderated
	api_url	API URL address to group
	created_by_id	Numeric identifier for which user created the group
	created_by_type	Type of object responsible for creating the group, e.g. “User”
	created_at	Date and time for when the group was created
	updated_at	Date and time for when the details about the group was updated (if applicable)
	deleted	TRUE or FALSE value depending on if the group has been deleted

Table 3.3 - Description of dataset attributes

[7] http://developer.yammer.com/v1.0/docs/data-export-api

As can be seen from Table 3.3, each message contains, if applicable, the message ID for which the original message is a reply to. This entails a hierarchical replied-to structure which allows for the analysis of the dataset by user-to-user connections, and it allows for the messages themselves to be grouped into threads.

3.1.3 Data filtering

To ensure that the dataset used for further analysis was suitable with respect to the stated objectives for this research, each group within the dataset was subjected to an evaluation based on certain criteria as listed beneath. The evaluation was conducted with basis in the characteristics listed in Table 3.2. As there was little basis for defining a cut-off value for which the evaluation criteria could be compared against, each group was compared to each other in order to make an informed decision on whether the groups could be deemed suitable for further analysis. The criteria used to evaluate each group included;

- Each group within the dataset must have a sufficient amount of posts and users to support a static analysis of the group individually. If the amount of users or messages is too low, then one runs the risk of not being able to conduct a quantitative analysis of the activity within the group and, in turn, make inferences based on the results of the analysis.
- Each group must have sustained activity over time in order to create snapshots of the group activity which are not absent of both users and posts. Groups who remain idle over a certain period of time, and only contribute content sporadically, would not satisfy a longitudinal analysis since the available timeslots for sampling would be too small.
- Each group must have been active for a sufficient amount of time in order to create snapshots of the activity with a certain amount of time in between them.

After comparing the groups too each other with respect to activity, count of users, and duration, ABX6 and ABX7 were removed from the dataset. ABX6 had been active for more than 1000 days, but had the lowest score in terms of messages per day compared to the other groups. In addition, the group has 5 monthly time periods in which there was no activity amongst the users and only 53 users had been active during the >1000 days of activity, the lowest score amongst the 7 groups. This indicated that the communication between users are not continuous over time, further implying that the group was less than an optimal candidate for further analysis. ABX7 was selected for removal based on its duration. The group had only been active for a total of 245 days (~8 months) by the time the dataset was collected, which was not deemed appropriate in order to sample time periods with sufficient time in between them. Out of these 8 months, there were 4 monthly time periods that had less than 10 messages. The message activity in the group was also moderate/low compared to the other groups with an average of 1.04 messages per day.

3.1.4 Pre-processing the data

In order to enhance the data available for each object contained within the dataset, several scripts were written in R to calculate additional attributes. More specifically, one script was written to calculate additional attributes for users, one for messages, and one for the group dataset. The objectives behind pre-processing the data was 1) calculate additional attributes needed for creating the networks and for designing more efficient network creation algorithms, 2) expand the data available to conduct statistical analysis of the dataset. The additional attributes calculated for each of the three separate datasets are listed in Table 3.4.

Dataset	Attribute	Description
Users	sent_messages	Number of messages sent per user
	received_messages	Number of received messages per user
	count_groups	Number of groups each user has participated in, either by receiving or posting a message
	count_threads	Number of threads each user has participated in
	duration	The amount of time each user has been active
	sent_rec_ratio	The count of sent messages divided by the count of received messages per user
	start_threads	Number of threads each user has initiated
	response_rate	Number of posts for each user responded to by other users
Messages	receiver_id	The user ID of the receiver for each message. The dataset includes only the ID of the message replied to and not the user ID itself, which can be retrieved by doing a lookup-operation for each replied_to message and querying for the corresponding user ID
	count_replies	Numeric value depicting the number of replies ach message has received (if any)
	stand_alone	Boolean value indicating whether a message a message is neither a reply or is a reply itself
Groups	count_users	Numeric value depicting how many users have been active within each group
	count_messages	Numeric value depicting how many messages each group contains
	users	Array containing all user ID's contained within the group
	messages	Array containing all message ID's contained within the group

Table 3.4 - Additional attributes calculated

3.2 Selection of analysis tools

This sub-chapter will give a brief introduction to the different tools and software that were applied during this research project.

3.2.1 The R programming language

R is an open source programming environment aimed towards facilitating statistical analysis and graphics, and is freely available under the GNU General Public License (R Development Core Team, 2013). The tools that are offered within the environment comes in the form of *packages*, which is a collection of functions aimed at a specific purpose such as building graphs or text mining. These packages are compiled by contributors of the project and can be downloaded and installed by any user who might benefit from the functions contained within the package. Given the rising popularity of R amongst statisticians and alike over the last years (D. Smith, 2012;

Vance, 2009), the library of available packages now caters to a wide range of needs with respect to statistical analysis. As such, the R programming language made a strong candidate for the purposes of this research project. Table 3.5 shows a brief extract of simple operations written in R.

```
# create vector of numbers 1 through 5
> t<-c(1,2,3,4,5)
# print content of vector t
> print(t)
[1] 1 2 3 4 5
# multiply each value in vector by 2
> t*2
[1] 2 4 6 8 10
# multiply each value in vector t by 2 and store results in new vector p
> p<-t*2
# for every value in vector p, print «Hello» and the current value
> for(i in 1:length(p)){print(c(«Hello»,p[i]))}
[1] «Hello» «2»
[1] «Hello» «4»
[1] «Hello» «6»
[1] «Hello» «8»
[1] «Hello» «10»
```

Table 3.5 - Examples of R syntax

Several packages in R include functions aimed at social network analysis, each containing different sets of functions applied during different phases of the analysis. The selection of packages applied in this research project comprises *igraph* (Csárdi & Nepusz, 2006), *sna* (Butts, 2014), *plyr* (Wickham, 2011), *tnet* (Opsahl, 2009), and *network* (Butts, Handcock, & Hunter, 2014).

3.2.2 UCINET

UCINET is a software which allows for statistical analysis of social network data through a number of built-in tools (Borgatti et al., 2002). For the purposes of this project the software was especially well-equipped to perform statistical tests on the dataset such as correlation analysis, as the tools available in the software are specifically designed to handle network data. Some of the statistical tools in UCINET can also be found in the different packages available in R, but the graphical user interface in UCINET offers a more efficient way of manipulating and visualizing data through NetDraw.

3.2.3 Gephi

Gephi is another software package aimed towards exploring social network data (Bastian, Heymann, & Jacomy, 2009). The software is plug-in-based and open-source, i.e. contributors are welcome to write their own plug-ins which is made available for all users of Gephi through their Gephi MarketPlace website[8]. The software differs from UCINET in the way that Gephi has a more limited selection of statistical tools when it comes to network data, but rather a large range of tools aimed at graphical manipulation, incorporating several ways of visually representing

[8] http://marketplace.gephi.org

graphs and networks. All figures containing graphs in this research project was produced using this software.

3.3 Genre analysis

3.3.1 Data sampling

Since the Yammer dataset consists of a total of 10,868 messages, beyond what was feasible to manually analyze during the research project, the messages were sampled in order to conduct the qualitative genre analysis. During this sampling process, the threads were used as the unit of sampling instead of the messages. This approach ensures that the context of each post, i.e. the thread, is captured in addition to the post itself (K. Riemer, Overfeld, et al., 2012), which helps ensure that a post is correctly categorized within a given genre repertoire.

Initially, the three first months of activity plus two additional random three-month periods per group were selected as basis for the sampling. A total of 1296 posts were sampled, spread across 159 threads, all 5 groups and 253 unique users. The average duration of the sampled threads were 9.16 days, with an average of 4.28 participants and 8.15 posts in each thread. The mean number of characters per message was 201.5. After an initial investigation of the preliminary results, it was decided that each of the three month time periods were to be divided into two separate ~6 week periods each, resulting in a total of six time periods per group instead of three. Table 3.6 shows an overview of the sampled time periods, each of which represents a snapshot of the network. Several previous studies have used this sampling method in order to longitudinally analyze social network data (e.g., Kossinets & Watts, 2006; Panzarasa et al., 2009).

The reasoning behind the post-processing of the sampling was the need for additional time periods in order to increase the number of sampled snapshots and, in turn, provide a more reliable statistical analysis. Another consideration that was taken into account was that taking snapshots of a network with a too large timespan might contribute to over-smoothing the activity in the snapshots, i.e. short periods with markedly low or high activity will become less visible (Kossinets & Watts, 2006). At the same time, it was considered important to not choose a too small sampling window since this could result in each sample containing an insufficient number of posts with respect to the genre analysis. By splitting up the initial three-month sample periods into two ~6-week sample periods each, the average number of posts per period remained sufficient (43.2) while at the same time increasing the amount of time periods available for further analysis.

Group	Time period	Start	End	Duration (weeks)	Mean thread duration (days)	Total count of posts
ABX1	1	05.09.2011	13.10.2011	5.43	2.00	44
	2	14.10.2011	30.11.2011	6.71	2.67	50
	3	01.03.2013	14.04.2013	6.29	14.67	37
	4	15.04.2013	31.05.2013	6.57	3.25	41
	5	01.09.2013	15.10.2013	6.29	3.50	45
	6	16.10.2013	30.11.2013	6.43	21.33	55
ABX2	1	20.06.2012	26.07.2012	5.14	3.29	46
	2	27.07.2012	31.08.2012	5.00	12.00	28
	3	01.03.2013	15.04.2013	6.43	3.00	22
	4	16.04.2013	31.05.2013	6.43	1.00	18
	5	01.06.2014	15.07.2014	6.29	29.80	59
	6	16.07.2014	31.08.2014	6.57	24.20	76
ABX3	1	15.10.2010	25.11.2010	5.86	3.33	53
	2	26.11.2010	31.12.2010	5.00	4.67	12
	3	01.04.2011	15.05.2011	6.29	13.45	82
	4	16.05.2011	30.06.2011	6.43	1.50	28
	5	01.07.2012	15.08.2012	6.43	8.00	16
	6	16.08.2012	30.09.2012	6.43	1.33	20
ABX4	1	13.04.2011	08.06.2011	8.00	12.13	42
	2	09.06.2011	30.06.2011	3.00	18.00	28
	3	01.07.2012	15.08.2012	6.43	5.89	77
	4	16.08.2012	30.09.2012	6.43	5.20	50
	5	01.01.2013	15.02.2013	6.43	18.00	27
	6	16.02.2013	31.03.2013	6.14	19.70	35
ABX5	1	07.10.2012	14.11.2012	5.43	10.33	56
	2	15.11.2012	31.12.2012	6.57	6.00	31
	3	01.07.2013	15.08.2013	6.43	3.00	40
	4	16.08.2013	30.09.2013	6.43	6.17	98
	5	01.07.2014	15.08.2014	6.43	15.50	51
	6	16.08.2014	30.09.2014	6.43	11.50	29
Total				**184,14**		**1296**

Table 3.6 - Sampled time periods

The variations in duration for each time period in Table 3.6 can be explained by the process of splitting up each 3-month period into two ~6 week periods. This process resulted in that some posts belonging to the same thread became dispersed over two separate time periods, which inhibited the full context of each thread from being captured. As previously stated in this section, this was deemed important in order to limit the risk of erroneously categorizing posts. Thus, the start- and end-date for those periods were shifted in either direction to ensure that no threads became separated within. The genre analysis conducted in this thesis is unaffected by the duration of each respective time period, hence this action was deemed appropriate.

Table 3.6 also depicts a variation within mean thread duration for each time period. This can be explained by how the duration of a given thread is calculated, which is calculated as the total

duration between the first and the last post in a thread. Thus, if a thread had 25 posts in a single day and then only one more post three weeks later, the duration was calculated to 21 days. The Yammer platform does not have the functionality of closing a thread permanently, which means that this sort of data is not available in the dataset. The duration was therefore calculated as described above.

3.3.2 Selection of genre repertoire

The genres used in the content analysis were developed *a priori*, i.e. the set of genres were set prior to the actual genre coding of the posts took place. This way of conducting a genre analysis is often referred to as deductive coding, as opposed to inductive coding which involves defining the genres based on the content that is being analyzed (Stemler, 2001; Weber, 1990). The *a priori* coding was done in collaboration with a fellow researcher and the different steps were followed as outlined in Mayring (2000) (see Figure 3.5).

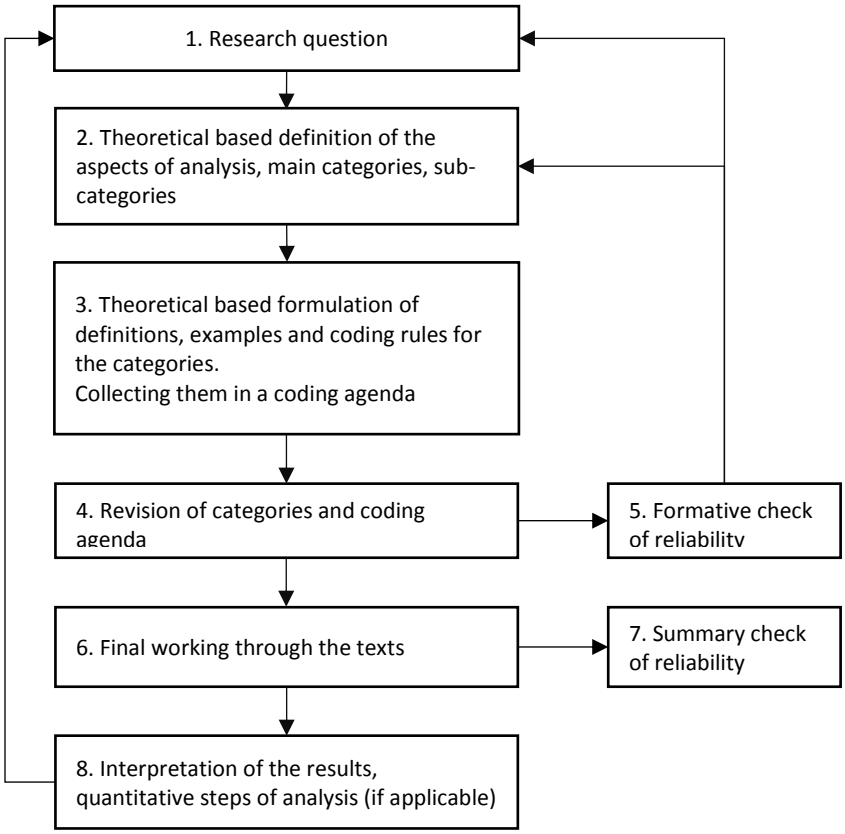

Figure 3.5 - Outlined steps for deductive genre analysis (Mayring, 2000)

The first step involves reviewing the research question and determining what genres are appropriate for use in the study. This step was conducted in collaboration with a fellow researcher by reviewing previous research and determining what previously used coding schemes would be suitable within the context of the research at hand. A main consideration was the environment in which the ESN had been used, i.e. the company and the nature of its business. Seeing that ABX is active within consulting, it was asserted that the communication between

employees would be professional in nature and therefore a coding scheme that was able to capture the nuances of communication that exists within such environments was needed. The assertions made during this step provided guidance for further literature study which can be found in the Theoretical Background-chapter.

Step 2 is where the results based on the literature study were formulated into broad categories which were deemed suitable for this research. The different coding schemes were evaluated and the appropriate main categories were extracted and compiled into a preliminary repository of genres for conducting the genre analysis. This repository mainly consisted of the union of previously applied genre schemes within the context of ESN, and provided the basis for further refinement. An overview of the studies that together constituted this repository can be found in Table 2.1 in the Theoretical Background-chapter.

During step 3 the main categories selected were further evaluated and the definitions of the categories were determined based on the existing literature depicted in Table 2.1. The result was a genre scheme comprising nine different genres; Information input, Discussion, Idea generation, Problem solving, Notification, Status update, Social, Praise, and Other. Table 3.7 provides a list of the different genres, their definition, and specific examples of each genre extracted from the genre analysis carried out in step 6.

Genre	Definition	Examples from dataset
Information input	Sharing a link/quote/video/resource (non-social), specific facts of professional value.	"Fantastic collection of information graphics (…) Great for inspiration. <link>" "Infographic: A Gargantuan Map Of The Internet. 196 countries, 350,000 sites. 2,000,000 links. 1 giant picture. <link>"
Discussion	Exchanging opinions on a specific subject introduced in a thread. Differs from Idea generation and Problem solving in the way that discussions are more opinionated in nature and less reliant on facts.	"True, but I thought Windows RT was a big waste of time (…)" "I think we should be empowering our managers to drive workflow with support from partners and directors." "Gets my vote too! Though centrally administered a wiki format would foster greater collaboration and knowledge sharing."
Idea generation	Sharing an idea or inviting others to participate in an exchange of creative ideas. Such messages are characterized by being visionary in nature, forward-thinking, and resemble a brainstorming session to a large degree.	"(…) What is an awesome idea for an app that we could work together as a group to create?" "Also I have some NFC stickers lying around that are trivial to program, although I haven't thought up any use cases on what I could do with them. Any thoughts?" "Yeah I was thinking about this idea over the weekend, I think there should be some sort of fun convention to the naming scheme, I like pop culture villains, what else can we think of?"
Problem solving	Actively providing or asking advice on how to solve a specific (professional) problem introduced in the thread. Such as answering somebody's inquiry about how to	"Anyone know how to put a z axis on a bubble chart, so I can have a 3D representation of risks vs. impact and time scale? Seems such simple thing, but struggling to find a means to do it."

Genre	Definition	Examples from dataset
	share a specific type of info, use a specific software etc. Also includes referencing to experts who might be able to assist with the inquiry.	"(…) If you need some inspiration on how the interactive version might work, YouTube Hans Rosling's TED 2006 speech. A very powerful story told using exactly this type of visualisation." "Did it to me too. To fix, find a pdf file, right-click, open with, choose default program, select pdf exchange viewer. No idea why it's done this, but that's the fix."
Notification	Make other people aware of upcoming company events, presentations, general news etc. Also includes messages where the poster responds to a (non-social) invitation to acknowledge his /hers participation or absence.	"I am running a call next week to test the concept of [redacted], basically a champions network across the country that can bring in thoughts and observations (…)" "Our first research paper [redacted] launched yesterday (…) "Count me in please"
Status update	Updates on the current state of a project, research or any other ongoing process within the company.	"I'm gonna be AWOL pretty much for the duration of my current project. Happy for anyone to pick up any of my tasks." "Here's where 15 minutes of coding got me last night - not at all how I want the end product to look but a decent start"
Social	All messages containing non-professional content but that serves a social purpose, such as invitations to concerts, after work-drinks etc.	"That would be great [redacted] lets grab a coffee" "Seem to be a large number of Samuel Adams varieties on here :) "
Praise	Acknowledging someone's contribution inside or outside the community. Also includes short statements expressing gratitude.	"Nice whitepaper mate!" "Wow. Thanks for the link - really interesting how various data can be visually represented."
Other	Any residual messages that does not fall into any of the other categories. Typically short messages that are either ambiguously formulated, include unknown jargon, or contain unreadable symbols or text.	"#NAME?" "Dobber." "Jealous..."

Table 3.7 - Genre repertoire

Step 4 and 5 included a preliminary analysis, in close collaboration with a fellow researcher, of the posts in the dataset by applying the genre repertoire in Table 3.7. The results showed that all the different genre categories were present, and hence, none of the categories were dropped before proceeding with the genre analysis. Furthermore, the count of messages within each category were revised in order to evaluate whether some of the genre categories could be merged into a common category. However, since each category had been developed *a priori* based on existing literature, no categories were merged since this would compromise the definitions by which each message had been categorized. For example, the categories Social, Praise, and Other all contained a relatively low frequency of posts, but since these categories each were tied to different definitions in the literature for which the repertoire was based on, they were not merged.

Step 6 was carried out by a final analysis of all the posts in the dataset. It is important to note that each message did not necessarily fall into only one category, but could be categorized using two or more categories depending on the content in the post. For example, a post that showed gratitude towards someone's contribution but also was part of an ongoing discussion would be categorized as both Praise and Discussion;

"Great points [redacted names]! (...) But also, unemployment occurs when people are without work and are actively SEEKING work. In a world of abundance, when technology can avail the wants and needs of every individual, would 'seeking work' be overrated?"

Another example would be posts that included both links to specific resources while participating in an Idea generation-thread, such as;

"This may seem a bit "phonist", but how about an app that makes use of NFC? http://developer.android.com/guide/topics/nfc/nfc.html. Automagical information exchange without having to have anything turned on!"

On average, 1.24 genres were assigned to each message during the genre analysis. This result is comparable with previous genre analysis studies carried out in similar contexts, e.g. Richter and Riemer (2013) who did a comparison study of previously conducted genre analysis of ESNs that showed that the average number of assigned genres per message across their five different case studies was 1.27 genres per post.

3.3.3 Inter-rater reliability

After all the 1296 sampled messages had been categorized, the results were revised by a second researcher which gave feedback on which posts needed to be re-categorized, and to/from which category this correction should be done. The initial genre analysis from researcher 1 was recorded along with the feedback from researcher 2, which allowed an investigation of the discrepancies between the two different versions. The initial analysis by researcher 1 included 1497 genres distributed among the 1296 posts (1.16 genre categories per post on average), while the final, revised version from researcher 2 included 1533 genres (1.18 genres per post on average). The union of the two versions were 1610 genre categories (1.24 genres per post on average), of which 1427 were identically coded by researcher 1 and researcher 2.

Based on the outcome of the genre analysis in step 6, Cohen's kappa (Cohen, 1960) was used to measure the inter-rater agreement between researcher 1 and researcher 2. Cohen's kappa differs from other inter-coder agreement measurements such as percentage-wise agreement in the way that Cohen's kappa takes into account the probability of two coders agreeing by chance. The formula used to calculate Cohen's kappa is defined as (Cohen, 1960);

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (3.1)$$

where Pr(a) is the percentage of agreement between the raters, and Pr(e) is the hypothetical probability that both raters select the same category by chance. Table 3.8 shows the results from each of the raters.

Genre	Rater 1 total	Rater 2 total	Agreement	By chance
Information Input	221	229	205	35.51
Discussion	668	701	663	328.52
Problem Solving	215	198	194	29.87
Idea Generation	73	71	70	3.64
Status Update	68	69	62	3.29
Notification	76	99	71	5.28
Praise	131	122	119	11.21
Social	9	8	8	0.05
Other	36	36	35	0.91
Total	**1497**	**1533**	**1427**	**418,26**

Table 3.8 - Cohen's kappa results

Pr(e) is calculated for each category by multiplying the relative distribution of the category for each rater, and multiplying the product with the total number of genres. As an example, the relative distribution of the genre Information input for rater 1 is 221/1497 = 0.148, and for rater 2 the relative distribution is 229/1533 = 0.149. The chance of the two raters agreeing by chance is therefore 0.148*0.149*1610 = 35.51, where the number 1610 refers to the total number of genres coded between the two raters. By summing up for all the categories, the result is the number of genres in the dataset that could have been assigned to the same category by both raters due to chance. Cohen's kappa can then be calculated as;

$$\frac{\frac{1427}{1610} - \frac{418,26}{1610}}{1 - \frac{418,26}{1610}} = 0,8464$$

The interpretation of Cohen's kappa differs in earlier studies. Landis and Koch (1977) presented a set of guidelines where Cohen's kappa-values >0.81 were considered "almost perfect", and Fleiss (1981) submitted that values >0.75 are to be considered "Excellent". Hence, based on the Cohen's kappa value of ~0.85 it was concluded that the result of the genre analysis was suitable for further use in this research project without further modifications.

3.4 Social network analysis

3.4.1 Creating the networks

Wasserman and Faust (1994) describes social networks as a set of nodes (i.e. users) and a set of ties (i.e. edges) that connect the nodes to each other. Furthermore, the edges connecting the nodes can either be directed or undirected, which is used to specify, if applicable, the *direction* of the connection between two nodes in the network.

The Yammer dataset contains several types of meta data that can be used to constitute ties between users in the network. In turn, the selection of which type to use dictates what algorithm is used to construct the networks. This section will discuss the different types of meta data available in the Yammer dataset, and based on this discussion, select a network creation algorithm for further use in this study.

Evaluating potential constituents for edges between actors

There are multiple ways of connecting the actors in the network depending on what type of network is subjected to analysis and what information is available on the communication data between users. This also applies to social network data, where the attributes used to constitute ties between users, may vary across different datasets extracted from different types of online communities. The Yammer dataset includes three distinct attributes which can be used to infer relationships between users; group affiliation, thread participation, and direct communication between users.

One way of constituting edges between actors is by looking at direct communication between users (Fisher, Smith, & Welser, 2006; Heidemann et al., 2010; Mislove, Koppula, Gummadi, Druschel, & Bhattacharjee, 2008; Shi, Zhu, Cai, & Zhang, 2009). In such cases, each interaction between user A and user B will result in a directed edge from user A to user B.

Berger, Klier, Klier, and Richter (2014) uses a combination of attributes to constitute edges between users in the network, namely the group ID in which a message was sent, and the receiver ID of the message. In their paper, a directed edge between a sender and a receiver is created if the receiver ID is included in the message, and if not, a directed edge is created between the sender ID and all members of the group in which the message was sent. The same way of creating the edges can be applied using the thread ID-attribute, i.e. a directed edge is created between a sender and all other participants in the same thread.

Furthermore, one can implement the timestamps of each message in combination with the thread ID to create the edges. For example, since the metadata in the Yammer dataset allows for messages in each thread to be ordered chronologically, an edge can be created from a sender to *previous* or *subsequent* participants in the same thread. The former method was used to construct a network in Toral, Martínez-Torres, & Barrero (2010), where weighted and directed edges were

constructed from a poster to all previous posters in the same thread. Petrovčič, Vehovar, and Žiberna (2012) designed an algorithm to incorporate the time elapsed between posts in the same thread, where the weight of the edges were based on the time elapsed and the number of posts between two posts in the same thread.

Table 3.9 gives an overview of the different ways of constructing networks as discussed above, henceforth referred to as algorithms. This table is not an exhaustive list, but rather specific to the Yammer data set as other social network data sets might offer different types of meta data.

Algorithm #	Sender ID	Receiver ID	Thread ID	Group ID	Timestamp (dd.mm.YYYY hh:mm:ss)	Resulting edges
1	A	N/A	N/A	100	N/A	A → {all participants in group ID: 100}
2	A	B	N/A	N/A	N/A	A → B
3	A	N/A	1	N/A	N/A	A → {all participants in thread ID: 1}
4	A	N/A	1	N/A	15.10.2014 12:11:11	A → {all participants prior to timestamp in thread ID: 1}

Table 3.9 - Examples of network creation pseudo-algorithms

Selection of network creation algorithm

For further use in this project, algorithm 2 was selected to create the networks, i.e. ties were created only between users who communicated with each other directly. The reasoning behind selecting this algorithm was mainly based on a combination of the research questions stated in Chapter 1 and the way the message data was sampled as described in section 3.3.1. Since the unit of analysis used for the genre analysis is the individual messages, it makes sense to define the ties based on user-to-user interactions which is done using algorithm 2.

By using any of the other algorithms (one-to-many interactions) one would have to include a larger set of assumptions, since all of these algorithms assume that every message has more than one receiver. Hence, this way of constructing the networks rests on making interpretations that are not explicitly reflected in the dataset since the Yammer data structure only allows for a single user to be specified as the receiver in a message.

Algorithm 2 does however assume that the users do in fact reply to the specific message they are reacting to within a thread, and not just the thread starter by default. If the latter was the case, then the chosen algorithm would not necessarily reflect the actual ties between users, since it is reasonable to assume that users also communicate between themselves within a thread and not only towards the thread starter. To investigate to what extent this assumption held true, the data was further analyzed in the following manner; 1) threads that contained three or more posts (including the thread starter-post) were extracted, 2) these threads were then grouped by the number of posts they contained in total, 3) for each of these groups, it was calculated how many percentage of the replies within these threads that were not addressed to the original thread

starter, i.e. to another participant within the thread. Figure 3.7 shows the result from this analysis.

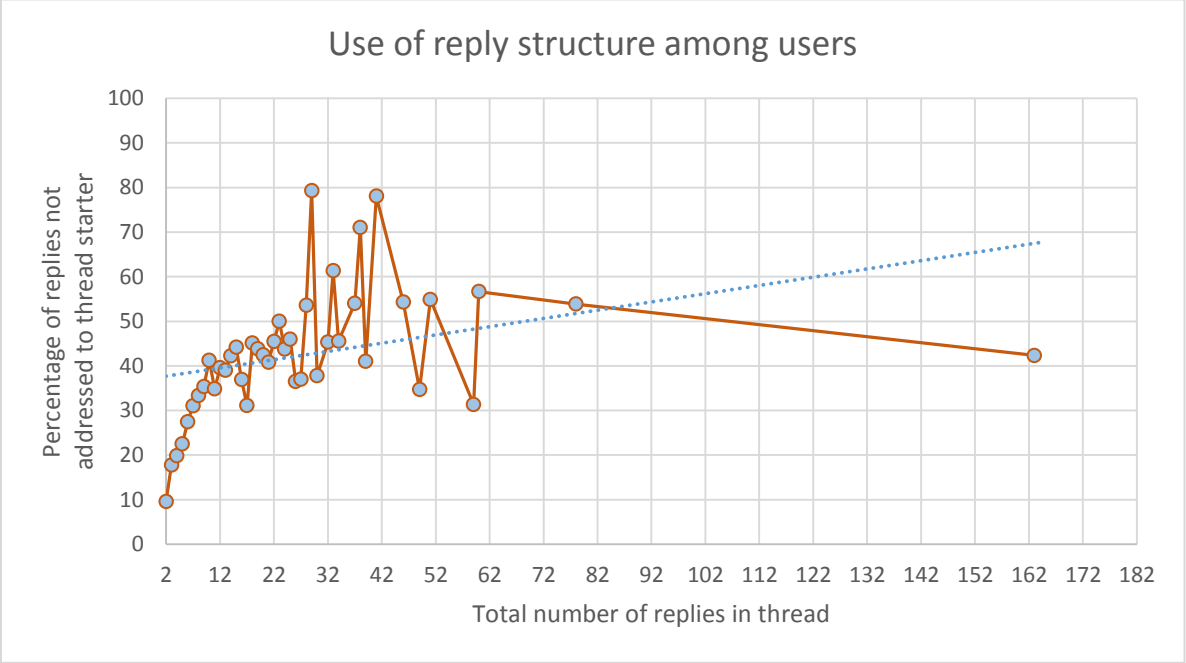

Figure 3.6 - Use of reply structure among users

For the whole dataset, the mean percentage of replies within threads that were not addressed to the thread starter was 42.62%. The remaining 57.38% of replies were posts that were addressed to the original thread starter. As the trendline in Figure 3.6 shows (blue dotted line), it appears that the more posts that a thread consisted of, the higher percentage of intra-thread conversations took place. This result implies that the users do make use of the proper reply-to functionality on the Yammer platform, and adds further weight to the choice of algorithm 2.

3.4.2 Analyzing the networks

Selection of network-level SNA metrics

In order to answer RQ1, there is first a need for defining which metrics to use in order to describe the structural characteristics of the network. The term "structural characteristics" is a broad term, as there are several metrics that can be applied in order to describe the structure of a network. However, RQ1 was formulated in such broad terms because of the lack of research on the relationship between network structure and conversational topics. For the same reason, it was determined that it would be necessary to employ an explorative approach by including several metrics in the analysis. Including all available network metrics would however be unfeasible. Hence, the selected metrics depicted in Table 3.10 were chosen based on the following considerations:

- The metrics should be widely accepted and applied in existing SNA-related research. The most commonly cited sources in existing SNA research were used as an indicator of their

relevance for this project, i.e. Hanneman and Riddle (2005), and Wasserman and Faust (1994).

- The metrics should be available in existing SNA software and tools (e.g., Bastian et al., 2009; Borgatti et al., 2002; Butts, 2014; Csárdi & Nepusz, 2006).
- As this study takes into account the weights of the ties between actors, the metrics must be able to accommodate weighted networks.

A thorough description of each metric is provided in chapter 2.

Metric	Description	Applied tool/software	R syntax (if applicable)
Network density	The ratio between the number of edges in the network and the number of potential edges in the network (Wasserman & Faust, 1994)	igraph (R package)	`graph.density(graph, loops=FALSE)`
Average weighted degree	The average sum of weights per node in the network (Barrat et al., 2004)	Gephi	N/A
Average clustering coefficient	The number of closed triplets in a network divided by the total number of triplets, generalized for weighted networks (Opsahl & Panzarasa, 2009)	tnet (R package)	`clustering_w(net, measure = "am")`
Average path length	The average path length from a node to all other nodes in the network, generalized for weighted networks (Opsahl et al., 2010)	tnet (R package)	`distance_w(net, directed=NULL, gconly=TRUE, subsample=1, seed=NULL)`
Reciprocity	The proportion of ties in the network that are reciprocated (Hanneman & Riddle, 2005)	igraph (R package)	`reciprocity(graph, ignore.loops = TRUE, mode = c("default"))`

Table 3.10 – Metrics applied to measure the structural characteristics of the network

Selection of actor-level SNA metrics

Table 3.11 shows the metrics that were applied to describe the centrality of the actors in the networks. As with the network-level metrics, the metrics in Table 3.11 were selected based on their common use in previous research (e.g., Berger, Klier, Klier, & Richter, 2014; Newman, 2001; Opsahl et al., 2010) as shown in Table 2.2, their availability in existing SNA software, and their ability to incorporate tie weights.

Metric	Description	Applied tool/software	Syntax (if applicable)
Weighted in-degree centrality	The number of incoming connections to a node, multiplied by their weights (Barrat et al., 2004)	Gephi	N/A
Weighted out-degree centrality	The number of out-going connections from a node, multiplied by their weights (Barrat et al., 2004)	Gephi	N/A
Betweenness centrality	How often a node appears on the shortest path between other nodes in the network, generalized for weighted networks (Brandes, 2001)	igraph (R package)	`betweenness(graph, v=V(graph), directed = TRUE, weights = NULL, nobigint = TRUE, normalized = FALSE)`
Closeness centrality	The sum of distances from a node to all other nodes in the network, generalized for weighted networks (Newman, 2001b)	igraph (R package)	`closeness(graph, vids=V(graph), mode = c("out", "in", "all", "total"), weights = NULL, normalized = FALSE)`
Eigenvector centrality	The centrality of a node measured by the connectedness by its neighboring nodes (Bonacich & Lloyd, 2001)	Gephi	N/A

Table 3.11 - Metrics applied to measure the centrality of nodes

Overlap analysis within and between groups

RQ2 aims to explore the overlap between central positions within and between groups in an ESN. To answer these questions, the networks have been created according to the algorithm stated in section 3.4.1, and the centrality of actors in all groups have initially been calculated according to Table 3.11. The result is a list of users with their corresponding centrality measures; weighted in-degree centrality, weighted out-degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. These ranked lists of users according to their centrality provides the basis for measuring the overlap of users between central positions in the different group networks.

To measure the overlap between central positions *within* a group, one table for each group is first produced. Each table consists of five columns and five rows, where each column- and row-name refer to one of the five centrality metrics. Furthermore, each row and column is partitioned into two sub-rows or sub-columns, which refer to the top 5% and the top 10% of users within each centrality metric, respectively. By inspecting the resulting table, one can investigate the overlap between two sets of central positions within each group by looking at the value of each intersecting cell between two centrality metrics. The values are depicted in percentages, e.g. if the

intersecting cell between the row "Eigenvector centrality (top 5%)" and the column "Closeness centrality (top 10%)" in group ABX1 holds a value of 60.8, then this means that 60.8% of the top 5% users according to eigenvector centrality in ABX1 are also present among the top 10% of users according to closeness centrality in the same group.

To measure to what extent users are central in *multiple* groups, a similar approach as the one described in the previous paragraph is employed. However, the cell values will depict, on average, how many *other* groups the top users according to the centrality measure in the row-names are central in, and which central position they hold in other groups. E.g. if the intersecting cell between the row "Eigenvector centrality (top 5%)" and the column "Closeness centrality (top 10%)" in group ABX1 holds a value of 0.8, then this means that, on average, users who are among the top 5% according to eigenvector centrality in at least one group, are among the top 10% according to closeness centrality in 0.8 (-1) other group.

The pseudo code that was implemented in R in order to calculate the overlaps can be found in Appendix B.

3.5 Inter-correlation analysis between genre analysis results and SNA metrics

3.5.1 Determining the relation between the network structure and the conversational nature in the ESN

As described in section 3.3.1 and 3.3.2, the messages within the given time windows (~6 weeks) have been categorized according to the genre repertoire presented in Table 3.7, and the structural characteristics of the networks in the corresponding time windows have been calculated. See Appendix C for a complete overview. In order to determine the relation between network structures and the conversational nature in these networks, a Pearson's r product-moment correlation analysis was conducted. The research question which this method aims to satisfy is RQ1, which aims to determine the relation between these two constructs.

After all messages had been categorized according to the genre repertoire in Table 3.7, the values of each genre was aggregated which resulted in a percentage-wise relative distribution of genres for each group within each time period. Having calculated the structural characteristics of the networks according to Table 3.9, both the results from the genre analysis and the SNA were imported into R for further analysis. The resulting table consisted of 14 variables (nine genre categories plus five SNA metrics) across 30 time periods ($n = 30$). All the columns in the resulting table were then correlated using the following command in R (R Development Core Team, 2013);

```
cor(x, y = NULL, use = "everything", method = c("pearson", "kendall",
"spearman"))
```

where the parameter `method` was set to `"pearson"`. No NA values were detected in the dataset, so no further adjustment of the R syntax was deemed necessary.

The resulting correlation table was then investigated in order to detect significant correlations between the variables included in the analysis. For this project, all correlation coefficients with a p-value >0.05 were deemed as significant. Moreover, if a correlation coefficient between two variables were detected as significant, a scatterplot was constructed in order to visually investigate the distribution of correlated data points. This was mainly done to detect outliers which could potentially have an impact on the significance of the correlation. If any outliers were detected in the data distribution, then the correlation analysis between the two variables at hand was re-done in order to detect any potential discrepancy in the outcome of the analysis.

3.5.2 Determining the relation between the structural characteristics of users and their conversational nature

RQ2 puts focus on the relation between the conversational topics and the centrality of users in the network. After the genre analysis was completed, all users included in the sample were compiled into a separate table. For each user, their relative percentage-wise distribution of genres was calculated in addition to their centrality metrics. This calculation was done both on the group-level as well as the network-level. For the group-level, the genre metrics and the SNA metrics were calculated by first partitioning all messages according to which group the messages resided in. The centrality metrics and the genre metrics were then measured for each user within each group, i.e. the groups themselves were treated as independent networks in this regard. Their centrality metrics and their genre metrics were calculated based on their involvement in the ABX network in its entirety. As users could be members of several groups, the first approach might result in some of the same users being redundant across groups. The second approach ensures that the same user is only represented once in the network, by including all users' activity over all groups in calculating the genre- and centrality-metrics.

To determine the relation between users' structural characteristics and their message topics (i.e. genre metrics), a linear regression model based on permutations was conducted. Regression analysis *"allows prediction or estimation of the value of one variable (the criterion, dependent, or predicted variable; traditionally called Y) from one or more predictor variables (called X_n)."*(Tompkins, 1992). This method for testing the relation between users' attributes and their centrality in a network is described in Hanneman and Riddle (2005, p. 300-304). In this case the users' attributes corresponded to their relative distribution of messages within each genre, and the centrality metrics corresponded to the metrics depicted in Table 3.10. As described by Hanneman and Riddle (2005), a researcher can apply a regression model to test the relation between a set of independent variables and a dependent variable, e.g. closeness centrality. The main difference when dealing with social network data, is that the relations between actors cannot be deemed as independent, and must therefore comply to a different statistical approach when testing for the relation between user attributes and their structural characteristics (Hanneman & Riddle, 2005, p. 303). Furthermore, the distribution of the population is unknown which excludes a set of commonly applied statistical tests in order to test for significance. The solution proposed by Hanneman and Riddle (2005) is to use permutation tests, which means that the observation included in the test are permuted (i.e. shuffled) in order to calculate standard errors towards

which the observed relation between variables can be tested against. The specific implementation of such a test suitable for the objective at hand is the multiple linear regression-method (with permutations). In practice, the statistical tests between genre metrics and centrality metrics for users were done using the `lmorigin` –command included in the `ape`-package in R(Paradis, Claude, & Strimmer, 2004). For all groups, and the network, the following command was executed in order to test the relation between individual genre metrics and the individual centrality metrics of users;

```
lmorigin(formula, data, origin=FALSE, nperm=10000, method="raw", silent=FALSE)
```

where `formula` was set to describe the model that was being tested using the command, `data` was set to the list of users with their corresponding genre- and centrality-metrics, `origin` was set to `FALSE`, `nperm` (number of permutations) was set to 10000, and `method` was set to `"raw"`. The syntax for defining the `formula` parameter is; `y ~ x1 + x2 + x3 + ... +xn`, where `y` is the dependent variable, and `x1...xn` is the set of independent variables (predictors).

Some genres were completely absent in some of the groups, e.g. the genres "Problem solving" and "Social" were not represented at all in the group ABX2. Since the permutation tests relies on non-null data in order to conduct the permutation tests (there has to be *something* to actually permute), these variables which consisted only of null-data were removed from the permutation tests.

In total the `lmorigin`-command was run 30 times; 5 centrality metrics per group * 5 groups + 5 centrality metrics * 1 complete network. For clarity, an example of a model that tested for the relation between genre metrics and betweenness centrality among ABX1-users is given below;

```
lmorigin(formula = ABX1_users[,"betweenness_w"] ~
ABX1_users[,"Information.input"] + ABX1_users[,"Discussion"] +
ABX1_users[,"Problem.solving"] + ABX1_users[,"Idea.Generation"] +
ABX1_users[,"Status.updates"] + ABX1_users[,"Notifications"] +
ABX1_users[,"Praise"], origin = FALSE, method = "raw", nperm = 10000)
```

In the example above, the different genre variables were regressed onto betweenness centrality. When running the command above for all centrality metrics in all groups (and the complete network), the `lmorigin`-command returns a set of test statistics. For clarification on how to read the result, and on how the results from the MLR analysis were used in this project, an example of the R output from the `lmorigin`-command is given in Table 3.12.

```
Coefficients and parametric test results
                                Coefficient Std_error t-value Pr(>|t|)
(Intercept)                        321.7946 1339.1030  0.2403   0.8112
ABX1_users[, "Information.input"]    2.9862   16.1537  0.1849   0.8542
ABX1_users[, "Discussion"]           1.6689   13.7723  0.1212   0.9041
ABX1_users[, "Problem.solving"]     17.3150   24.5642  0.7049   0.4847
ABX1_users[, "Idea.Generation"]      6.7515   27.4901  0.2456   0.8072
ABX1_users[, "Status.updates"]      76.0196   59.0116  1.2882   0.2046
ABX1_users[, "Notifications"]       10.6022   18.7545  0.5653   0.5748
ABX1_users[, "Praise"]              -1.9099   19.0206 -0.1004   0.9205

Two-tailed tests of regression coefficients
                                Coefficient p-param  p-perm
(Intercept)                        321.7946  0.8112      NA
ABX1_users[, "Information.input"]    2.9862  0.8542 0.71743
ABX1_users[, "Discussion"]           1.6689  0.9041 0.87981
ABX1_users[, "Problem.solving"]     17.3150  0.4847 0.30147
ABX1_users[, "Idea.Generation"]      6.7515  0.8072 0.66553
ABX1_users[, "Status.updates"]      76.0196  0.2046 0.06799 .
ABX1_users[, "Notifications"]       10.6022  0.5748 0.31147
ABX1_users[, "Praise"]              -1.9099  0.9205 0.79692
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1437.884 on 43 degrees of freedom
Multiple R-square: 0.05826552   Adjusted R-square: -0.09504009

F-statistic: 0.3800613 on 7 and 43 DF:
   parametric p-value   : 0.9089546
   permutational p-value: 0.4988501
after 10000 permutations of residuals of full model
```

Table 3.12 – Example of R-output from MLR analysis

The upper table depicted in Table 3.12 depicts the test statistics for the MLR analysis, including regression coefficients, standard errors, and t-values for all predictors (independent variables). The lower table depicts the test statistics from the permutation test, where all observations of the betweenness centrality and the different genres for the group at hand have been shuffled 10000 times in order to calculate a distribution against which the observed coefficients can be compared.

The lower table in Table 3.12 shows the comparison between permuted coefficients and observed coefficients. The information used to determine the significance of predictors can be found in the "p-perm"-column, which depicts the proportion of permutation tests that resulted in a p-parameter ("p-param"-column) equal or higher than the observed value. As an example, in Table 3.12, it can be seen that out of the 10,000 permutations of observations between the Status updates-genre and the betweenness centrality of users, only 680 (10,000 * 0.06799) of these permutations resulted in an observed value equal or higher than the value found in the "p-param"-column. This result was only significant on the 0.1-level (two-tailed), and was therefore not deemed as significant since the significance level applied in this project is 0.05 (two-tailed). Furthermore, information about the R^2-values and the F-statistics can be found on the bottom of

the R output in Table 3.12. This information is only relevant if the goal was to fit a complete model, which is not the case in this project, as the objective is to single out individual genres that are significant in predicting the betweenness centrality of users. Thus, only the information found in the "Two-tailed test of regression coefficients"-table was used to test for significance.

As previously describes, MLR analyses test for the significance of predictors when controlling for all other predictors, i.e. it determines the effect of each predictor when all other predictors are held constant. This way of conducting a significance test of the relationship between variables are therefore not equivalent to conducting zero-order correlation analysis between variables (e.g. Pearson product-moment correlation analysis), as this test for the linear relationship between two independent variables without considering the effect of other co-variables (Nathans, Oswald, & Nimon, 2012). In other words, MLR and zero-order correlation analysis tells two different sides of the story. A common way to add more nuance to the testing of relationships between variables is by both conducting both a MLR and a correlation analysis between predictors and the independent variable, as these two types of analysis complement each other and allows for a more informed interpretation of the relationship between the variables (Nathans et al., 2012). Hence, in order to add more nuance to the relationship between users' genre metrics and their centrality measures, a Pearson correlation analysis was done in conjunction with the MLR analysis, where each of the predictors (genres) were individually correlated with each of the centrality metrics.

Chapter 4

Results

This chapter will present the findings from the data analysis phase outlined in Chapter 3. The first sub-chapter will report on the results from the genre analysis and social network analysis conducted for all five groups. These results are based on the aggregated metrics for the genre analysis and the social network analysis, meaning that the different metric values have been calculated based on the total accumulated activity during the entirety of each group's duration. The second subchapter will address the results of the longitudinal analysis conducted based on the temporal snapshots of each group's activity over time. Here, the metrics representing the structural characteristics of each temporal network are correlated with the relative distribution of genres within the same time period as each temporal network. The third sub-chapter will present the findings from the analysis of actors' roles in each network and the topics of their communication. The results from studying central actors' different roles within the same online community, their roles in multiple online communities, and their ability to maintain their central roles over time are presented in sub-chapter 4.3, 4.4, and 4.5, respectively.

4.1 The network structure and conversational nature of ABX

This sub-chapter will provide a summary of the structural characteristics and the conversational nature found within the different groups in ABX, and provides the fundament for the statistical analysis conducted in sub-chapter 4.2. Since the study of the communication content within each group were based on a sampling of messages within certain time periods, the results shown in this sub-chapter represent the aggregated genre-values over all time periods for each groups. In turn, the mean of the frequencies for each specific genre within each group is used to represent distribution of genres within the whole network, i.e. all groups combined.

Table 4.1 shows the frequencies found for each genre within each group, in addition to the aggregated total count for each genre.

	ABX1	ABX2	ABX3	ABX4	ABX5	Grand total
Social	0	0	10	0	0	10
Information input	31	77	58	41	39	246
Praise	13	14	53	39	14	133
Other	10	9	7	4	8	38
Idea Generation	10	23	5	2	33	73
Notification	16	45	36	4	11	112
Discussion	209	140	78	124	174	725
Status updates	5	12	14	18	24	73
Problem solving	21	0	29	86	64	200
Grand total	315	320	290	318	367	1610

Table 4.1 – Frequency count for each genre within each group

As Table 4.1 shows, the frequency of genres vary between groups, implying that the different groups serve different purposes. There is also a notable variation between genres when counted for the whole dataset (all groups combined), implying that the Yammer platform is dominantly used for certain purposes (mainly discussion). The lowest score can be found in the Social genre, which only accounts for a total of 10 messages in the total dataset. Furthermore, one group accounts for the entirety of this genre's frequency count; ABX3. The highest count of occurrences can be found in the Discussion genre which accounts for a total of 725 messages, almost three times as many as the succeeding genre in terms of frequency count which is information input. As can be seen from Table 4.1, the number of genre categories allocated within each group were not equivalent to each other, which was a direct consequence of two factors; 1) the messages within each group were sampled based on their thread ID and not their message ID, resulting in a different number of messages being sampled in each group as a result of each thread containing a different number of messages, 2) each message could get allocated more than one genre, resulting in the average genres/message-ratio deviating across the groups. The frequencies of each genre within each group therefore needed to be normalized in order to give a fair impression of how the different genres were represented, taking into account the total number of genres allocated for each group. Figure 4.1 shows the relative frequency of genres for the whole dataset, and Figure 4.1 shows the relative frequencies per group expressed in percentages.

As can be noted from Figure 4.1, Discussion was the most prominent genre in all of the five groups, ranging from 66.3% in ABX1, and 26.9% in ABX3. Information input, being the second highest allocated genre category, ranged from 24.1% in ABX2 to 9.8% in ABX1. It was an interesting observation that only one group contains messages in all of the nine different genre categories, which is ABX3. All other genre categories are present in all of the groups with the exception of Problem solving, which is absent in ABX2. This category however, represents 27.0% of genres in ABX4, and 17.4% of genres in ABX5. The above-mentioned observations acted as a testimony the composition of communication topics across the groups, which implicitly can be described as diverse with respect to all of the different genre categories.

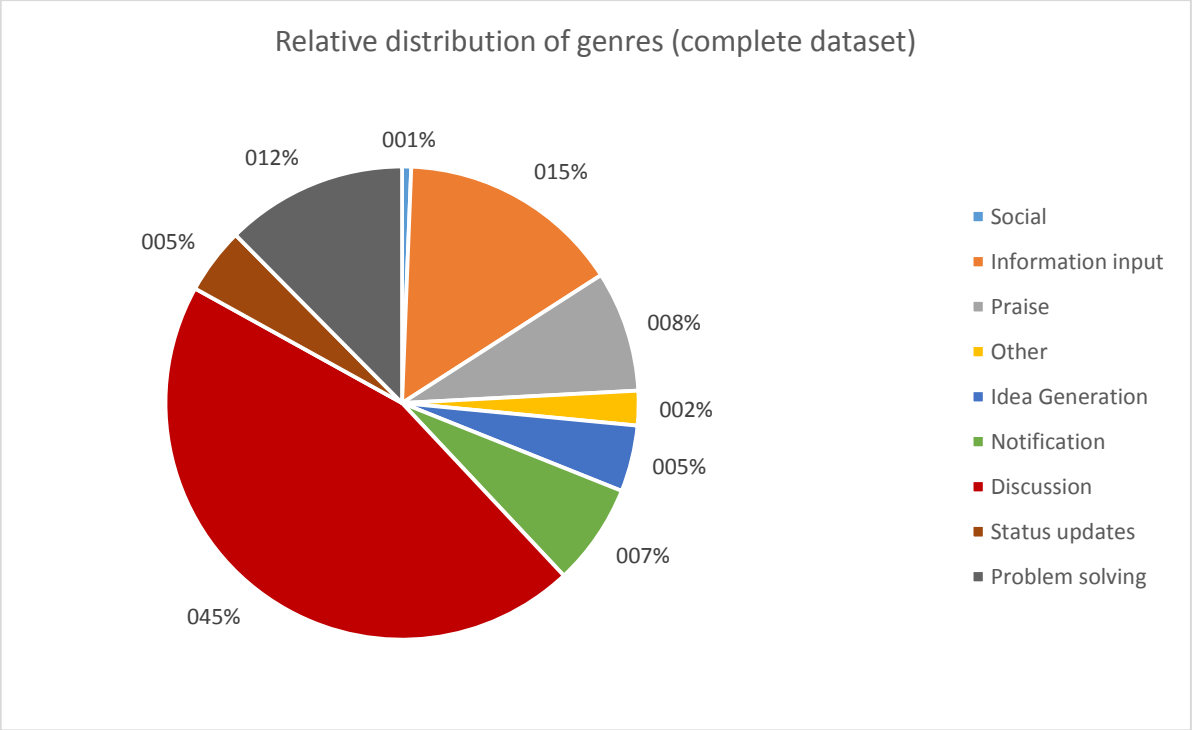

Figure 4.1 - Relative distribution of genres for the complete dataset

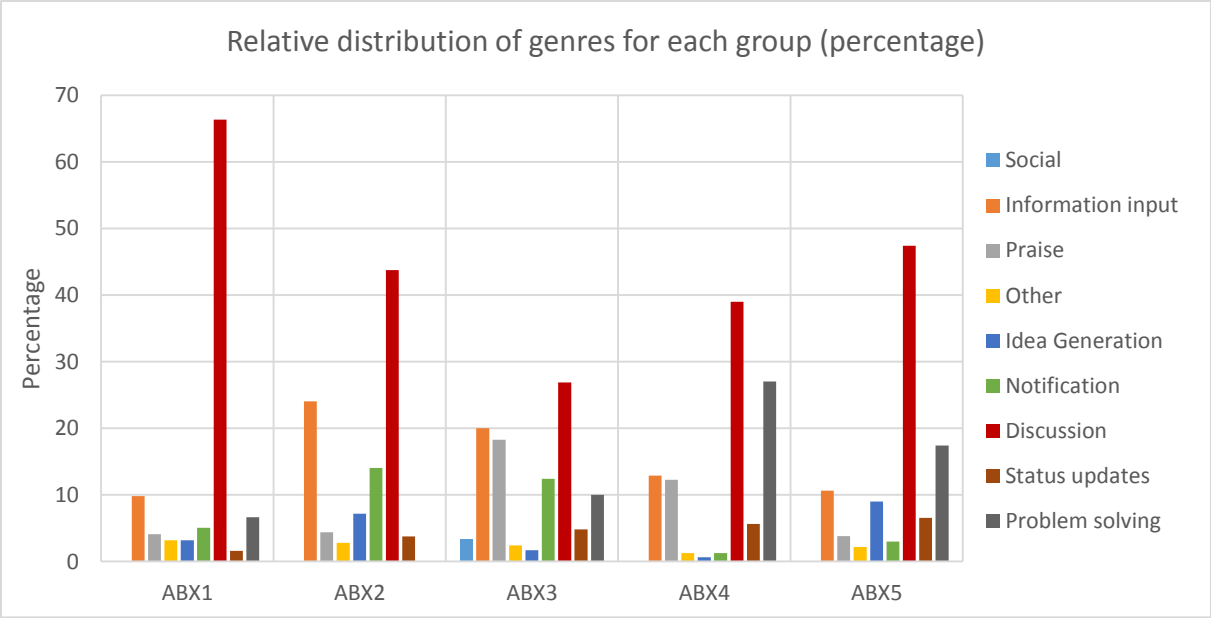

Figure 4.2 - Relative distribution of genres within all groups

Having displayed the composition of genres within the groups, the structural characteristics of each group were also computed and visualized according to the methods described in Chapter 3. As stated in the mentioned chapter, the metrics used to describe the networks studied in this research were average weighted degree centralization, graph density, average clustering coefficient, reciprocity, and average path length. Table 4.2 shows a summary of these calculated metrics for each group along with the relative distribution of genres expressed in percentage. To give an impression of the size of the network, the count of nodes (actors) and the count of edges (ties) are also included in the table.

		ABX1	ABX2	ABX3	ABX4	ABX5
Genre categories	**Social**	0.00	0.00	3.45	0.00	0.00
	Information input	9.84	24.06	20.00	12.89	10.63
	Praise	4.13	4.38	18.28	12.26	3.81
	Other	3.17	2.81	2.41	1.26	2.18
	Idea Generation	3.17	7.19	1.72	0.63	8.99
	Notification	5.08	14.06	12.41	1.26	3.00
	Discussion	66.35	43.75	26.90	38.99	47.41
	Status updates	1.59	3.75	4.83	5.66	6.54
	Problem solving	6.67	0.00	10.00	27.04	17.44
Network metrics	**Nodes**	174	121	240	191	155
	Edges	721	250	621	702	660
	Average weighted degree	9.85	3.98	3.48	5.78	16.75
	Graph density	0.024	0.017	0.011	0.019	0.028
	Average clustering coefficient	0.25	0.15	0.11	0.23	0.35
	Average path length	2.55	2.85	3.29	3.15	2.54
	Reciprocity	0.59	0.28	0.29	0.26	0.69

Table 4.2 - Network metrics and relative genre distribution per group

The largest network was ABX3, with a total count of 240 nodes. The network with the most active users was ABX5, with an average weighted degree per node of 16.75. This number represents the average number of in- and out-going ties per node in the network multiplied by their individual weights. Compared to ABX3, the average weighted degree centralization in ABX5 was almost 5 times as large, and over twice as large as the average over all groups (7.97). ABX5 was also the group with highest density (2.8%), highest average clustering coefficient (0.35), smallest average path length (2.54), and the highest degree of reciprocity (0.69). ABX5 as a network could therefore be said to be the most active on average, the most tightly connected, the most clustered, the network with the least degree of separation between the nodes, and the network where the highest proportion of posts are reciprocated from the receiver On the opposite end is ABX3, which got the opposite testimonial according to all the aforementioned metrics.

Figure 4.3 depicts the visual representation of all the five groups. In each of the networks, the circles represent nodes and the lines between the circles represent the ties. For visual effect, the nodes' color and size is determined by their weighted degree centrality, where a large, red node corresponds to a large weighted degree centrality score, and the blue, small circles corresponds to a small weighted degree centrality score. The same legend is applied to all of the following social graphs in this thesis.

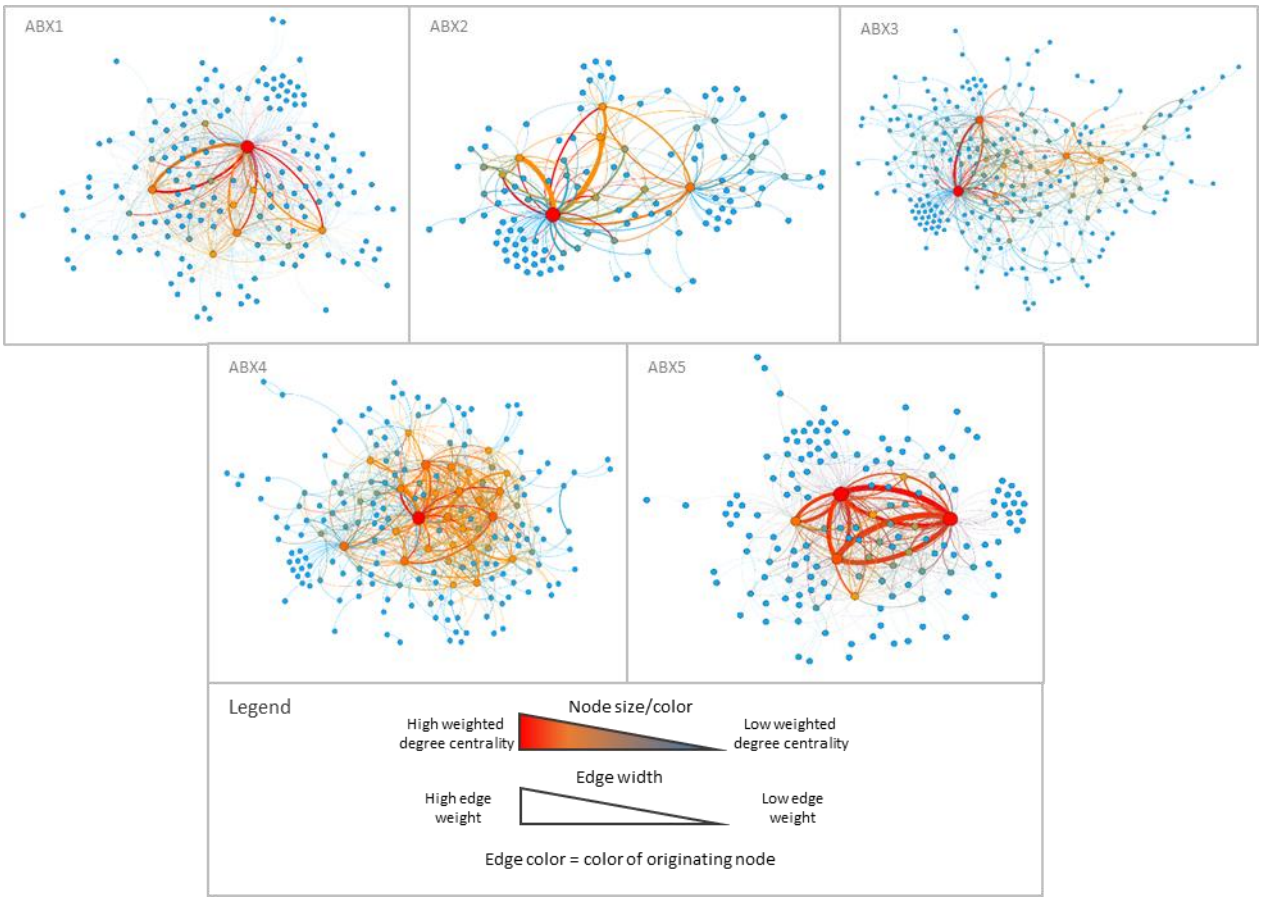

Figure 4.3 - Visual representation of the different groups

The visual representation of each group allows for a more intuitive interpretation of the structural composition of the networks. As an example, one can visually inspect and get an impression of how the weighted degree centralization is distributed amongst the actors in the network. For this purpose, the width of the ties in Figure 4.3 are determined by their weights, and the color is determined by the ties' originating nodes. As can be seen in the figure, the structure of ABX5 seem to be characterized by a relatively low number of central actors, but with a large weight of the ties that connect these central actors to each other. Moreover, by comparing the visual impression of the groups, one can infer that all of the groups seem to comprise a relatively small number of central actors (according to weighted degree centrality) who amongst themselves are tightly connected, and that the majority of actors in the different groups seems to be inactive to a large degree as reflected by their blue color and small size. The Force Atlas layout algorithm applied also allows for the visual detection of local clusters within the different groups, which further contributes to conveying the complexity and composition of such networks in an intuitive manner.

4.2 The relation between network structure and conversational nature in the ABX network

This sub-chapter will present the results from the longitudinal analysis, that is, the genre- and structural analysis that was conducted by partitioning the activity of each group into time

periods, i.e. snapshots. The results in this sub-chapter provides the main body of results needed to answer RQ1.

Figure 4.4 shows a stacked bar-chart which represents the relative distribution of each genre within each snapshot of all the groups. A reference for which time period each snapshot represents can be found in Table 3.6 in Chapter 3.

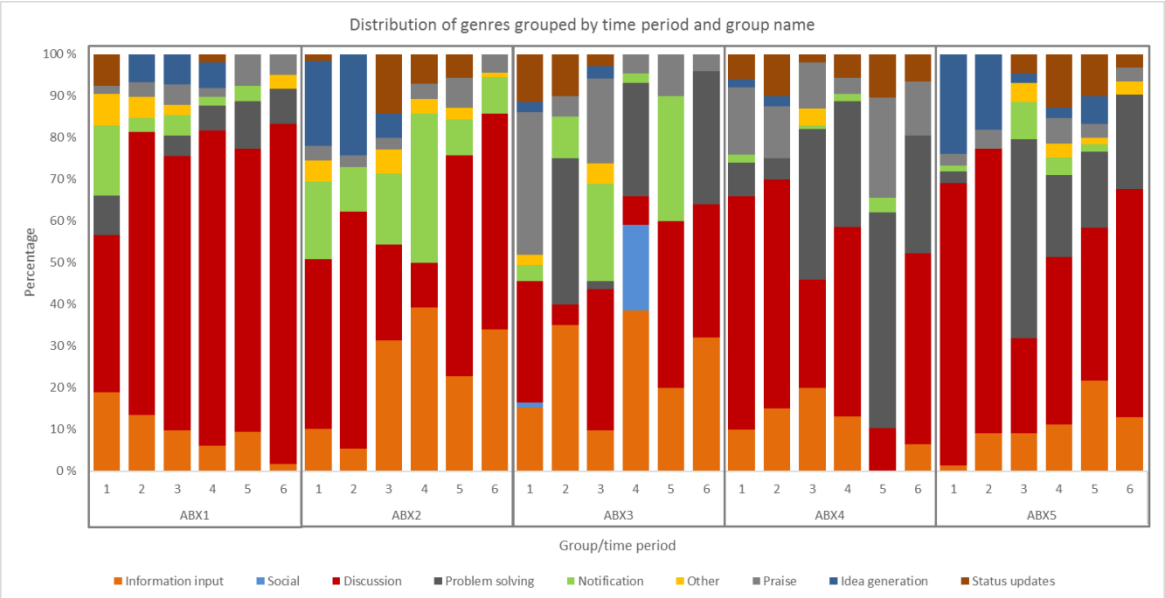

Figure 4.4 - Relative distribution of genres per time period for each group

The visual representation in Figure 4.4 provides further nuance to how the distribution of genres fluctuated over time within each group, as opposed to Figure 4.2, which only represents the aggregated values during the entirety of each group's duration. It is apparent from inspecting the chart that the distribution of genres are not static and uniform within each group over time, but rather a volatile value. As an example, the occurrence of the Discussion genre within ABX1 seemed to be a steadily rising value throughout its duration, while as for ABX3 it fluctuated between 5.0% in the second time period and 40.0% in the fifth time period. And while Information input seemed to be a rising occurrence in ABX5, the same genre steadily diminished in ABX1 as time went on. It was also an interesting observation that none of the time periods comprised all of the nine genre categories. In fact, the average number of genres present in each time period was 6.13, leaving ~1/3 of the genre categories absent. Discussion was the only genre category that had presence in all the different time periods, while the Social genre were absent from 28 out of the 30 time periods.

To explore if the aforementioned distribution of genres were associated with the structure of the networks in each snapshot, a Pearson's product-moment correlation analysis was conducted by correlating the values for each genre in each time period with the network metrics calculated for the same time periods. Each temporal network were also visually constructed in Gephi, and the results are shown from Figure 4.5 through Figure 4.9. In these figures, the size and color is determined by their weighted degree centralization as in Figure 4.3. The user ID's are also

included, which allows for the visual detection of which users holds central positions in the different time periods.

The visual representation of the networks seemed to confess the same reality as with the distribution of genres, that is, to a fluctuating network structure depending on the time period. Both the size and the composition of the temporal networks, as can be observed, varied both across and within the different groups.

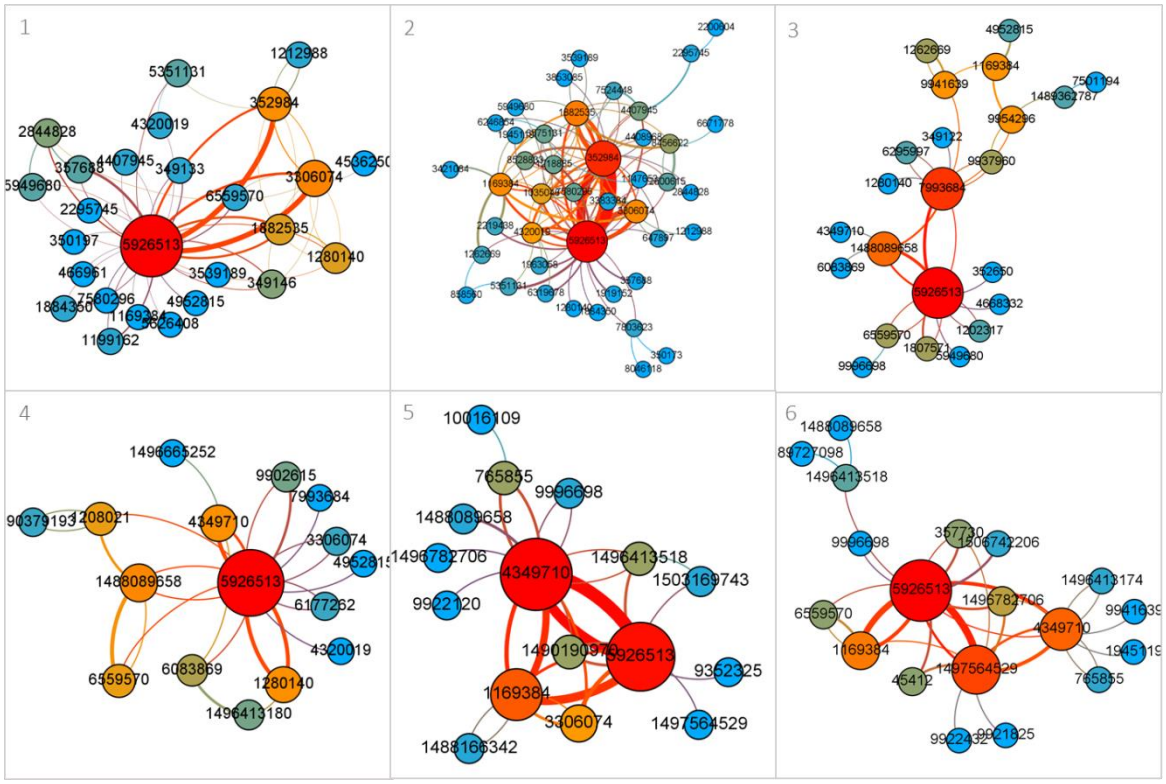

Figure 4.5 – Network snapshots (ABX1)

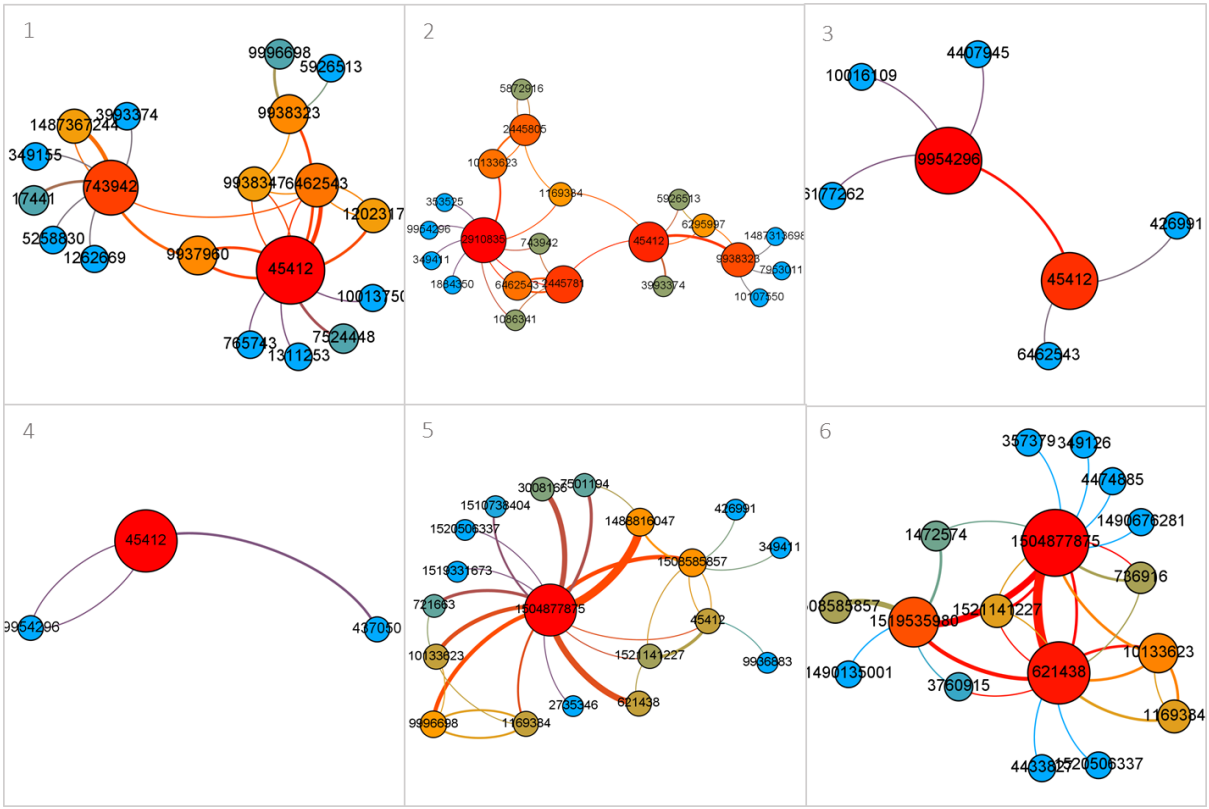

Figure 4.6 - Network snapshots (ABX2)

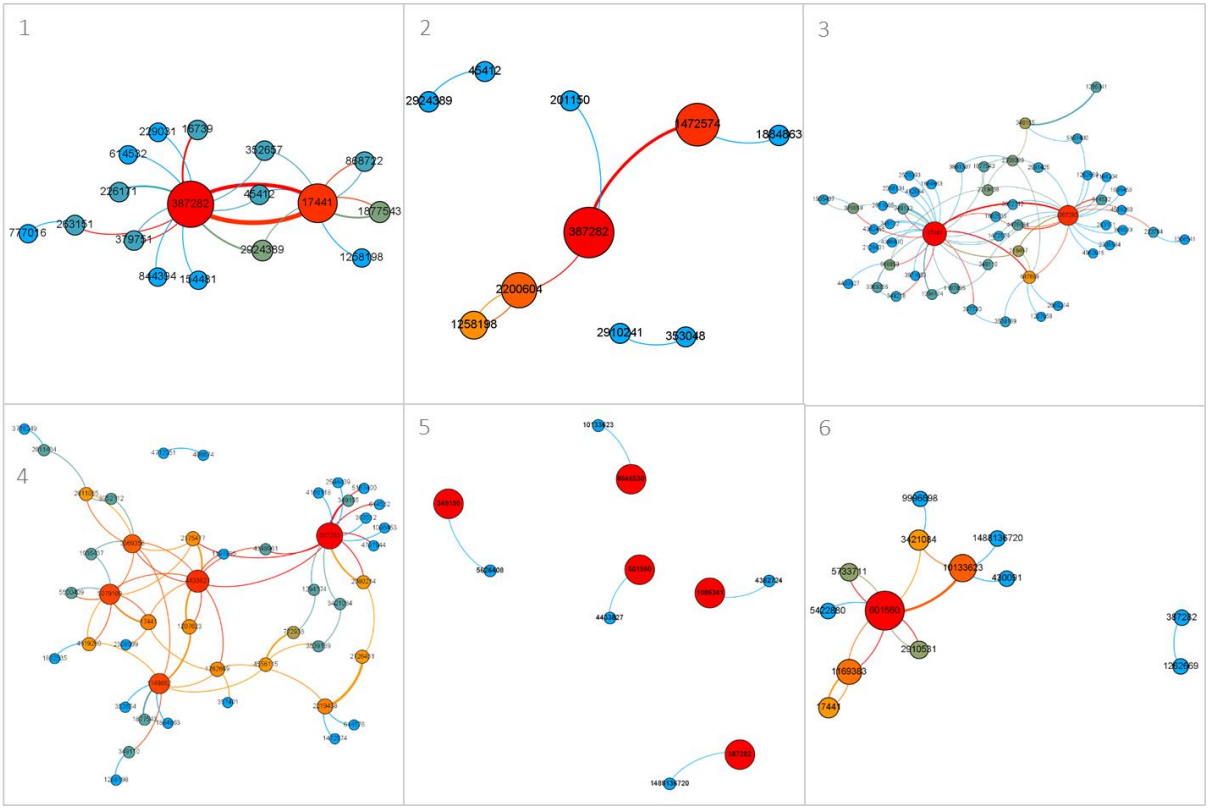

Figure 4.7 - Network snapshots (ABX3)

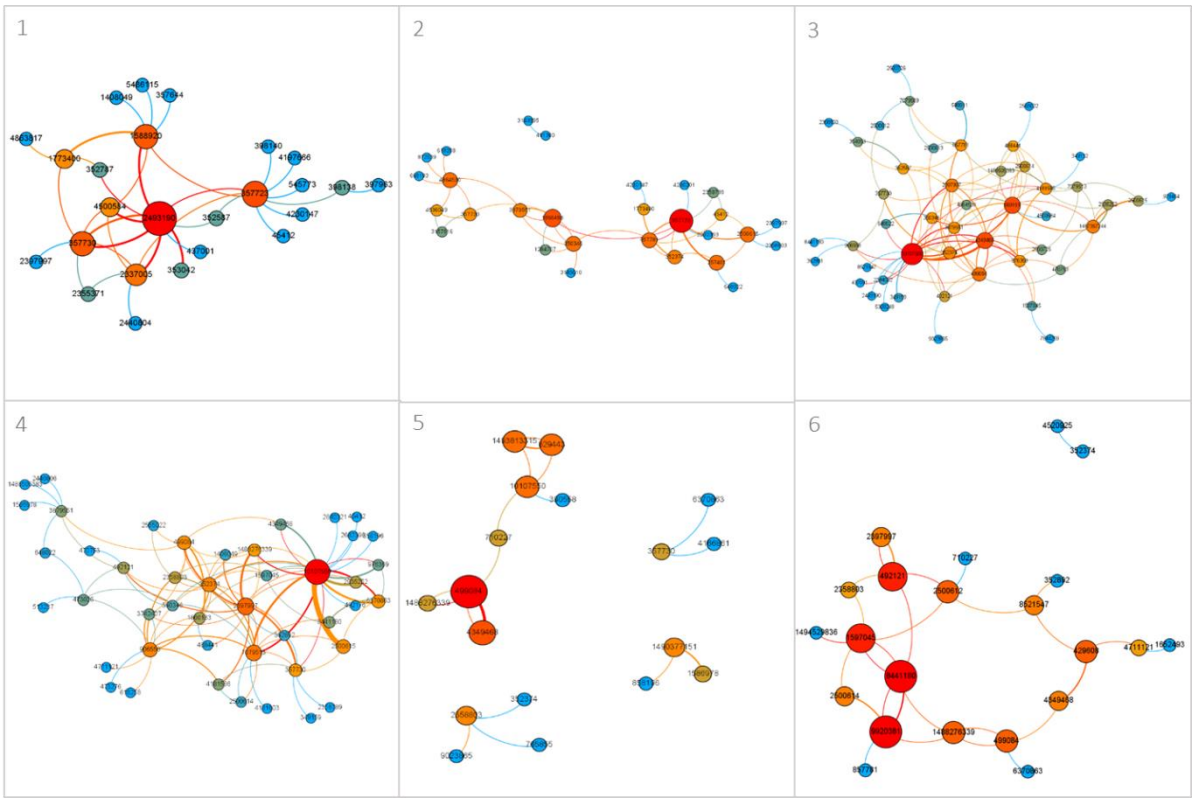

Figure 4.8 - Network snapshots (ABX4)

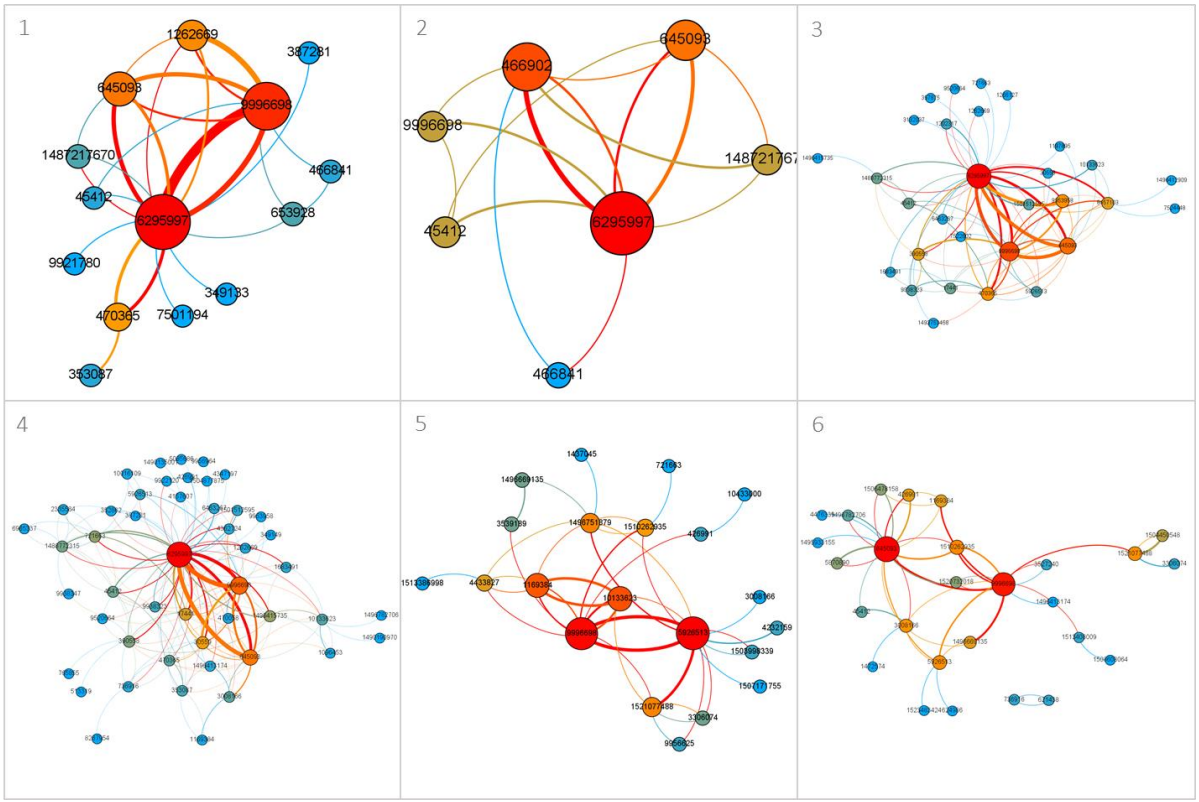

Figure 4.9 - Network snapshots (ABX5)

Table 4.3 presents the results from correlating the structural characteristics (SNA metrics) and the conversational nature (genre categories) of ABX (see Appendix C for the data values used), and the correlation analysis between the genre categories themselves. Although the intra-correlation between genres are not specifically of any interest with respect to the research questions, they are relevant to consider for the following reason; the correlation analysis between the genre categories and the network metrics only considers the relative percentage-value of each genre *individually*. Since the values are measured in percentages, it means that if one genre increases or decreases in value from one observation to the next, this will cause the percentage-value of the other genres to increase or decrease as well. These interdependencies are not accounted for in a correlation analysis. Thus, it makes sense to take into account the intra-correlation between the genre categories as well in order to provide a better ground for discussing the results. As these intra-correlations are not directly tied to the research questions, they are not presented as part of the results in this chapter but are rather incorporated into the discussion presented in Chapter 5.

		Genre categories								
		Information input	**Social**	**Discussion**	**Notification**	**Other**	**Praise**	**Problem solving**	**Idea generation**	**Status update**
Network metrics	**Reciprocity**	-0.157	-0.139	0.195	-0.130	0.232	-0.337	0.142	-0.023	-0.136
	Average weighted degree centralization	-0.277	-0.114	0.166	-0.215	0.342	-0.311	0.191	0.050	0.057
	Graph density	0.260	-0.145	-0.051	**0,363***	0.020	-0.230	-0.275	0.174	0.006
	Average clustering coefficient	**-0,487****	-0.157	**0,474****	**-0,446***	-0.022	-0.149	-0.042	**0,409***	-0.195
	Average path length	-0.123	0.210	0.221	**-0,460***	0.133	-0.035	0.145	-0.166	-0.184
Genre categories	**Information input**	1								
	Social	**0,378***	1							
	Discussion	**-0,585****	0.338	1						
	Notification	0.355	0.110	**-0,383***	1					
	Other	0.076	0.167	-0.138	**0,392***	1				
	Praise	-0.213	0.035	-0.220	-0.050	-0.131	1			
	Problem solving	-0.008	0.155	**-0,508****	-0.356	-0.170	0.042	1		
	Idea generation	**-0,387***	0.118	0.352	-0.018	-0.070	-0.248	**-0,409***	1	
	Status update	0.159	0.166	**-0,493****	0.076	0.198	0.335	0.197	-0.293	1

Table 4.3 - Correlation analysis between genre categories and network metrics (n=30; *-significant, α = 0.05, two-tailed; **-significant, α = 0.01, two-tailed)

Two correlations were found to be significant at the 0.01-level:

- There was a negative correlation between the amount of Information input-messages and the average clustering coefficient in the temporal networks, r(28) = -.487, p = < .01, two-tailed.
- There was a positive correlation between the amount of Discussion-messages and the average clustering coefficient in the temporal networks, r(28) = .474, p = < .01, two-tailed.

Figure 4.10 and 4.11 shows the scatterplots of the correlations found to be significant at the 0.01-level.

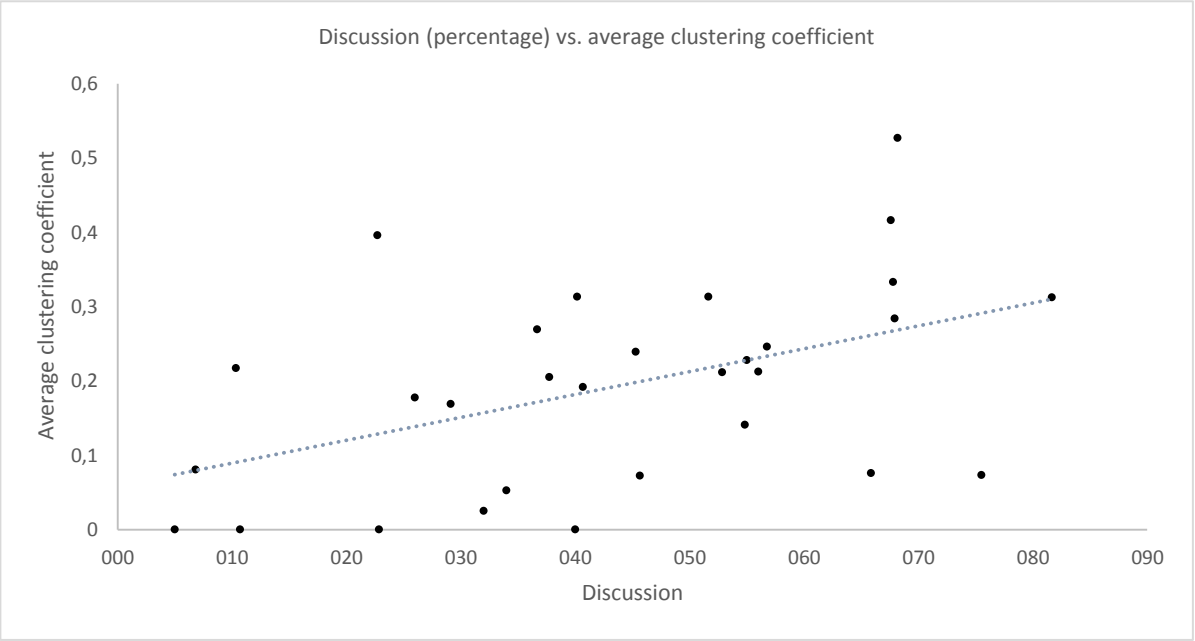

Figure 4.10 - Discussion (percentage) vs. average clustering coefficient (scatterplot)

The correlation between the amount of Discussion-messages and the average clustering coefficient in the temporal networks imply that time periods that consist of more discussions are correlated with a higher number of closed triplets in the network, i.e. a higher degree of clustering. The metric used to measure the global clustering coefficient in this research does also take into account the weight of the ties that connects the nodes that are part of closed triplets, meaning that the total *count* of triplets is not the only criteria used to calculate the clustering coefficient itself, as is done when dealing with un-weighted networks. It should also be noted that the clustering coefficient can only give *one* value per network. In temporal networks where there exists more than one component of connected nodes, e.g. ABX4 (fifth time period), the clustering coefficient will be calculated only for the giant component, i.e. the largest set of connected nodes in the network. In temporal networks where there is multiple components consisting of only two nodes, e.g. ABX3 (fifth time period), the algorithm for calculating the clustering coefficient will yield the value of 0, as there are no components with a sufficient number of nodes to calculate the coefficient. As can be observed from Figure 4.10, there were a total of four temporal networks that had a clustering coefficient equal to 0 due to the mentioned reasons, namely ABX2 (3rd and 4th time period), and ABX3 (2nd and 5th time period). In each of these cases, there were not enough nodes in either of the components to calculate the ratio

between closed triplets and the total number of possible triplets, resulting in a clustering coefficient equal to 0.

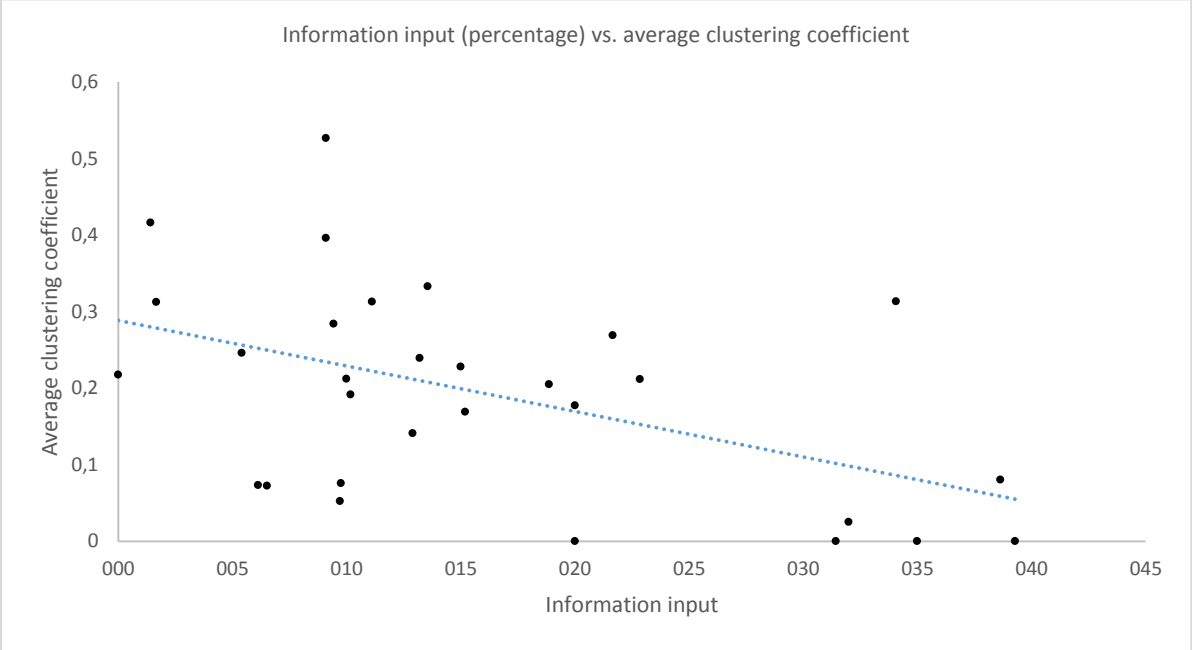

Figure 4.11 – Information input (percentage) vs. average clustering coefficient (scatterplot)

The observations found in Figure 4.11 are comparable to the observations in Figure 4.10, but then in the opposite direction. There seemed to be a significant, negative correlation between the amount of information input-messages and the clustering in the temporal networks.

Five correlations were found to be significant at the 0.05-level:

- There was a negative correlation between the amount of Notification-messages and the average clustering coefficient, $r(28) = -.446$, $p = < .05$, two-tailed.
- There was a negative correlation between the amount of Notification-messages and the average path length, $r(28) = -.460$, $p = < .05$, two-tailed.
- There was a positive correlation between the amount of Notification-messages and the graph density, $r(28) = .363$, $p = < .05$, two-tailed.
- There was a positive correlation between the amount of Idea generation-messages and the average clustering coefficient, $r(28) = -.409$, $p = < .05$, two-tailed.

Figure 4.12 through 4.15 shows the scatterplots of the correlations found to be significant at the 0.05-level.

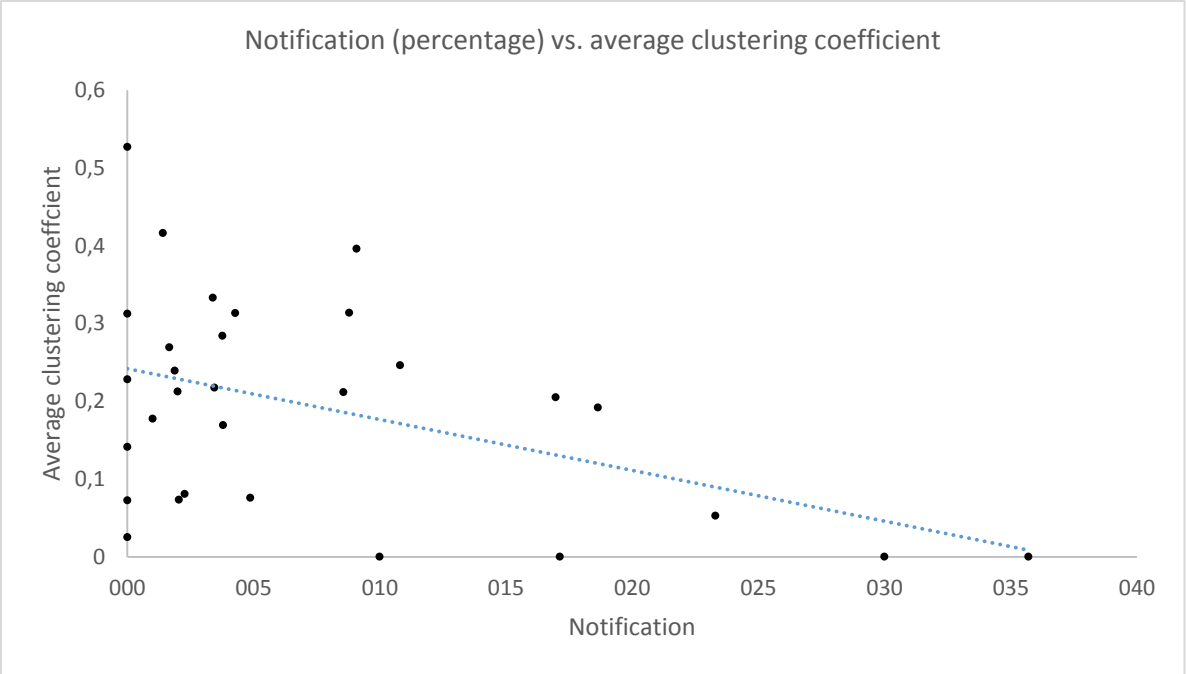

Figure 4.12 - Notification (percentage) vs. average clustering coefficient (scatterplot)

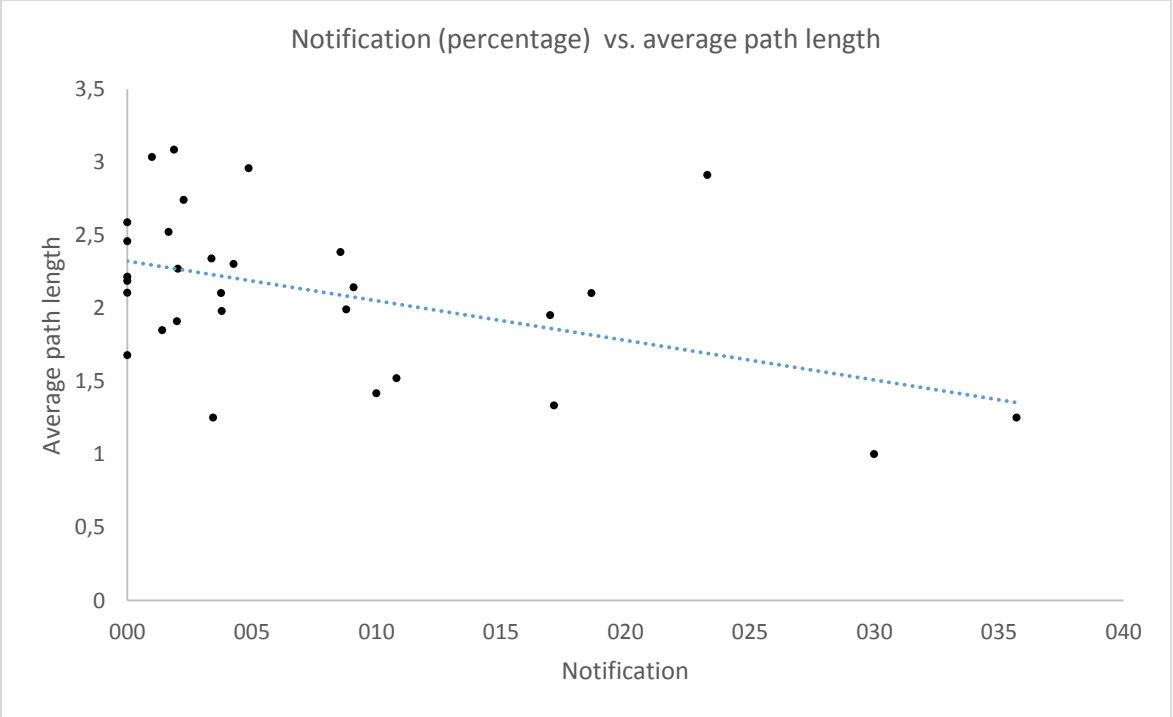

Figure 4.13 - Notification (percentage) vs. average path length (scatterplot)

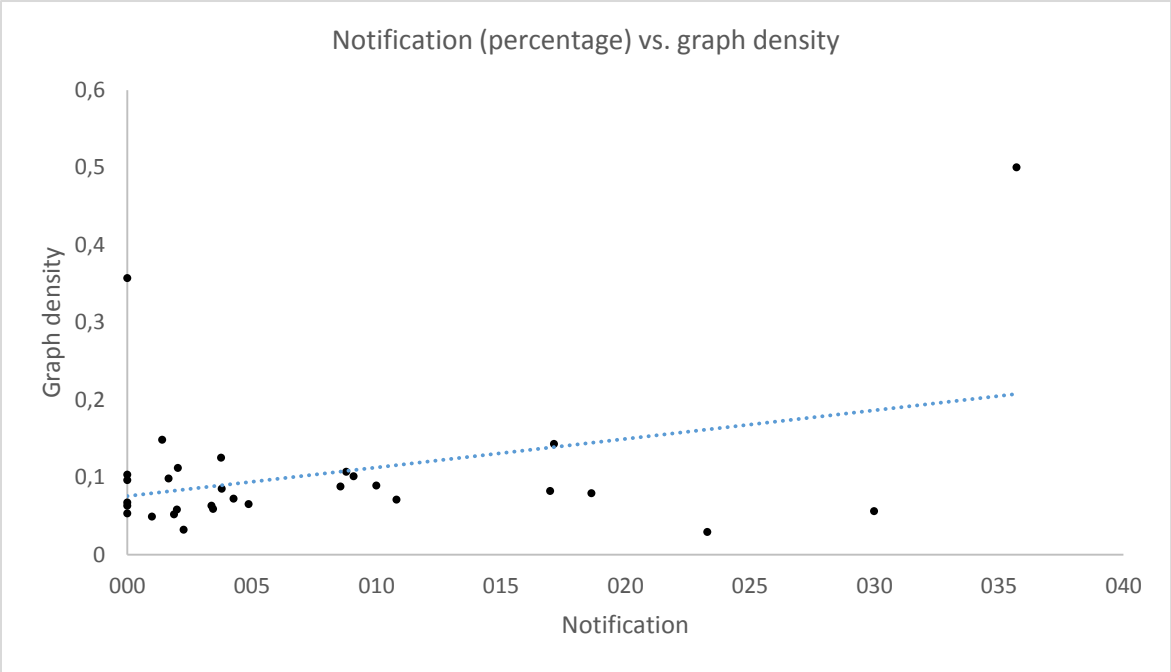

Figure 4.14 - Notification (percentage) vs. graph density (scatterplot)

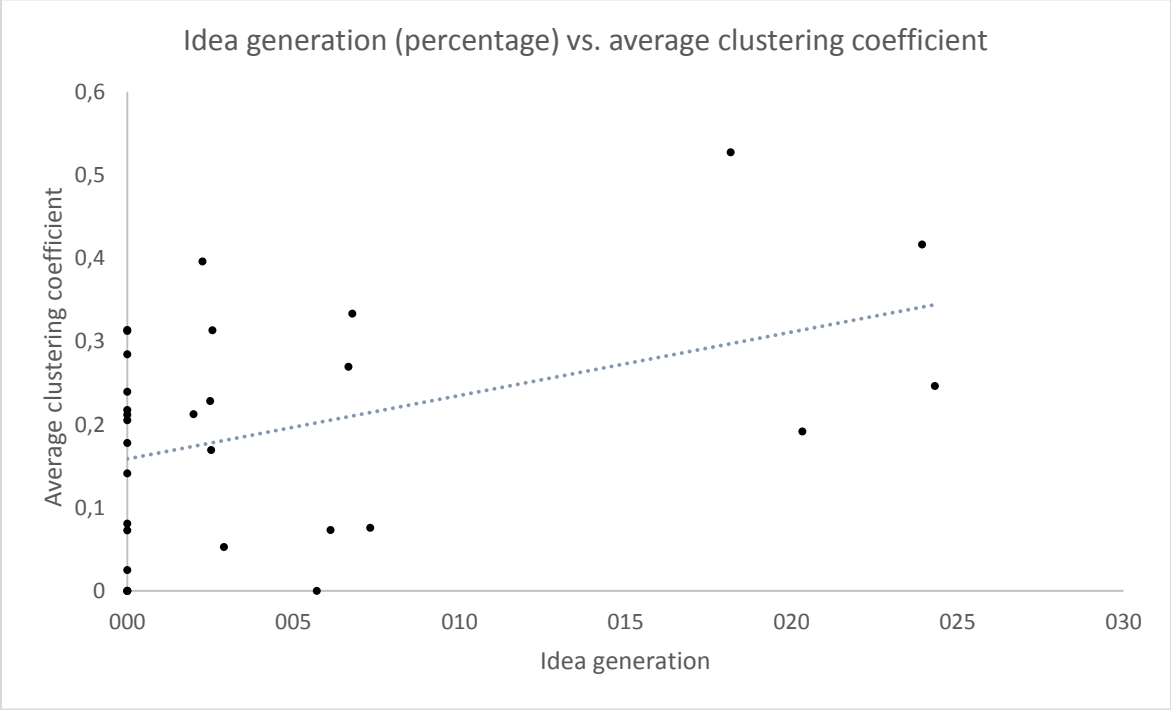

Figure 4.15 – Idea generation (percentage) vs. graph density (scatterplot)

By further visual inspection of Figure 4.14, an extreme value of the graph density was observed which required further evaluation of the correlation coefficient. The extreme value in this figure reports a graph density of 0.5, while the mean value of graph density in the dataset is 0.1034. To inspect how much effect this data point had on the correlation coefficient, the data point was removed and a new correlation analysis was run. After removing the data point, no significant correlation was observed between the amount of Notification-messages and the graph density in

the temporal networks (r(28) = -.175, two-tailed). Furthermore, by removing the data point, it was observed that the direction of the correlation was reversed from positive to negative. Hence, observing such an effect by simply removing one data point should therefore be incorporated into the interpretation of the results depicted in Figure 4.14.

Overall, Notification was the genre that occurred most often as having significant correlations with the structural characteristics of the temporal networks. According to the results depicted in Table 4.3, networks that are observed having a relatively low degree of clustering, and short path length between the nodes in the network, are also the same networks that have a relatively low amount of these types of messages.

Along with the amount of Discussion-messages, the amount of Idea generation-messages were found to have a significant, positive correlation with the degree of clustering in the temporal networks. However, by inspecting the scatterplots in Figure 4.15 and in Figure 4.10, it is apparent that the correlation between Idea generation-messages and the degree of clustering is more dependent on a small set of extreme values, as opposed to the correlation between Discussion-messages and the degree of clustering, where the data points seemed to be more evenly distributed across the trend line. Hence, as with the observations made regarding the correlation between the variables Notification and graph density, it was important to adjust the level of confidence in the observed results accordingly.

4.3 The relation between the structural characteristics and conversational nature of ABX actors

This sub-chapter will aim to answer the first half of RQ2 by analyzing the relationship between what the different actors talk about, and their centrality within each group in ABX. As described in Chapter 3, these relationships were explored using multiple linear regression with permutations as described in Hanneman and Riddle (2005), which takes into account the non-independence of the attributes of actors in a social networks such as the one studied in this research.

A summary of the results are depicted in Table 4.4, where the columns represent each of the different centrality metrics upon which the genre categories have been regressed. The values in the first sub-column ("Predictor p-values") only reports the p-values from the two-tailed permutation tests, which represent the proportion of permutation tests that resulted in a regression coefficient as high as or higher than the observed coefficient calculated without permutations. As an example, the p-value of the predictor Idea generation when regressed on the dependent variable Closeness centrality in ABX1 is 0.66553. Since the regression analysis was run using 10,000 permutations, this number reflects that 66.55% of the permuted coefficients were as high as or higher than the coefficient found when applying standard multiple linear regression without permutations, i.e. ~ 6655 results out of 10,000 yielded those results. This implies that the correlation between Idea generation and Closeness centrality, when controlling for all the other predictors, cannot be said to be significant as the same, or higher correlation coefficients can be

expected by randomly shuffling the observations of the variables Idea generation and Closeness centrality. The second sub-column ("Pearson's r") reports the correlation coefficient obtained when individually correlating each genre category with each of the centrality metrics. To sum up, the value in the first sub-column under each centrality metric reflects the significance of each genre category as a *predictor*, i.e. its significance obtained when altering its value while holding all other predictors constant, while the second sub-column only reflects the strength of the relationship between each genre and centrality metric *independently* of all other genre categories. By combining these two values, one obtains a more nuanced impression of how each genre category is related with each centrality metric, as opposed to only applying one of the aforementioned methods to describe these relationships.

As stated in chapter 3, not all genre categories, depending on their number of occurrences among the users for which the regression analysis was carried out on, were included as predictors in the formula. When running the regression analysis, predictors that are absent (zero occurrences) such as the Social category in numerous cases (e.g. for users in ABX1), were exempted as they do not contain any values that can be permuted. Hence, the regression analysis employed using the *ape*-package in R does not allow for such predictors to be part of the model specified using the *lmorigin*-command, and were dropped from the set of predictors included in the model.

When looking at the whole network, i.e. the network constructed when not partitioning the messages into groups, the following significant relationships are observed;

- A negative relationship between the variables Discussion and Closeness centrality ($r(251) = -.135$, $p = .05$, two-tailed)
- A negative relationship between the variables Idea generation and Closeness centrality ($r(251) = -.133$, $p = .05$, two-tailed)
- A negative relationship between the variables Praise and Eigenvector centrality ($r(251) = -.124$, $p = .05$, two-tailed)

		Betweenness centrality		Closeness centrality		Eigenvector centrality		Weighted in-degree centrality		Weighted out-degree centrality	
		Predictor p-value	Pearson's r	Predictor p-value	Pearson's r	Predictor p-value	Pearson's r	Predictor p-value	Pearson's r	Predictor p-value	Pearson's r
Whole network (n = 253)	**Discussion**	.450	-.016	.255	**-.135***	**.049***	.064	.237	.000	.176	.025
	Idea generation	.645	-.002	**.082**	**-.133***	**.061**	.077	.251	.047	.277	.043
	Status update	.166	.067	.573	.035	**.039***	.085	.103	.082	.131	.063
	Problem solving	.419	.028	.919	.064	.261	-.031	.293	.027	.263	.025
	Notification	.339	.020	.600	.051	.338	-.038	.429	-.009	.540	-.030
	Information input	.147	.096	.539	.075	**.018***	.076	.239	.040	.262	.033
	Praise	.832	-.086	.801	.062	.497	**-.124***	.835	-.072	.740	-.075
	Social	.933	-.036	.927	.020	.765	-.050	.951	-.037	.926	-.043
	Other	.512	-.011	.786	.030	.220	-.011	.471	-.012	.424	-.012
ABX1 (n = 51)	**Information input**	.717	.033	.151	.164	.728	.021	.975	.019	.977	.015
	Discussion	.880	-.073	.312	-.079	.771	-.029	.970	-.067	.981	-.045
	Problem solving	.301	.102	.742	-.012	.804	-.001	.322	.111	.387	.097
	Idea generation	.666	.019	.368	.054	.985	-.032	.658	.033	.886	.009
	Status update	**.068**	.178	.208	.125	**.066**	.238	**.093**	.174	**.074**	.225
	Notification	.311	.088	.332	.058	.338	.133	.538	.050	.634	.037
	Praise	.797	-.049	.463	-.023	.875	-.073	.629	-.040	.503	-.056
ABX2 (n = 56)	**Information input**	**.091**	**.319***	.380	.237	.199	**.324***	.118	.217	.118	**.415****
	Discussion	.773	-.170	.700	.049	.678	-.190	.807	-.152	.803	-.127
	Idea generation	.680	-.082	**.009****	**-.520****	.462	-.073	.541	-.077	.551	-.132
	Status update	.398	.033	.982	.110	.992	.007	.477	.017	.485	-.026
	Notification	.151	.227	.535	-.078	.155	**.343***	**.069**	**.284***	**.073**	.134
	Praise	.680	-.099	.624	.182	.418	-.132	.657	-.079	.665	-.095
	Other	.852	-.048	.266	-.121	.481	-.113	.975	-.049	.974	-.079
ABX3 (n = 79)	**Information input**	.200	.122	**.013***	**-.301****	.123	.219	.171	.139	**.079**	.095
	Discussion	.618	.043	.354	-.015	.923	-.039	.699	.008	.402	.055
	Problem solving	.518	-.026	.301	.095	.996	.083	.619	.003	.332	.019
	Status update	**.052**	**.231***	.908	.053	**.052**	**.222***	**.037***	**.269***	**.037***	.219
	Notification	.847	-.107	.400	-.004	.811	-.122	.985	-.081	.518	-.152
	Praise	.857	-.039	.623	.104	.937	-.066	.931	-.060	.509	.006
ABX4 (n = 70)	**Information input**	.922	.128	.737	.117	.495	**.241***	.970	.177	.951	.074
	Discussion	.307	.016	.338	-.135	.474	-.131	.134	-.081	.850	.005
	Problem solving	.341	.017	.745	.097	.971	.210	.302	.080	.926	.096
	Status update	.611	.035	.174	-.189	.935	.029	.563	.035	.891	.014
	Praise	.135	-.179	.759	.067	.223	**-.238***	**.098**	-.162	.407	-.170
ABX5 (n = 43)	**Information input**	.135	-.027	.267	-.149	**.018***	.116	**.073**	-.032	**.086**	-.011
	Discussion	.116	-.110	.423	.137	**.008****	**-.332***	**.066**	-.144	**.071**	-.136
	Problem solving	.164	.046	.308	-.154	**.017***	.080	**.088**	.052	**.095**	.048
	Idea generation	.191	-.002	.586	.077	**.041***	.054	.137	.030	.129	.000
	Praise	.677	.095	.286	-.108	.354	.100	.509	.098	.472	.075
	Notification	.975	.157	.939	.083	.559	.179	.774	.170	.858	.179
	Other	.256	.051	.845	.090	.197	.234	.194	.076	.180	.060

Table 4.4 – Results from multiple linear regression analysis with permutations (number of permutations = 10000) and Pearson's r correlation analysis (*-significant, α = .05, two-tailed; **-significant, α = .01, two-tailed)

- A positive relationship between the predictor Discussion and the independent variable Eigenvector centrality when controlling for all other predictors (p = .05, two-tailed)
- A positive relationship between the predictor Status update and the independent variable Eigenvector centrality when controlling for all other predictors (p = .05, two-tailed)
- A positive relationship between the predictor Information input and the independent variable Eigenvector centrality when controlling for all other predictors (p = .05, two-tailed)

When looking at the whole network, none of the relationships between genres and centrality metrics were significant in *both* the MLR analysis *and* the correlation analysis. Discussion and Idea generation were individually negatively correlated with the closeness centrality of users in the network as a whole, implying that users with a lower closeness centrality in the network are characterized by a higher, relative amount of these two genres in their messages, and vice versa. The same observation can be made on the relationship between the variables Praise and Eigenvector centrality. When controlling for all the other predictors, the relative amount of Status update and Discussion in a user's appears to act as predictors onto the dependent variable Eigenvector centrality, i.e. the eigenvector centrality of users appears to rise as the amount of Status update and Discussion in their messages rise while holding the other genre categories constant.

For ABX1, none of the genre categories were found to have a significant relationship with any of the centrality metrics neither in the correlation analysis or the multiple linear regression analysis. Four of the five centrality metrics did show a noticeable relationship with the variable Status update as a predictor, but these relationships were only significant with α = .10, which implies a more narrow confidence interval than what has been employed in testing for significance in this research.

The analysis of the ABX2 network revealed one genre that was significant both in the correlation analysis and the regression analysis, and five genres that were significant only in the correlation analysis. These significant relationships consisted of;

- A positive relationship between the variables Information input and Betweenness centrality ($r(54)$ = .319, p = .05, two-tailed)
- A positive relationship between the variables Information input and Eigenvector centrality ($r(54)$ = .324, p = .05, two-tailed)
- A positive relationship between the variables Information input and Weighted out-degree centrality ($r(54)$ = .415, p = .01, two-tailed)
- A positive relationship between the variables Notification and Eigenvector centrality ($r(54)$ = .343, p = .05, two-tailed)
- A positive relationship between the variables Information input and Weighted in-degree centrality ($r(54)$ = .284, p = .05, two-tailed)
- A negative relationship between the variables Idea generation and Closeness centrality ($r(54)$ = -.520, p = .01, two-tailed)
- A positive relationship between the predictor Idea generation and the independent variable Closeness centrality when controlling for all other predictors (p = .01, two-tailed)

It appeared that in ABX2, Information input was the most noticeable genre in terms of relationships with the centrality of the actors. Information input was not significant as a predictor in the regression analysis, but only when individually correlated with the different centrality metrics. Moreover, all the relationships between Information input and the centrality metrics were positive, indicating that actors whose messages fell into this category more often, also were the most central. Specifically, the results imply that actors with a higher relative amount of Information input, are more often on the shortest path between other actors in the network (betweenness centrality), more often are connected to other actors who themselves are central (eigenvector centrality), more often the most active (weighted out-degree centrality), and more often the most popular (weighted in-degree centrality). Another result worth noting was the relationship between Idea generation and Closeness centrality. This genre was negatively correlated with the closeness centrality of an actor in the Pearson product-moment correlation analysis, but positively associated when included in the regression analysis. In other words, when controlling for all other genres (keeping their values constant), an increase in the value of the variable Status update was associated with an increase in the closeness centrality of an actor in ABX2. When ignoring all other genres, i.e. correlating Idea generation with Closeness centrality individually, this relationship appeared to be negative.

The results from running the correlation analysis and MLR analysis from groups ABX3, BX4, and ABX5 can be read and interpreted in the same manner as stated above. Table 4.5 shows a compilation of significant results found in the analysis.

Group	Genre	Centrality metric	Correlation analysis (CA)/Regression analysis (MLR)	Significance level (two-tailed, negative correlations in parentheses)
Whole network	Discussion	Closeness centrality	CA	(.05)
Whole network	Discussion	Eigenvector centrality	MLR	.05
Whole network	Idea generation	Closeness centrality	CA	(.05)
Whole network	Status update	Eigenvector centrality	MLR	.05
Whole network	Information input	Eigenvector centrality	MLR	.05
Whole network	Praise	Eigenvector centrality	CA	(.05)
ABX2	Information input	Betweenness centrality	CA	.05
ABX2	Information input	Eigenvector centrality	CA	.05
ABX2	Information input	Weighted out-degree	CA	.01
ABX2	Idea generation	Closeness centrality	CA/MLR	(.05)
ABX2	Notification	Eigenvector centrality	CA	.05
ABX2	Notification	Weighted in-degree	CA	.05
ABX3	Information input	Closeness centrality	CA/MLR	(.05)
ABX3	Status update	Betweenness centrality	CA	.05
ABX3	Status update	Eigenvector centrality	CA	.05
ABX3	Status update	Weighted in-degree	CA/MLR	.05
ABX3	Status update	Weighted out-degree	MLR	.05
ABX4	Information input	Eigenvector centrality	CA	.05
ABX4	Praise	Eigenvector centrality	CA	(.05)
ABX5	Information input	Eigenvector centrality	MLR	.05
ABX5	Discussion	Eigenvector centrality	CA/MLR	(.05)
ABX5	Problem solving	Eigenvector centrality	MLR	.05
ABX5	Idea generation	Eigenvector centrality	MLR	.05

Table 4.5 – Overview of significant results from correlation analysis and MLR analysis

4.4 The overlap of central users within ABX groups

This sub-chapter will report on the findings from studying the overlap between the different central roles in each of the groups. In the result as presented in Table 4.5, the numbers in the cells represent the percentage-wise overlap between the top users according to the respective centrality metric noted in the rows of each table, and the respective centrality metric noted in the columns of the table. Thus, the intersecting cells between the same centrality metrics will always equate to 100. Furthermore, each centrality metric is partitioned into two sub-labels; top 5 percent and top 10 percent. These numbers represent, respectively, the overlap between the top 5 percent and the top 10 percent of users within each centrality metric. For presentational purposes, the top 10 percent of overlap-values are highlighted in green in each of the tables. To give a more specific example on how the tables should be read: if the top 10 percent of users according to eigenvector centrality consists of user A, B, C, and D, and the top 10 percent of users according to closeness centrality in the same group consists of user B, C, E, and F, then the overlap in terms of percentage would be 50 % (user B and C).

Table 4.5 presents the mean values for the overlap between users in all five groups. The overlap analysis for each group on which the values in Table 4.5 are based on can be found in Appendix

D. The largest overlap can be found between the top 5 percent of users according to eigenvector centrality, and the top 10 percent of users according to weighted in-degree centrality. This implies that, on average, more than nine out of ten users who are in the top 5 percent according to eigenvector centrality, can be found amongst the top 10 percent according to weighted in-degree centrality within the same group. The top 5 percent according to eigenvector centrality are also the users who, on average, have the highest degree of overlap with all the other centrality metrics (76.97 %). The second largest overlap, on average, is between the top 5 percent according to weighted out-degree centrality, and the top 10 percent of users according to betweenness centrality (95.48%). It should be noted that none of the intersecting overlaps have a value below 33%, indicating that in any intersecting overlap between centrality metrics, at least 1/3 of the top users according to the centrality metric stated in the row names are also present in the list of top users according to the centrality metric stated in the column names. Another notable observation is that the average overlap between all lists of top users according to all centrality metrics is 63.64%, implying that, on average, ~2/3 users can be found to be redundant across any two top lists of users according to any two combinations of centrality metrics.

All groups combined		Weighted in-degree centrality		Weighted out-degree centrality		Betweenness centrality		Closeness centrality		Eigenvector centrality		
		Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Average
Weighted in-degree centrality	Top 5			68.33	78.89	67.90	88.25	58.81	74.64	73.53	93.81	75.52
	Top 10			40.89	65.60	41.45	70.03	34.89	59.62	46.55	81.08	55.01
Weighted out-degree centrality	Top 5	68.33	85.00			72.34	95.48	58.81	73.53	62.42	82.14	74.76
	Top 10	37.78	65.60			43.33	70.25	34.45	63.17	39.01	62.37	52.00
Betweenness centrality	Top 5	67.90	86.03	72.34	89.92			58.73	73.25	67.90	84.37	75.05
	Top 10	42.50	70.03	46.05	70.25			36.77	59.17	41.95	66.35	54.13
Closeness centrality	Top 5	58.81	72.22	58.81	71.11	58.73	76.11			66.67	81.94	68.05
	Top 10	36.00	59.62	35.56	63.17	35.29	59.17			38.78	68.51	49.52
Eigenvector centrality	Top 5	73.53	96.67	62.42	81.11	67.90	87.22	66.67	80.28			76.97
	Top 10	45.22	81.08	39.56	62.37	40.61	66.35	39.62	68.51			55.42
Average		53.76	77.03	53.00	72.80	53.44	76.61	48.59	69.02	54.60	77.57	**63.64**

Table 4.6 - Overlap between central positions (mean values of all groups)

4.5 The overlap of central users between ABX groups

This sub-chapter will report on the analysis of the redundancy of central users *across* groups. As the previous sub-chapter addressed the redundancy of users across the different central positions within the same group, the results presented here will depict how users who held a central position in at least one group, also were represented among the central users in their other groups. The number of groups a user is part of was not taken into account when calculating the values presented in this sub-chapter, i.e. if a user was part of the top 10 percent users according to betweenness centrality in one group, and was also amongst the top 10 percent of users according to the same centrality measure in one more group, then this user's score would be 1, regardless of how many groups the user had been active in. If the user held a central position in one group, and was not a member of any other groups, then the user would not be included when calculating the

average overlap across groups for central users. Thus, the base of users for which the overlap values were calculated only include a subset of users. This subset consisted of users who were central in at least one group, *and* had been active in at least one more group. Users who had only been active in one group *or* had been active in more than one group but held no central positions in any of them, were ignored.

Table 4.7 presents the results from the analysis conducted. As with the tables presented in the previous sub-chapter, the top 10 percent of overlaps are highlighted in green. Each number in each cell represents, on average, the count of other groups that a user held a central position in, *in addition* to their other group in which the user also held a central position. As can be read from Table 4.7, the highest inter-group overlap of central users can be found between the top 5 percent of central users according to weighted out-degree centrality, and the top 10 percent of central users according to the same centrality metric (1.37). This result implies that, on average, users that were part of at least one group, and were amongst the top 5 percent of users according to weighted in-degree centrality in any of their groups, were also present amongst the top 10 percent of users according to the same centrality metric in 1.37 *more* groups.

The second largest overlap between groups were between the top 5 percent of users according to betweenness centrality and the top 10 percent of users according to weighted out-degree centrality (1.30), followed by the overlap between weighted out-degree centrality (top 5 percent) and weighted in-degree centrality (top 10 percent)(1.26). On average, users who were part of the top 5 percent of users according to weighted out-degree in at least one of their groups were the most likely to appear among the central users according to any centrality metric in their other groups, with an average overlap value of 1.05. The lowest average overlap value could be found among users who were part of the top 5 percent users according to weighted in-degree centrality, who had the least likelihood of possessing other central positions in their other groups (0.63).

It was noteworthy that the average overlap between all centrality metrics was 0.80, implying that users who held any central position in one of their groups also held a central position in ~1 more group. This result entails that central users tended to be redundant with respect to their centrality, and were not likely to hold a central position in one group alone.

		Weighted in-degree centrality		Weighted out-degree centrality		Betweenness centrality		Closeness centrality		Eigenvector centrality		
		Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Average
Weighted in-degree centrality	Top 5	0.29	0.81	0.57	0.86	0.62	0.76	0.57	0.81	0.38	0.67	0.63
	Top 10	0.50	0.88	0.71	0.94	0.68	0.82	0.74	0.97	0.53	0.79	0.76
Weighted out-degree centrality	Top 5	0.63	1.26	0.95	1.37	0.95	1.26	0.95	1.26	0.68	1.16	1.05
	Top 10	0.50	0.89	0.72	0.94	0.72	0.92	0.72	0.94	0.56	0.83	0.78
Betweenness centrality	Top 5	0.62	1.10	0.86	1.24	0.86	1.14	0.81	1.10	0.62	1.05	0.94
	Top 10	0.42	0.74	0.63	0.87	0.63	0.84	0.55	0.71	0.42	0.66	0.65
Closeness centrality	Top 5	0.60	1.25	0.90	1.30	0.85	1.05	0.80	1.15	0.65	1.00	0.96
	Top 10	0.49	0.94	0.69	0.97	0.66	0.77	0.66	0.86	0.54	0.74	0.73
Eigenvector centrality	Top 5	0.42	0.95	0.68	1.05	0.68	0.84	0.68	1.00	0.42	0.79	0.75
	Top 10	0.47	0.90	0.73	1.00	0.73	0.83	0.67	0.87	0.50	0.73	0.74
Average		0.49	0.97	0.74	1.05	0.74	0.92	0.71	0.97	0.53	0.84	**0.80**

Table 4.7 - Centrality overlap amongst users

Chapter 5

Discussion

This chapter will discuss the results shown in the previous chapter within the scope of the stated research questions. The chapter will then conclude with a discussion about the limitations that were identified during this project.

5.1 RQ1: What, if any, is the relation between the conversational nature in an ESN and its network structure?

The first main research question in this project aimed at analyzing the relation between the conversational nature and the structural characteristics of ESN networks. The method employed to accomplish this was to create snapshots of all the five groups over a set time period. Selected SNA metrics where then applied to describe the structure of each of the networks in these snapshots, and a genre analysis to describe the conversational nature of each network. A correlation analysis was then conducted to discover any linear relationships that might exist between the two constructs.

Two of the SNA metrics applied did not appear to have any significant relationships with any of the genre categories; average weighted degree centrality and reciprocity. I.e., it seems that the mean sum of weighted ties among all actors in the ESN networks is not related to the conversational nature in the networks, and the same goes for the proportion of reciprocated ties in the network.

One interesting result was the strong positive correlation between the relative amount of discussion-messages and the clustering in a network. Discussion-messages have in previous studies been shown to be one of the dominating genres in ESNs (K. Riemer & Richter, 2012; K. Riemer, Scifleet, et al., 2012), implying that these networks are primarily used for discussing opinions and ideas within the organization. The high correlation between this genre and the degree of clustering adds to these results, as it also seems that the overall interconnectedness between the users seem to be higher in networks where this genre is more dominant. Although this research paper has not been able to find any research on exactly *what* positive or negative effects a high clustering coefficient might have in an ESN, it is not unreasonable to assume that a high degree of clustering is beneficial from an organizational standpoint. This because the coefficient itself can be seen as an expression of how well-connected the users are in the network seen as a whole. Say, if an organization decides that their primary goal for implementing an ESN is to increase the level of collaboration, it could be beneficial to actively encourage open discussions in order to facilitate this goal.

A positive correlation was also found between the relative amount of idea generation-messages and the degree of clustering in the network. It is interesting to read the results depicted in this and the previous paragraph in light of previous research related to the conversational nature in ESN networks. Riemer and Tavakoli (2013) compiled a set of group archetypes based on their research, and found that in networks with a high amount of discussion posts can be described as "*virtual water-coolers*", and networks with a relative high amount of idea generation-messages are typically "*networks of expertise*". From the results presented in this study, it seems as if both of these archetypes are characterized by having strongly interconnected users in the network. It is worth pointing out that the border between the two genre categories "Discussion" and "Idea generation" might be hard to draw as well, as an "Idea generation" conversation does necessarily require some form of discussion. The inherent overlap between these two genres could indicate that they tend to bear the same characteristics, which in turn might help explain why they are both strongly correlated with the clustering coefficient. Furthermore, these results might provide further weight to the validity of the archetypes described by Riemer and Tavakoli (2013), as a it seems that these sorts of networks tend to consist of users that are more tightly clustered.

Two other genre categories, Information input and Notification, were also found to be correlated with the degree of clustering in the network, but then in the form of a negative correlation. A possible interpretation of these results can be that these two genres are inherently forms of broadcasting, and do not directly encourage an exchange of opinions or ideas. Typical messages of these types are just informative, short announcements of some sort, or short messages just containing a link to a resource. Hence, they typically do not contain questions or any other form of inquiry, which is the case with discussion- and idea generation-messages. Furthermore, one can infer from these results that, given that having a strongly interconnected users in an informal network is in the best interest of an organization, that engaging in discussions and idea generation might assist in accomplishing this as opposed to using ESNs only for passive forms of information broadcasting.

The Notification-genre was also found to be negatively correlated with the average path length. The average path length provides an impression of how the actors in the network are connected. As discussed in the previous paragraph, networks with a high relative amount of notification-messages were found to be less tightly clustered. However, it appears that the average distance between actors is smaller in these networks as well. A possible inference from this observation is that in time periods with a high amount of notifications, one can expect to find a smaller user base but where the users are more densely connected. However, as was shown in sub-chapter 4.2 in this paper, the positive correlation between graph density and the relative amount of notification-messages was highly affected due to a single outlier. By removing this outlier it was demonstrated that the correlation both was negative instead of positive, and not significant. This specific result should therefore be interpreted with caution knowing the sensitivity of the correlation test with respect to outliers in the data.

A limitation of the correlation analysis used to bring forth these results, is that it only correlates each genre individually, and does not take into account the conversation pattern of the network

as a whole. Since the values for each genre category is presented as a percentage value, an increase or decrease of one genre will cause one or more other genres to also increase or decrease as a result. The genres were therefore also correlated with each other in order to see if there were such interdependencies present in the observations.

The results showed that several genre categories were correlated with each other. E.g. the genre "Discussion" was negatively correlated with the genres "Notification", "Problem solving", and "Status updates". This could provide grounds for saying that the correlation between the genre "Discussion" and the degree of clustering as discussed in the previous paragraphs might not be isolated to the "Discussion"-genre only, but it might be that a high degree of clustering is related to the genres "Discussion", "Notification", "Problem solving", and "Status updates" in concert. For the purpose of this research however, it would not make sense to engage in a detailed discussion about each one of these correlations. The reason is that the correlation analysis is inept when it comes to considering a pattern which consists of several variables at once such as with the genre categories. A regression- or clustering analysis would be more aptly suited for this purpose, but this would require a much larger set of observations in order to produce reliable results. However, the correlation analysis does provide an indication of the genres that are *most* strongly correlated with the SNA metrics and it should therefore, despite of its limitations, be seen as a valid indicator of how specific genres are related to the structural characteristics of a network.

5.2 RQ2: What are the key characteristics of central actors in an ESN with respect to their conversational nature and their redundancy across different central positions within and between groups?

RQ2 aims to investigate what conversational topics central actors tend to be characterized by, and to what extent these actors hold several central positions within and across groups. A statistical analysis was conducted in this project in order to determine the relation between the centrality of ESN users and the topics of their communication within the network. This analysis was carried out using two different approaches in conjunction with each other; 1) a multiple linear regression (MLR) analysis was conducted according to the steps outlined in Hanneman and Riddle (2005, p. 300), using permutations in order to account for the non-independence of relations between the actors in the network, 2) a Pearson's correlation analysis was conducted in conjunction with the MLR analysis (Nathans et al., 2012) to add more nuance to the interpretation of the relation between message topics and centrality. Both the MLR- and the correlation analysis were run both on the group level as well as the network-level in order to investigate if and how the relation between the two constructs might deviate depending on the context.

Regarding the conversational nature of actors and their centrality within the ABX network as a whole, the result have shown that actors with a low closeness centrality typically have a high relative amount of discussion- and idea generation-messages in their communication. These results were brought forth by a correlation analysis between these genres and the closeness centrality score of actors in the network. The MLR analysis produced a slightly different

impression, as actors with a high relative amount of discussion-, status update-, and information input-messages were found to have a high eigenvector centrality in the network. These results a in a sense more directly relevant in describing the characteristics of central actors, since the MLR analysis controls for all of the other genres when determining whether there is a relation between what users talk about and their centrality. Hence, with respect to central actors according to their eigenvector centrality, these actors are typically characterized by;

- active involvement in discussions where the conversation revolves around exchanging personal opinions
- active involvement in updating other actors in the network about projects and other ongoing work-related process within the organization
- active involvement in distributing information about work-related resources which other actors in the network can make use, or to contribute with facts as part of an ongoing discussion

The same type of analysis as described above was also carried out on the level of the groups. The results showed that in this context, the relation between message genres and centrality metric is highly dependent on the context of the group. E.g. in ABX1, none of the genres appeared to have a significant relation to any of the centrality metrics for users, while in all the other four groups there were several such relations found. In ABX2, the information input-genre appeared to be positively correlated with three centrality metrics; eigenvector centrality, weighted out-degree centrality, and betweenness centrality. This genre was also positively correlated with eigenvector centrality in ABX4 and ABX5. In ABX3, the amount of information input-messages were negatively correlated with closeness centrality. It was an interesting observation that there was a relation between closeness centrality and a specific genre in two more groups, but in each of these cases they were negative. This implies that in any case where a genre can be used to either predict the closeness centrality of an actor, or be used to individually inspect the relation between a genre and the closeness centrality score of an actor, one usually finds that there is a negative relation and that these users are not centered close to the middle of the network. Of the genres that were most often found to have a relation with a centrality score, information input was the most prominent one, which was present in all of the four groups were a significant relation was found. This observation is not without relevance for the context of this project. As Riemer & Tavakoli (2013) argued in their research; "*information sharing groups can be seen as the first evolutionary step of emerging groups(...)*". Thus, within the local context of groups, the users who contribute with this type of communication can also be found to be situated in a central position within the group in one way or another, and play a central role in how the group develops over time.

It is necessary to point out the caveats that comes with the analysis conducted on the level of the groups, especially with respect to the MLR analysis. As the number of actors included in the MLR analysis for the groups ABX1, ABX2, ABX3, ABX4, and ABX5 were 51, 56, 79, 70, and 43, respectively, there are considerations about the sample size that should not go unnoticed. There are different "rules-of-thumb" that exist when considering the sample size that is needed in order to conduct a reliable MLR analysis, all of which differs from each other but are nonetheless

important to discuss in order to get a general impression of how well the MLR analysis conducted in this project falls within these recommendations. As an example, Green (1991) suggested that a sample size N should be larger than 104 + m, where m is the number of individual predictors. For example, in the case of ABX1, the sample size should exceed 104 + 7 = 111, which clearly is not the case as the sample size in this group is only 51, roughly half of what Green (1991) suggests. However, other "rules of thumb" have been suggested which might draw a different picture. Strauss and Harris (1975) suggested that there should be 10 observations per predictor, which in the case of ABX1 would indicate that the sample size should be 70. In the specific case of ABX1 this requirement is still not met, but it is however with respect to ABX3 and ABX4, where the number of sampled users are higher. Hence, it is clear that the sample size present in each of the groups are in the grey-zone when it comes to what is necessary in order to produce an MLR analysis that can be interpreted with a certain degree of confidence. However, the main focus of the analysis conducted was on the ESN in its entirety, where the sample size (N = 253) was notably higher and therefore can be expected to produce more reliable results.

An overlap analysis was conducted within each group between the different central positions in order to find the redundancy of users across these different positions. The results from each of the groups were then compiled into a single table in order to find the mean overlap across all groups in the network.

The results showed a relatively high degree of overlap between all intra-group central positions. On average, the results showed a 63% overlap between the central positions within the groups. The highest overlap was found between the top 5% of users according to eigenvector centrality and all other central positions, where on average, 76.97% of these users could be found holding at least one of the other four central positions in the network. Users who were among the top 10% according to closeness centrality were found to be the least likely to also hold other central positions in the network, with an average overlap of 49.53%. The single biggest overlap was between the top 5% of users according to eigenvector centrality, and the top 10% of users according to weighted in-degree centrality. These results imply that the actors that together occupy the most important positions in the network constitutes a relatively small part of the total user-base. It would therefore not be unreasonable to refer to these users as a sort of "elite" who, although small in numbers, has the access and power to control much of the agenda within these networks. It is also interesting to see these results in the light of what was discussed in the previous section, since actors with a high eigenvector centrality were found to be most active in the other central positions as well. From this it could be inferred that eigenvector centrality is not only a suitable centrality metric when it comes to identifying actors who play a central role in terms of the information flow within the network, but also to identify users who are more likely to occupy several other central positions in the network.

In the same way as was done with analyzing the overlap of central actors within a group, an overlap analysis was also conducted between groups. This analysis was carried out by identifying users who were not only active in more than one group, but that also occupied central positions in more than one group.

The results showed that, on average, actors who held a central position in at least one group also held a central position in 0.8 (-1) more groups. The highest overlap was found between users who were among the top 5% according to weighted out-degree centrality in at least group, and among the top 10% of users according to the same centrality metric in other groups. These users were also found to be the most likely to occupy any of the other central positions in other groups, as the results showed that on average, these users could be found holding central positions in 1.05 other groups. It was also an interesting observation that actors did not necessarily hold the same central position in all the groups they were active in, in fact this was most often not the case. This is not surprising seen in light of the results presented in the previous section, as actors seem to be redundant across several central positions within the same group, and it is therefore reasonable to expect the same observation being made across groups.

5.3 Limitations

As is typical with any research, this project was also confined by certain limitations. First, with respect to the data collection, the dataset only included ESN data from a single company. As Richter and Riemer (2013) showed, the use of ESN can vary greatly depending on the context of the organization in which the ESN is deployed. Thus, the results presented in this project should be read within the context of ABX and does not necessarily allow for generalizations outside of this context. However, the ESN consisted of actors in a large variety of departments and from several countries. For similar types of multi-national consultancy companies for whose business context is similar it would therefore be reasonable to expect similar results. Second, only five groups were included in the dataset. As shown in previous research (e.g., Berger et al., 2014), some ESNs can consist of hundreds of groups, which would also have been ideal in this research as a portion of the analysis was conducted on the group-level. Third, the inferences about the importance of actors are purely reliant on their online communication and their position within the informal network. It should be stressed that employees' online behavior is not necessarily a 1:1 mapping of their "offline" importance within the organization, and should not be interpreted as such. For a more wholeheartedly evaluation of the central positioning within an organization, a researcher would also need to gather data about how employees communicate with each other outside of the ESN, which could produce a different picture of who is important with respect to the information flow within the company.

Chapter 6

Conclusion and Future work

6.1 Conclusion

This project has explored the state-of-the-art of genre analysis and SNA of ESNs, the relation between the conversational nature and structural characteristics of the ABX ESN, and the characteristics of central users regarding their conversational nature and redundancy across central positions within and across ESN groups.

Based on the results brought forth in this project, it can be concluded that several structural characteristics ESNs are related to several individual genres when looking at an ESN over time. With respect to the degree of clustering in an ESN network in a specific time-period, this characteristic is positively correlated with the relative amount of discussion-messages and idea-generation messages, and negatively correlated with the relative amount of information input- and notification messages. From the definitions applied in the genre repertoire, one can infer that in time periods where employees engage more in exchanging opinions and discussing certain topics one also finds that the employees are more tightly clustered in the informal network. The same inference holds for time-periods where employees engage more in discussing new ideas and having conversations that are more creative in nature. In time-periods where the conversations on the network-level are more characterized by the broadcasting of short messages including links to resources etc., and by notifications about events and other happenings, employees are less tightly clustered. In time-periods where the conversations are characterized by a relative large amount of notification-messages, the networks also have a higher density and a lower average path length. I.e. a higher proportion of possible links are present in the network, and the average geodesic distance between the actors in the network are shorter.

Of all the centrality metrics applied in measuring the centrality of actors in the network, eigenvector was the only one that was found to have a significant relation with individual genres when controlling for all other genres by conducting a MLR analysis. The eigenvector centrality of an actor can, according to the results, be predicted by the following three genres; Discussion, Information input, and Status updates. Three distinct characteristics can therefore be found among actors with a high eigenvector centrality; they often engage in the exchange of personal opinions, they often provide links to resources of professional value for other employees to make use of, and they often provide updates about the current status of ongoing projects and alike. Moreover, central users in an ESN can be characterized by occupying several other central positions simultaneously. Again, actors who have a high eigenvector centrality are the most likely to occupy other central positions in the network, which points towards that important actors in

the network is typically characterized by being connected to a high number of other central actors.

In summary, this project has made the following contributions to the body of knowledge related to the study of ESNs;

- As far as the literature review conducted in this project has been able to discover, this is the first research to combine genre analysis and SNA in order to discover how the conversational nature of an ESN relates to its structural characteristics. An understanding of how these two constructs are related can help decision-makers in the implementation and monitoring of ESNs in the future, as this technology can be expected to gain increasing interest among businesses and organizations in the time to come.

- This project has shown how central users are characterized by their conversational profiles. Understanding not only who are the important players in an informal network, but how they communicate with other employees, is vital information for any manager or decision-maker that wishes to derive insights about how these important actors can be identified. As the importance of ESN platforms can be expected to rise, it can also be reasonably assumed that acquiring more knowledge about the actors within them will rise.

- The results regarding the redundancy of users across central positions in an ESN provide insights about the power and influence is distributed within these networks. The results have shown that there is a relatively large overlap between the central positions, which is useful information for organizations in analyzing the internal structures of its ESN. E.g. in assessing the health of an ESN, it is valuable for decision-makers to realize that the activity within the networks might be sustained only by a fraction of the total user base, and that if these actors were to disappear it could potentially have damaging effects for the sustainability of the ESN.

- This project has shown that central actors are normally not redundant across groups, but are rather highly active within a single community or group. This is valuable to know with respect to building an ESN community within an organization, as each group revolving around a different topic can typically be expected to be comprised of a different user base than other groups. The valuation of central users therefore needs to be conducted within the context of each group, and not necessarily on the level of the whole network.

6.2 Future work

This study has mainly focused on exploring the characteristics of ESN users and networks within the context of SNA and genre analysis. One aspect that is lost when only SNA and genre analysis methods are applied are the organizational and personal factors that might play a role in how users are positioned in the network and what they talk about. Such factors could include job title and function, geographical location, departmental affiliation, age, gender, or other pieces of meta-data that are tied to each individual in an organization. Future studies should find a way to include these factors in order to create a more comprehensive understanding of the relationship between what users talk about, their characteristics, and their position within an ESN.

References

Adamic, L., Zhang, J., Bakshy, E., Ackerman, M., & Arbor, A. (2008). Knowledge Sharing and Yahoo Answers : Everyone Knows Something. In *Proceedings of the 17th International Conference on World Wide Web* (pp. 665–674). Beijing, China. doi:10.1145/1367497.1367587

Ahn, Y.-Y., Han, S., Kwak, H., Moon, S., & Jeong, H. (2007). Analysis of Topological Characteristics of Huge Online Social Networking Services. In *Proceedings of the 16th International Conference on World Wide Web* (pp. 835–844). Banff, Canada. doi:10.1145/1242572.1242685

Almack, J. (1922). The influence of intelligence on the selection of associates. *School and Society*, *16*, 529–530.

Angeletou, S., Rowe, M., & Alani, H. (2011). Modelling and analysis of user behaviour in online communities. *Lecture Notes in Computer Science*, *7031*, 35–50. doi:10.1007/978-3-642-25073-6_3

Archambault, A., & Grudin, J. (2012). A longitudinal study of facebook, linkedin, & twitter use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2741–2750). Austin, USA. doi:10.1145/2207676.2208671

Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(11), 3747–3752. doi:10.1073/pnas.0400087101

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. In *Third International AAAI Conference on Weblogs and Social Media* (pp. 361–362). San Jose, USA. doi:10.1136/qshc.2004.010033

Bellman, R. (1958). On a Routing Problem. *Quartlerly of Applied Mathematics*, *16*, 87–90.

Berger, K., Klier, J., Klier, M., & Richter, A. (2014). Who is Key?-Value Adding Users in Enterprise Social Networks. In *Proceedings of the 2014 European Conference on Information Systems* (pp. 1–16). Tel Aviv, Israel. Retrieved from http://ecis2014.eu/E-poster/files/0745-file1.pdf

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, *2*(1), 113–120. doi:10.1080/0022250X.1972.9989806

Bonacich, P., & Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, *23*(3), 191–201. doi:10.1016/S0378-8733(01)00038-7

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies. doi:10.1111/j.1439-0310.2009.01613.x

Brandes, U. (2001). A faster algorithm for betweenness centrality. *The Journal of Mathematical Sociology*, *25*(2), 163–177. doi:10.1080/0022250X.2001.9990249

Brandes, U., Lerner, J., & Snijders, T. A. B. (2009). Networks evolving step by step: statistical analysis of dyadic event data. In *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining* (pp. 200–205). Athens, Greece. doi:10.1109/ASONAM.2009.28

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1), 107–117. Retrieved from http://infolab.stanford.edu/~backrub/google.html

Butts, C. T. (2014). sna: Tools for Social Network Analysis. Retrieved from http://cran.r-project.org/package=sna

Butts, C. T., Handcock, M. S., & Hunter, D. R. (2014). network: Classes for Relational Data. Retrieved from http://statnet.org/

Cao, J., Gao, H. Y., Li, L. E., & Friedman, B. (2013). Enterprise Social Network Analysis and Modeling: A Tale of Two Graphs. In *Proceedings of the IEEE Infocom* (pp. 2382–2390). Turin, Italy. doi:10.1109/INFCOM.2013.6567043

Carley, K. M. (2003). *Dynamic Network Analysis* (pp. 133–145). Retrieved from http://www.chronicdisease.org/files/public/2009Institute_NA_Track_Carley_2003_dynamicnetwork.pdf

Cohen, J. (1960). A coefficient of agreement of nominal scales. *Educational and Psychological Measurement*, *20*, 37–46. doi:10.1177/001316446002000104

Csárdi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, *Complex Sy*, 1695. Retrieved from http://igraph.sf.net

De Laat, M., Lally, V., Lipponen, L., & Simons, R. (2007). Investigating patterns of interaction in networked learning and computer supported collaborative learning: a role for social network analysis. *The International Journal of Computer-Supported Collaborative Learning*, *2*(1), 87–103. doi:10.1007/s11412-007-9006-4

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, *1*(1), 269–271. doi:10.1007/BF01386390

DiMicco, J. M., & Millen, D. R. (2007). Identity Management. In *Proceedings of the 2007 international ACM conference on Supporting group work* (pp. 383–386). Sanibel Island, USA. doi:10.1145/1316624.1316682

DiMicco, J., Millen, D. R., Geyer, W., Dugan, C., Brownholtz, B., & Muller, M. (2008). Motivations for social networking at work. In *Computer supported cooperative work CSCW'08* (pp. 711–720). San Diego, USA. doi:10.1145/1460563.1460674

Efimova, L., & Grudin, J. (2007). Crossing boundaries: A case study of employee blogging. In *Proceedings of the 40th Hawaii International Conference on System Sciences* (p. 86a). Hawaii, USA. doi:10.1109/HICSS.2007.159

Ehrlich, K., Lin, C. Y., & Griffiths-Fisher, V. (2007). Searching for experts in the enterprise: combining text and social network analysis. In *Proceedings of the 2007 international ACM conference on Supporting group work* (pp. 117–126). Sanibel Island, USA. doi:10.1145/1316624.1316642

Fisher, D., Smith, M., & Welser, H. T. (2006). You Are Who You Talk To: Detecting Roles in Usenet Newsgroups. In *Proceedings of the 39th Hawaii International Conference on System Sciences* (Vol. 3, p. 59b). Hawaii, USA: IEEE. doi:10.1109/HICSS.2006.536

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). Wiley.

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, *40*, 35–41. doi:10.2307/3033543

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239. doi:10.1016/0378-8733(78)90021-7

Fulk, J., & Yuan, Y. C. (2013). Location, motivation, and social capitalization via enterprise social networking. *Journal of Computer-Mediated Communication*, *19*(1), 20–37. doi:10.1111/jcc4.12033

Halatchliyski, I., & Cress, U. (2014). How Structure Shapes Dynamics: Knowledge Development in Wikipedia - A Network Multilevel Modeling Approach. *PLoS ONE*, *9*(11).

Hanneman, R. A., & Riddle, M. (2005). *Introduction to social network methods*. Riverside, USA: University of California. doi:10.1177/000169939403700402

Heidemann, J., Klier, M., & Probst, F. (2010). Identifying key users in Online Social Networks: A PageRank Based Approach. In *Proceedings of the 31st International Conference on Information Systems* (Vol. 4801, pp. 1–22). St. Louis, USA. Retrieved from http://www.wi-if.de/paperliste/paper/wi-301.pdf

Holland, P. W., & Leinhardt, S. (1977). A dynamic model for social networks. *Journal of Mathematical Sociology*, *5*(1), 5–20. doi:10.1080/0022250X.1977.9989862

Huffaker, D. (2010). Dimensions of Leadership and Social Influence in Online Communities. *Human Communication Research*, *36*(4), 593–617. doi:10.1111/j.1468-2958.2010.01390.x

Johnson, D. B. (1977). Efficient Algorithms for Shortest Paths in Sparse Networks. *Journal of the ACM*, *24*(1), 1–13. doi:10.1145/321992.321993

Kazienko, P., Michalski, R., & Palus, S. (2011). Social Network Analysis as a Tool for Improving Enterprise Architecture. *Agent and Multi-Agent Systems: Technologies and Applications*, *6682*, 651–660. Retrieved from ‹Go to ISI›://WOS:000311841600067

Kossinets, G., & Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science*, *311*, 88–90. doi:10.1126/science.1116869

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology*. *Education* (2nd ed., Vol. 79). SAGE Publications. doi:10.2307/2288384

Kügler, M., & Smolnik, S. (2014). Uncovering the Phenomenon of Employees Enterprise Social Software Use in the Post-Acceptance Stage-Proposing a Use Typology. In *Proceedings of the 2014 European Conference on Information Systems*. Tel Aviv. Retrieved from http://aisel.aisnet.org/ecis2014/proceedings/track21/1/

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174. doi:10.2307/2529310

Leonardi, P. M., Huysman, M., & Steinfield, C. (2013). Enterprise social media: Definition, history, and prospects for the study of social technologies in organizations. *Journal of Computer-Mediated Communication*, *19*, 1–19. doi:10.1111/jcc4.12029

Lietdke, M. (2012). Microsoft Buys Yammer for $1.2 Billion. *Huffington Post*. San Francisco. Retrieved from http://www.huffingtonpost.com/2012/06/25/microsoft-buys-yammer-for_n_1625193.html

Lin, C. Y., Ehrlich, K., Griffiths-Fisher, V., & Desforges, C. (2008). SmallBlue: People mining for expertise search. *IEEE Multimedia*, *15*, 78–84. doi:10.1109/MMUL.2008.17

Lospinoso, J., McCulloh, I., & Carley, K. M. (2009). Utility seeking in complex social systems: An applied longitudinal network study on command and control. In *Proceedings of the 23rd Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour* (pp. 46–51). Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84859044066&partnerID=tZOtx3y1

Luce, R., & Perry, A. (1949). A method of matrix analysis of group structure. *Psychometrika*, *14*(2), 95–116. doi:10.1007/BF02289146

Majchrzak, A., Wagner, C., & Yates, D. (2006). Corporate wiki users: results of a survey. In *Proceedings of the 2006 international symposium on Wikis* (Vol. pp, pp. 99–104). Odense, Denmark. doi:10.1145/1149453.1149472

Mandarano, L. A. (2009). Social Network Analysis of Social Capital in Collaborative Planning. *Society & Natural Resources*, *22*(3), 245–260. doi:10.1080/08941920801922182

Mann, J., Austin, T., Drakos, N., Rozwell, C., & Walls, A. (2012). *Predicts 2013: Social and Collaboration Go Deeper and Wider*.

Mayring, P. (2000). Qualitative Content Analysis. *Forum Qualitative Sozialforschung*, *1*(2), Article 20.

Milgram, S. (1967). The Small-World Problem. *Psychology Today*, *1*, 61–67. doi:10.1007/BF02717530

Mislove, A., Koppula, H. S., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2008). Growth of the flickr social network. In *Proceedings of the 1st workshop on Online social networks - WOSP '08* (pp. 25–30). Princeton, USA. doi:10.1145/1397735.1397742

Moore, E. F. (1959). The shortest path through a maze. In *Proceedings of the International Symposium on the theory of Switching* (pp. 285–292). Harvard University Press.

Nathans, L. L., Oswald, F. L., & Nimon, K. (2012). Interpreting multiple linear regression : A Guidebook of variable importance. *Practical Assessment, Research & Evaluation*, *17*, 1–19. Retrieved from http://www.dspace.rice.edu/handle/1911/71096

Neuendorf, K. (2002). *The Content Analysis Guidebook* (1st ed.). London, UK: Sage Publications. doi:10.1016/j.compmedimag.2009.12.010

Newman, M. E. (2001a). Clustering and preferential attachment in growing networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *64*(2). doi:10.1103/PhysRevE.64.025102

Newman, M. E. (2001b). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, *64*(1), 16132. doi:10.1103/PhysRevE.64.016132

Newman, M. E. (2001c). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 404–409. doi:10.1073/pnas.021544898

Newman, M. E. (2004). Analysis of weighted networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *70*(5). doi:10.1103/PhysRevE.70.056131

Nielsen, J. (2013). *Participation Inequality : Encouraging More Users to Contribute*. *Neilsen Norman Group*. Retrieved October 15, 2014, from http://www.nngroup.com/articles/participation-inequality/

Nonnecke, B., & Preece, J. (2000). Lurker demographics: Counting the silent. In *Proceedings of the SIGCHI Conference on Human factors in computing systems* (pp. 73–80). The Hague, the Netherlands. doi:10.1145/332040.332409

Opsahl, T. (2009). *Structure and Evolution of Weighted Networks*. University of London (Queen Mary College), London, UK. Retrieved from http://toreopsahl.com/tnet/

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, *32*(3), 245–251. doi:10.1016/j.socnet.2010.03.006

Opsahl, T., & Panzarasa, P. (2009). Clustering in weighted networks. *Social Networks*, *31*(2), 155–163. doi:10.1016/j.socnet.2009.02.002

Panzarasa, P., Opsahl, T., & Carley, K. M. (2009). Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *Journal of the American Society for Information Science and Technology*, *60*(5), 911–932. doi:10.1002/asi.21015

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*(2), 289–290. doi:10.1093/bioinformatics/btg412

Petrovčič, A., Vehovar, V., & Žiberna, A. (2012). Posting, quoting, and replying: A comparison of methodological approaches to measure communication ties in web forums. *Quality and Quantity*, *46*(3), 829–854. doi:10.1007/s11135-011-9427-z

Pfeil, U., & Zaphiris, P. (2009). Applying qualitative content analysis to study online support communities. *Universal Access in the Information Society*, *9*(1), 1–16. doi:10.1007/s10209-009-0154-3

R Development Core Team. (2013). *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing*. doi:10.1007/978-3-540-74686-7

Richter, A., & Riemer, K. (2013). The Contextual Nature Of Enterprise Social Networking : A Multi Case Study Comparison. In *Proceedings of the 2013 European Conference on Information Systems* (p. Paper 94). Utrecht, the Netherlands.

Richter, D., Riemer, K., & vom Brocke, J. (2011). Internet Social Networking: Research State of the Art and Implications for Enterprise 2.0. *Business & Information Systems Engineering*, *3*(2), 89–101. doi:10.1007/s12599-011-0151-y

Riemer, K., Diederich, S., Richter, A., & Scifleet, P. (2011). Short Message Discussions: On The Conversational Nature Of Microblogging In A Large Consultancy Organisation. In *Proceedings of the Pacific Asia Conference on Information Systems* (p. Paper 158). Brisbane, Australia. Retrieved from http://aisel.aisnet.org/pacis2011/158

Riemer, K., Diederich, S., Richter, A., & Scifleet, P. (2011). *Tweet Talking-Exploring The Nature Of Microblogging at Capgemini Yammer*. Sydney, Australia. Retrieved from http://ses.library.usyd.edu.au/handle/2123/7226

Riemer, K., Overfeld, P., Scifleet, P., & Richter, A. (2012). *Oh, SNEP! The Dynamics of Social Network Emergence-the case of Capgemini Yammer*. Sydney, Australia. Retrieved from http://prijipati.library.usyd.edu.au/handle/2123/8049

Riemer, K., & Richter, A. (2012). *SOCIAL-Emergent Enterprise Social Networking Use Cases: A Multi Case Study Comparison*. Sydney, Australia. Retrieved from http://seslibraryusydedu.empfi.net/handle/2123/8845

Riemer, K., Richter, A., & Seltsikas, P. (2010). Enterprise Microblogging: Procrastination or productive use? In *Proceedings of the 16th Americas Conference on Information Systems* (p. Paper 506). Lima, Peru. Retrieved from http://aisel.aisnet.org/amcis2010/506/

Riemer, K., Scifleet, P., & Reddig, R. (2012). *Powercrowd: Enterprise social networking in professional service work: A case study of Yammer at Deloitte Australia*. Sydney, Australia. Retrieved from http://prijipati.library.usyd.edu.au/handle/2123/8352

Riemer, K., & Tavakoli, A. (2013). *The role of groups as local context in large Enterprise Social Networks: A Case Study of Yammer at Deloitte Australia*. Sydney, Australia. Retrieved from http://prijipati.library.usyd.edu.au/handle/2123/9279

Risius, M. (2014). Is it Really About Facts? The Positive Side of Meforming for Turning Self-Disclosure into Social Capital in Enterprise Social Media. In *Proceedings of the 2013 European Conference on Information Systems*. Tel Aviv, Israel. Retrieved from http://aisel.aisnet.org/ecis2014/proceedings/track21/11/

Sabater, J., & Sierra, C. (2002). Reputation and social network analysis in multi-agent systems. In *Proceedings of the 1st international joint conference on Autonomous agents and multiagent systems* (pp. 475–482). Bologna, Italy. Retrieved from http://www2.iiia.csic.es/~jsabater/Publications/2002-AAMASa.pdf

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, *31*(4), 581–603.

Scott, J. (2000). *Social Network Analysis: A Handbook* (2nd ed.). London, UK: Sage Publications. doi:10.1370/afm.344

Shami, N. S., Ehrlich, K., Gay, G., & Hancock, J. T. (2009). Making sense of strangers' expertise from signals in digital artifacts. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 69–78). Boston, USA. doi:http://doi.acm.org.proxy.lib.umich.edu/10.1145/1518701.1518713

Shi, X., Zhu, J., Cai, R., & Zhang, L. (2009). User grouping behavior in online forums. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 777–785). Paris, France. doi:10.1145/1557019.1557105

Škerlavaj, M., Dimovski, V., & Desouza, K. (2010). Patterns and structures of intra-organizational learning networks within a knowledge-intensive organization. *Journal of Information Technology*, *25*(2), 189–204. doi:10.1057/jit.2010.3

Small, T. (2011). What the Hashtag? A Content Analysis of Canadian Politics on Twitter. *Information, Communication and Society*, *14*(6), 872–895. doi:10.1080/1369118x.2011.554572

Smith, D. (2012). R Tops Data Mining Poll. *Java Developer' Manual*. Retrieved from http://java.sys-con.com/node/2288420

Smith, M., Hansen, D. L., & Gleave, E. (2009). Analyzing Enterprise Social Media Networks. In *Proceedings of the 12th International Conference on Computational Science and Engineering* (pp. 705–710). Vancouver, Canada: IEEE. doi:10.1109/CSE.2009.468

Snijders, T. A. B. (1996). Stochastic actor-oriented dynamic network analysis. *Journal of Mathematical Sociology*, *21*(1-2), 149–172. doi:10.1080/0022250X.1996.9990178

Snijders, T. A. B. (2005). Models for Longitudinal Network Data. In B. Meyer (Ed.), *Models and methods in social network analysis* (Vol. 11, pp. 215–247). Springer.

Statista. (2015). *Number of monthly active Facebook users worldwide 2008-2014*. Retrieved March 05, 2015, from http://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation*, *7*(17), 137–146. Retrieved from http://pareonline.net/getvn.asp?v=7&n=17

Stieglitz, S., Riemer, K., & Meske, C. (2014). Hierarchy or Activity? The Role of Formal and Informal Influence in Eliciting Responses From Enterprise Social Networks. In *Proceedings of the 2014 European Conference on Information Systems*. Tel Aviv, Israel. Retrieved from http://aisel.aisnet.org/ecis2014/proceedings/track07/12/

Stokman, F., & Doreian, P. (1997). Evolution of Social Networks: Processes and Principles. In P. Doreian & F. Stokman (Eds.), *Evolution of social networks* (pp. 233–250). Amsterdam, the Netherlands: Gordon and Breach Publishers.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. (C. A. Chapelle & S. Hunston, Eds.). New York, USA: Cambridge University Press.

Toivonen, R., Kovanen, L., Kivela, M., Onnela, J. P., Saramaki, J., & Kaski, K. (2009). A comparative study of social network models: network evolution models and nodal attribute models. *Social Networks*, *31*(4), 240–254. doi:10.1016/j.socnet.2009.06.004

Tompkins, C. (1992). Using and interpreting linear regression and correlation analyses. *Clinical Aphasiology*, *21*, 35–46.

Toral, S. L., Martínez-Torres, M. R., & Barrero, F. (2010). Analysis of virtual communities supporting OSS projects using social network analysis. *Information and Software Technology*, *52*(3), 296–303. doi:10.1016/j.infsof.2009.10.007

Van der Aalst, W. M. P., & Song, M. (2004). Mining Social Networks: Uncovering interaction patterns in business processes. *Business Process Management - Lecture Notes in Computer Science*, *3080*, 244–260. doi:10.1007/978-3-540-25970-1_16

Vance, A. (2009). Data Analysts Captivated by R's Power. *New York Times*. New York, USA. Retrieved from http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?_r=0

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge, UK: Cambridge University Press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*, 440–442. doi:10.1038/30918

Weber, R. P. (1990). *Basic Content Analysis*. (S. McElroy, Ed.) (2nd ed.). Sage Publications. doi:10.2307/2289192

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data. *Journal of Statistical Software*, *40*(1), 1–29. doi:10.1.1.182.5667

Wooldridge, M. (2002). *Introduction to Multiagent Systems* (1st ed.). Chichester, England: John Wiley & Sons. Retrieved from http://www.csc.liv.ac.uk/~mjw/pubs/imas/

Yammer. (2014a). *Yammer Customer Site*. Retrieved December 19, 2014, from https://about.yammer.com/customers/

Yammer. (2014b). *Yammer Developer Center*. Retrieved November 21, 2014, from https://developer.yammer.com/v1.0/docs/data-export-api

Yammer. (2014c). *Yammer Home Page*. Retrieved December 19, 2014, from http://www.yammer.com

Yang, S., & Knoke, D. (2001). Optimal connections: Strength and distance in valued graphs. *Social Networks*, *23*(4), 285–295. doi:10.1016/S0378-8733(01)00043-0

Zhang, J., Ackerman, M. S., & Adamic, L. (2007). Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th International Conference on World Wide Web* (pp. 221–230). Banff, Canada. doi:10.1145/1242572.1242603

Zhang, J., Qu, Y., Cody, J., & Wu, Y. (2010). A Case Study of Micro-blogging in the Enterprise: Use, Value, and Related Issues. In *Proceedings of the 28th ACM Conference on Human Factors in Computing Systems* (pp. 123–132). doi:10.1145/1753326.1753346

Zhao, D., Rosson, M. B., Matthews, T., & Moran, T. (2011). Microblogging's impact on collaboration awareness: A field study of microblogging within and between project teams. In *Proceedings of the 2011 International Conference on Collaboration Technologies and Systems* (pp. 31–39). Philadelphia, USA: IEEE. doi:10.1109/CTS.2011.5928662

Appendix A. SNA metrics.

SNA descriptive metrics

Network-level metrics

Network density[1]:

$$D = \frac{|E|}{n(n-1)/2} \text{ (undirected network)}$$

$$D = \frac{2|E|}{n(n-1)/2} \text{ (directed network)}$$

where $|E|$ represents the absolute number of ties present in the network, and n represents the number of nodes.

Clustering coefficient[2]:

$$C_w = \frac{\sum_{\Delta\tau} w}{\sum_{\tau} w}$$

where $\sum_{\Delta\tau} w$ represents to the total weighted value of closed triplets in the networks, and $\sum_{\tau} w$ represents the total weighted value of triplets in the networks.

Node-level metrics

Weighted degree centrality[3]:

$$s_i = \sum_{j=1}^{N} s_{ij} w_{ij}$$

where N represents the total number of ties for node s_i, and w represents the tie weight of each tie.

---

[1]Coleman, T. F., & Moré, J. J. (1983). Estimation of Sparse Jacobian Matrices and Graph Coloring Blems. *SIAM Journal on Numerical Analysis*. doi:10.1137/0720013

[2]Luce, R., & Perry, A. (1949). A method of matrix analysis of group structure. *Psychometrika*, *14*, 95–116. doi:10.1007/BF02289146

[3]Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 3747–3752.

Betweenness centrality[4]:

$$C_B(v) = \sum_{v \neq s \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the number of shortest paths between node *s* and node *t*, $\sigma_{st}(v)$ is the number of times node s appears on the shortest paths between node *s* and node *t*, and *V* is the set of nodes in the network.

Closeness centrality[5]:

$$C_C(v) = \frac{1}{\sum_{t \in V} d_G(v,t)}$$

where $d_G(v,t)$ represents the distance from node v to node t, and V represents the set of nodes in the network.

Eigenvector centrality[6]:

$$C_E(v) = \lambda^{-1} \sum_t A_{(v,t)} t$$

where λ is a constant, and $A_{(v,t)}$ is the adjacency matrix between node v and node t.

---

[4]Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, *40*, 35. doi:10.2307/3033543

[5]Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, *31*(4), 581–603.

[6]Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*. doi:10.1080/0022250X.1972.9989806

Appendix B. Pseudo-code for overlap analysis within and between groups.

Pseudo-code for measuring overlap between central positions within a group

start
all_groups ← list of all groups with their respective users
i ← 1

start loop
1: for group[i] in *all_groups*
  1.1: *current_group* ← list of users in group[i]

  1.2: for every user in *current_group*
    1.2.1: calculate all centrality metrics and store in *current_group*

  1.3: for every centrality metric in *current_group*
    1.3.1: *current_centrality_metric* ← currently selected centrality metric
    1.3.2: order users in *current_group* in a descending order according to *current_centrality_metric*
    1.3.3: select top 5% of users according to *current_centrality_metric* and store in list *top5_metric[group_name]*
    1.3.4: select top 10% of users according to *current_centrality_metric* and store in list *top10_metric[group_name]*

  1.4: store all users in all *top5_metric[group_name]* and *top10_metric[group_name]* in list *top_users[group_name]*

  1.5: for every entry in *top_users[group_name]*
    1.5.1: *current_top_list* ← currently selected list of top users
    1.5.2: *other_top_lists* ← list of all other top lists except for *current_top_list*

  1.6: for every entry in *other_top_lists*
    1.6.1: *current_other* ← currently selected list from *other_top_lists*
    1.6.2: calculate the relative overlap between users in *current_top_list* and *current_other*
    1.6.3: store the result in table *current_group_overlap[i]*

  1.7: *i* ← *i* + 1
  1.8: if *i* larger than number of groups in *all_groups* then terminate loop, else return to step 1
  end loop

2: return all tables in *current_group_overlap[i]*
end

Pseudo-code for measuring the redundancy of central users across groups

```
start
all_groups ← list of all groups with their respective users
i ← 1

start loop 1
1: for group[i] in all_groups
  1.1: current_group ← list of users in group[i]

  1.2: for every user in current_group
    1.2.1: calculate all centrality metrics and store in current_group

  1.3: for every centrality metric in current_group
    1.3.1: current_centrality_metric ← currently selected centrality metric
    1.3.2: order users in current_group in a descending order according to current_centrality_metric
    1.3.3: select top 5% of users according to current_centrality_metric and store in list
           top5_ metric[group_name]
    1.3.4: select top 10% of users according to current_centrality_metric and store in list
           top10_metric[group_name]

  1.4: store all users in all top5_metric[group_name] and top10_metric[group_name] in list
       top_users[group_name]

  1.5: i ← i + 1
  1.6: if i larger than number of groups in all_groups then terminate, else return to step 1
  end loop 1

start loop 2
2: for every [group_name] in top_users[group_name]
  2.1: current_top_users ← list of users in currently selected [group_name]

  2.2: for every centrality metric in current_top_users
    2.2.1: current_centrality_users ← list of top users according to currently selected centrality metric

      2.2.1.1: for every user in current_centrality_users
        2.2.1.1.1: count all other groups where the current user holds a central position and record which
        2.2.1.1.2: store overlap table for current user in list group_overlap[userID]
end loop 2

start loop 3
3: for all tables in group_overlap[userID]
  3.1: divide all cell values in group_overlap[userID] by the number of entries in group_overlap
  3.1: add cell values to table group_overlap_final
end loop 3

4: return group_overlap_final

end
```

Appendix C. Longitudinal network data

Samples		Genres									Network statistics			Network metrics				
Group	TP	Information input %	Social %	Discussion %	Notification %	Other %	Praise %	Problem solving %	Idea generation %	Status updates %	Total message count	Nodes	Edges	Reciprocity	Average weighted degree centrality	Average path length	Graph density	Average clustering coefficient
ABX1	1	18.87	0.00	37.74	16.98	7.55	1.89	9.43	0.00	7.55	126	26	53	0.642	5.346	1.951	0.082	0.205
	2	13.56	0.00	67.80	3.39	5.08	3.39	0.00	6.78	0.00	220	43	113	0.425	4.930	2.340	0.063	0.333
	3	9.76	0.00	65.85	4.88	2.44	4.88	4.88	7.32	0.00	52	23	33	0.485	1.870	2.956	0.065	0.076
	4	6.12	0.00	75.51	2.04	0.00	2.04	6.12	6.12	2.04	47	16	27	0.593	2.438	2.270	0.112	0.073
	5	9.43	0.00	67.92	3.77	0.00	7.55	11.32	0.00	0.00	79	16	30	0.600	4.125	2.103	0.125	0.284
	6	1.67	0.00	81.67	0.00	3.33	5.00	8.33	0.00	0.00	92	19	33	0.642	3.000	2.215	0.096	0.312
ABX2	1	10.17	0.00	40.68	18.64	5.08	3.39	0.00	20.34	1.69	51	19	27	0.370	2.053	2.103	0.079	0.192
	2	5.41	0.00	56.76	10.81	0.00	2.70	0.00	24.32	0.00	61	21	30	0.200	1.762	1.519	0.071	0.246
	3	31.43	0.00	22.86	17.14	5.71	2.86	0.00	5.71	14.29	22	7	6	0.000	1.000	1.333	0.143	0.000
	4	39.29	0.00	10.71	35.71	3.57	3.57	0.00	0.00	7.14	18	3	3	0.667	1.333	1.250	0.500	0.000
	5	22.86	0.00	52.86	8.57	2.86	7.14	0.00	0.00	5.71	129	19	30	0.133	1.474	2.382	0.088	0.212
	6	34.07	0.00	51.65	8.79	1.10	4.40	0.00	0.00	0.00	114	17	29	0.370	3.118	1.990	0.107	0.314
ABX3	1	15.19	1.27	29.11	3.80	2.53	34.18	0.00	2.53	11.39	73	17	23	0.348	2.000	1.979	0.085	0.169
	2	35.00	0.00	5.00	10.00	0.00	5.00	35.00	0.00	10.00	15	10	8	0.250	1.000	1.417	0.089	0.000
	3	9.71	0.00	33.98	23.30	4.85	20.39	1.94	2.91	2.91	152	51	73	0.110	1.647	2.910	0.029	0.053
	4	38.64	20.45	6.82	2.27	0.00	4.55	27.27	0.00	0.00	129	46	66	0.212	1.652	2.739	0.032	0.081
	5	20.00	0.00	40.00	30.00	0.00	10.00	0.00	0.00	0.00	18	10	5	0.000	0.500	1.000	0.056	0.000
	6	32.00	0.00	32.00	0.00	0.00	4.00	32.00	0.00	0.00	29	13	16	0.500	1.462	2.105	0.103	0.025
ABX4	1	10.00	0.00	56.00	2.00	0.00	16.00	8.00	2.00	6.00	67	25	35	0.057	1.680	1.910	0.058	0.212
	2	15.00	0.00	55.00	0.00	0.00	12.50	5.00	2.50	10.00	32	28	40	0.150	1.643	2.587	0.053	0.228
	3	20.00	0.00	26.00	1.00	4.00	11.00	36.00	0.00	2.00	219	48	110	0.291	2.958	3.033	0.049	0.178
	4	13.21	0.00	45.28	1.89	0.00	3.77	30.19	0.00	5.66	236	45	102	0.235	3.356	3.084	0.052	0.239
	5	0.00	0.00	10.34	3.45	0.00	24.14	51.72	0.00	10.34	36	18	18	0.333	1.278	1.250	0.059	0.217
	6	6.52	0.00	45.65	0.00	0.00	13.04	28.26	0.00	6.52	58	22	29	0.276	1.500	2.186	0.063	0.072
ABX5	1	1.41	0.00	67.61	1.41	0.00	2.82	2.82	23.94	0.00	91	14	27	0.519	4.214	1.849	0.148	0.416
	2	9.09	0.00	68.18	0.00	0.00	4.55	0.00	18.18	0.00	35	7	15	0.267	3.571	1.677	0.357	0.527
	3	9.09	0.00	22.73	9.09	4.55	0.00	47.73	2.27	4.55	367	30	88	0.591	8.067	2.143	0.101	0.396
	4	11.11	0.00	40.17	4.27	3.42	5.98	19.66	2.56	12.82	555	48	162	0.617	9.458	2.301	0.072	0.313
	5	21.67	0.00	36.67	1.67	1.67	3.33	18.33	6.67	10.00	100	21	41	0.439	3.238	2.522	0.098	0.269
	6	12.90	0.00	54.84	0.00	3.23	3.23	22.58	0.00	3.23	105	27	47	0.468	2.889	2.458	0.067	0.141

.

Appendix D. Group overlap tables.

ABX1		Weighted in-degree centrality		Weighted out-degree centrality		Betweenness centrality		Closeness centrality		Eigenvector centrality		
		Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Average
Weighted in-degree centrality	Top 5			100.00	100.00	87.50	100.00	25.00	37.50	87.50	100.00	79.69
	Top 10			47.06	82.35	47.06	76.47	11.76	29.41	47.06	82.35	52.94
Weighted out-degree centrality	Top 5	100.00	100.00			87.50	100.00	25.00	37.50	87.50	100.00	79.69
	Top 10	47.06	82.35			47.06	76.47	11.76	29.41	47.06	76.47	52.21
Betweenness centrality	Top 5	87.50	100.00	87.50	100.00			25.00	50.00	87.50	100.00	79.69
	Top 10	47.06	76.47	47.06	76.47			11.76	29.41	47.06	70.59	50.74
Closeness centrality	Top 5	25.00	25.00	25.00	25.00	25.00	25.00			25.00	37.50	26.56
	Top 10	17.65	29.41	17.65	29.41	23.53	29.41			17.65	35.29	25.00
Eigenvector centrality	Top 5	87.50	100.00	87.50	100.00	87.50	100.00	25.00	37.50			78.13
	Top 10	47.06	82.35	47.06	76.47	47.06	70.59	17.65	35.29			52.94
Average		57.35	74.45	57.35	73.71	56.53	72.24	19.12	35.75	55.79	75.28	**57.76**

ABX2		Weighted in-degree centrality		Weighted out-degree centrality		Betweenness centrality		Closeness centrality		Eigenvector centrality		
		Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Average
Weighted in-degree centrality	Top 5			66.67	66.67	66.67	83.33	50.00	66.67	50.00	100.00	68.75
	Top 10			50.00	58.33	50.00	75.00	41.67	58.33	50.00	83.33	58.33
Weighted out-degree centrality	Top 5	66.67	100.00			100.00	100.00	83.33	83.33	83.33	100.00	89.58
	Top 10	33.33	58.33			50.00	66.67	50.00	58.33	50.00	66.67	54.17
Betweenness centrality	Top 5	66.67	100.00	100.00	100.00			83.33	83.33	83.33	100.00	89.58
	Top 10	41.67	75.00	50.00	66.67			41.67	58.33	50.00	66.67	56.25
Closeness centrality	Top 5	50.00	83.33	83.33	100.00	83.33	83.33			83.33	100.00	83.33
	Top 10	33.33	58.33	41.67	58.33	41.67	58.33			50.00	75.00	52.08
Eigenvector centrality	Top 5	50.00	100.00	83.33	100.00	83.33	100.00	83.33	100.00			87.50
	Top 10	50.00	83.33	50.00	66.67	50.00	66.67	50.00	75.00			61.46
Average		48.96	82.29	65.63	77.08	65.63	79.17	60.42	72.92	62.50	86.46	**70.10**

ABX3		Weighted in-degree centrality		Weighted out-degree centrality		Betweenness centrality		Closeness centrality		Eigenvector centrality		
		Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Average
Weighted in-degree centrality	Top 5			41.67	50.00	58.33	83.33	66.67	83.33	66.67	83.33	66.67
	Top 10			29.17	37.50	33.33	58.33	37.50	50.00	41.67	58.33	43.23
Weighted out-degree centrality	Top 5	41.67	58.33			58.33	91.67	66.67	83.33	33.33	58.33	61.46
	Top 10	25.00	37.50			37.50	62.50	37.50	62.50	25.00	37.50	40.63
Betweenness centrality	Top 5	58.33	66.67	58.33	75.00			58.33	58.33	41.67	58.33	59.38
	Top 10	41.67	58.33	45.83	62.50			41.67	62.50	29.17	54.17	49.48
Closeness centrality	Top 5	66.67	75.00	66.67	75.00	58.33	83.33			58.33	83.33	70.83
	Top 10	41.67	50.00	41.67	62.50	29.17	62.50			37.50	66.67	48.96
Eigenvector centrality	Top 5	66.67	83.33	33.33	50.00	41.67	58.33	58.33	75.00			58.33
	Top 10	41.67	58.33	29.17	37.50	29.17	54.17	41.67	66.67			44.79
Average		47.92	60.94	43.23	56.25	43.23	69.27	51.04	67.71	41.67	62.50	**54.38**

ABX4		Weighted in-degree centrality		Weighted out-degree centrality		Betweenness centrality		Closeness centrality		Eigenvector centrality		
		Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Average
Weighted in-degree centrality	Top 5			33.33	77.78	55.56	88.89	66.67	100.00	77.78	100.00	75.00
	Top 10			31.58	63.16	36.84	73.68	36.84	73.68	47.37	94.74	57.24
Weighted out-degree centrality	Top 5	33.33	66.67			44.44	100.00	33.33	77.78	22.22	66.67	55.56
	Top 10	36.84	63.16			42.11	78.95	26.32	78.95	26.32	57.89	51.32
Betweenness centrality	Top 5	55.56	77.78	44.44	88.89			55.56	88.89	55.56	77.78	68.06
	Top 10	42.11	73.68	47.37	78.95			42.11	78.95	36.84	73.68	59.21
Closeness centrality	Top 5	66.67	77.78	33.33	55.56	55.56	88.89			66.67	88.89	66.67
	Top 10	47.37	73.68	36.84	78.95	42.11	78.95			42.11	78.95	59.87
Eigenvector centrality	Top 5	77.78	100.00	22.22	55.56	55.56	77.78	66.67	88.89			68.06
	Top 10	47.37	94.74	31.58	57.89	36.84	73.68	42.11	78.95			57.89
Average		50.88	78.44	35.09	69.59	46.13	82.60	46.20	83.26	46.86	79.82	**61.89**

ABX5		Weighted in-degree centrality		Weighted out-degree centrality		Betweenness centrality		Closeness centrality		Eigenvector centrality		
		Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Top 5	Top 10	Average
Weighted in-degree centrality	Top 5			100.00	100.00	71.43	85.71	85.71	85.71	85.71	85.71	87.50
	Top 10			46.67	86.67	40.00	66.67	46.67	86.67	46.67	86.67	63.33
Weighted out-degree centrality	Top 5	100.00	100.00			71.43	85.71	85.71	85.71	85.71	85.71	87.50
	Top 10	46.67	86.67			40.00	66.67	46.67	86.67	46.67	73.33	61.67
Betweenness centrality	Top 5	71.43	85.71	71.43	85.71			71.43	85.71	71.43	85.71	78.57
	Top 10	40.00	66.67	40.00	66.67			46.67	66.67	46.67	66.67	55.00
Closeness centrality	Top 5	85.71	100.00	85.71	100.00	71.43	100.00			100.00	100.00	92.86
	Top 10	40.00	86.67	40.00	86.67	40.00	66.67			46.67	86.67	61.67
Eigenvector centrality	Top 5	85.71	100.00	85.71	100.00	71.43	100.00	100.00	100.00			92.86
	Top 10	40.00	86.67	40.00	73.33	40.00	66.67	46.67	86.67			60.00
Average		63.69	89.05	63.69	87.38	55.71	79.76	66.19	85.48	66.19	83.81	**74.10**