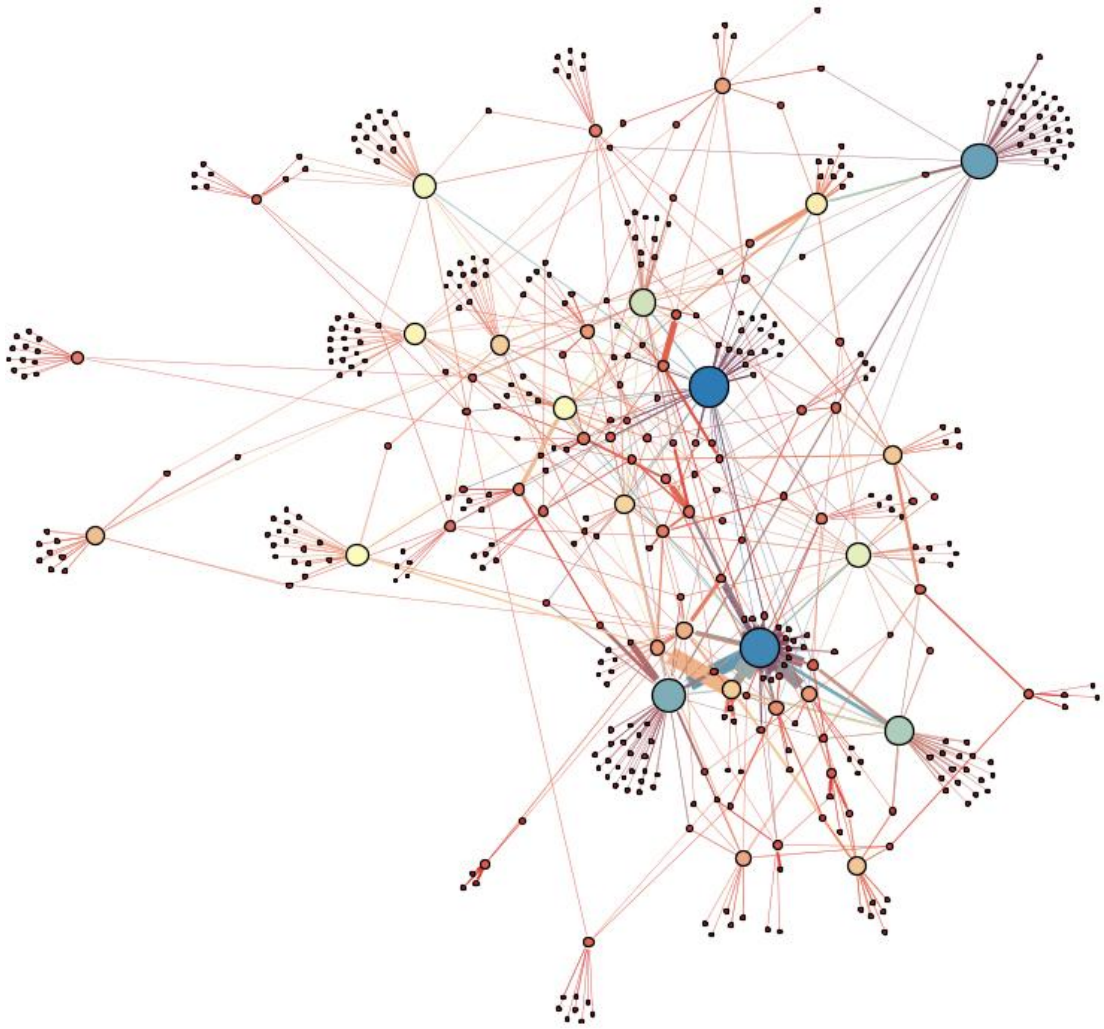
Structural Creation of a Social Network:

An Analysis of Longitudinal Development of Online Communities



Author:

Wensi Ai [4035348]
Master of Business Informatics
Institute of Information &
Computer Science
Utrecht University

First Supervisor:

Prof. dr. Remko Helms
Institute of Information &
Computer Science
Utrecht University

Second Supervisor:

Dr. Fabiano Dalpiaz
Institute of Information &
Computer Science
Utrecht University

Abstract

Over the past ten years, online communities have attracted significant attention from a wide range of organizations in various markets. Social interactions within these virtual communities have logged an enormous amount of digital traces, and opened up new opportunities to researchers in the field of social network studies. However, current research in this field seems mainly focused on the individual participant, while very little has been researched from a structural, time-based [longitudinal] network perspective. One of the crucial issues that comes with it, is that there is insufficient knowledge regarding the aspect of evolution in online communities. Such issues get aggravated when the data extracted from online communities are unstructured (i.e. without explicit relational references) and are cumulated over a long period of time.

With this in mind, in this study we carried out a series of experiments and analysis on a public online forum [online community] with a hierarchical network structure. The objective is to identify scientific methods from existing literature for creating network structures based on unstructured data, as well as scientific approaches for longitudinal SNA. Finally, we try to determine the validity and applicability of existing literature on social network analysis in online communities from a practical point of view, in order to contribute to a scientific approach in investigating longitudinal development of online communities, from a network's perspective.

As a result, we identified certain limitations in the existing literature and proposed our own methods and guidelines for social network creation and longitudinal SNA. The results of our work revealed the critical issues related to working with unstructured and longitudinal communication data retrieved from an online forum.. The analytical results of edge-ratio analysis suggest that the development of an online community can be monitored periodically based on fractions of the whole network structure (i.e. network snapshots). Indications are that the development of an online community is two-folded, i.e. participation of "old" users as well as fresh in-streams of new users is roughly equally important. Last but not least, an attempt to visualize an online community's development in a longitudinal manner with the Gephi software package has provided discernible insights, as compared to other methods.

Keywords: Online community, longitudinal development, social network analysis, unstructured data.

Acknowledgement

This thesis is the finishing line of a great journey for me, here at Utrecht university. It has been a wonderful two and a half years of top-level academic experience. First of all, I would like to express my gratitude towards my first supervisor Prof. Dr. Remko Helms and my second supervisor Dr. Fabiano Dalpiaz for their unreserved and professional guidance throughout this fascinating project. Secondly, I would like to thank Gábor Majdán and Steffen Bjerkenås for their inspirations on the subject of my thesis. Finally, I'd like to thank Danny Vandebeek for his thesis (language) review and support with Python programming, and my classmate / roommate Yudi Xu, for the morale support. I would not have made it this far without the help of you gentlemen. Thank you all!

Table of Contents

List of Tables

List of Figures

1 Introduction and problem statement

The interactions among social network participants have left an enormous amount of complex, digitalized and self-documenting records behind (Gleave & Welser, 2009), and have naturally formed virtual [online] communities, as they perform activities regarding connectivity, content creation, conversation and collaboration (Ang, 2011). Such online communities have drawn significant attention from a wide range of organizations in various markets. Many of them have quickly realized the potential benefits that could be gained by participating in online communities (Faasse, Helms, & Spruit, 2011), in order to gain business advantages (Culnan, McHugh, & Zubillaga, 2010).

It is reported that by engaging participants in such online communities, an organization could increase customer focus and understanding, improve customer service levels and decrease time-to-market (Jussila, Kärkkäinen & Leino, 2011). For instance, information gleaned from online communities could provide indications for new business opportunities as well as new ideas for products to support marketing campaigns (Gillin & Schwartzman, 2011) and help in establishing promotion mix oriented activities for advertisement (Mangold & Faulds, 2009). It can also strengthen relationships with customers and improve both internal and external collaboration (Hoffman & Fodor, 2010).

Major streams of research targeting social network and online communities tend to focus on characterizing and positioning different categories or classifications of their participants (Kaplan & Haenlein, 2010; Stanoevska-Slabeva & Schmid, 2001) or trying to identify their user groups (Correa, Hinsley & de Zúñiga, 2010; Smith, 2011). Another interesting research perspective tries to explain the motives for participation from a sociological point of view (Ridings, & Gefen, 2004; Wasko & Faraj, 2005). Other studies have been conducted to investigate the interactions and affects of social network participants among each other in online communities, in order to analyze and identify the driving factors of its popularity (Agichtein, 2008; Fischer & Reuber, 2011).

However, the current research streams seem to be focused on the individual level; very little is known or studied from the structural perspective (Faraj & Johnson, 2011). Interesting questions such as "*how are online communities structured?*" and "*what about the structural evolution [patterns] of those online communities over time?*" remain unanswered.

One of the critical issues that comes with it, is that there is a lack of knowledge regarding the aspect of longitudinal development in online communities, in terms of their structural

evolution over time. It is interesting to be acquainted with existing research perspectives that could provide solutions on solving this issue with this regard. Furthermore, in order to analyze the longitudinal development of the structural evolution of an online community, one perspective is to investigate this matter in the sense of an aggregated set of network fragments accumulated over time.

This process in itself requires data segmentation, yet how to appropriately segment the data is still ambiguous: do we measure these data [segments] in days, weeks, months, years even? How do we notice that the social network's growth is accelerating, and when does a social network implode, from a network perspective? The existing literature suggests that the time interval for data segmentation may differ drastically (Igarashi, 2005; Hansen et al., 2011; Petrovčič, Vehovar & Žiberna, 2012).

Moreover, the network structures are created upon the interaction [edge] patterns of people [nodes] that participate in a certain online community (Backstrom, Dwork & Kleinberg, 2007; Hansen, Shneiderman & Smith, 2011). There is not yet a standard manner or method to present and describe those particular network structures; it is even still unclear how these online community structures can be identified, measured and visualized.

Additionally, difficulties in describing this network structure are aggravated when the data generated by an online community is basically unstructured, meaning that there is a lack of relational references between participants (Sack, 2000), or the interaction patterns of participants might not be indicated explicitly (Petrovčič et al., 2012).

With the aforementioned argumentations and issues, we summarize the formal problem statement of this research as follows:

"The existing literature about using social network analysis techniques on research of online communities has contributed a great deal to our understanding of this field. However, very little has been investigated from the network perspective. There is no standard method to describe the network structure and its related data segmentation, nor for analysis of the network [structure]'s longitudinal development, and such issues get aggravated when the data generated by an online community is unstructured."

The objective of this research is to identify the important aspects and methods for structural creation of online communities, for the purpose of analyzing the longitudinal development of online communities with regard to their structural evolvement over time, from a scholarly perspective, in the context of a business-oriented social networking environment.

1.1 Research questions

Based on the problem statements from the previous section, the main research question of this study is formulated as follows:

"How does the social network structure of online communities evolve over time?"

To properly answer this research question, we plan to study the follow sub questions:

SQ1: What are the existing scientific methods for structure creation of online communities and social network analysis?
SQ2: What are the existing methods for network creation based on unstructured data?
SQ3: What are the existing research perspectives regarding longitudinal analysis of online communities?
SQ4: What are the important metrics for measuring the development of online communities?

Research (sub)question 1 will be completely based on phase 1 of this thesis, the literature review. The other [answers to] sub-questions will be based on an integration of findings from the literature review with the results of experimentation and social network analysis we gather from the online community's data set at hand.

1.2 Research approach

In this chapter, we elaborate the research approach of this thesis project. As stated in the problem statement, this study tends to focus on the longitudinal development of online communities and its related issues as the main objective.

Within the given research context, we define our research approach as an exploratory study, on the data set from an online community (an online forum). First, a literature study is conducted to review the existing scientific publications as well as ongoing research, to investigate the available sources regarding longitudinal social network analysis in online communities, from a theoretical perspective. Secondly, the findings of literature review are then validated in terms of their applicability and feasibility in practice, via a series of experimentations. The objective here, is to determine whether the findings from literature are indeed useable or adaptable to the context of our research. As a result, the data acquired from the online community is transformed into social network structures, and are then processed for longitudinal SNA. Finally, SNA is carried out on basis of the network structures created

from previous steps. Based on these results, we draw conclusions for this research. In the last section of this chapter, issues concerning validity of this study are addressed.

1.2.1 Literature review

The literature review is to be conducted for investigating the theoretical background on the subject of online communities, especially towards the aspects of social network structural creation & analysis techniques and its related subjects, such as network creation based on structured or unstructured data, and different research perspectives regarding longitudinal analysis of online communities. The goal is to answer or partially answer the sub-research questions, from a theoretical perspective.

In the context of this thesis project, for literature review, we plan to use the snowballing technique, instead of systematic literature review. The main reason for this choice is the fact that the first supervisor of this study has provided essential publications and results of ongoing research regarding the chosen subject.

Additionally, Jalali and Wohlin (2012) have compared the snowballing technique with systematic literature review technique. Results of this comparison indicate similar outcomes between the two literature review techniques. Webster and Watson (2002) defined the snowballing technique with the following steps:

1) Find at least one leading source which is relevant to the research objective(s).
2) Go backward by reviewing the citations for the articles identified in step 1 to determine prior articles that should be considered.
3) Go forward by using electronic libraries to identify articles citing the key articles identified in the previous steps. Determine which of these articles should be included in the review.

Webster et al. (2002) also emphasized that this three-step review technique should be carried out iteratively and recursively, until there are no more new concepts to be found. In our case, we have rather focused objectives; therefore, instead of finding all relevant concepts, we narrowed down our search towards publications and other sources that could directly contribute towards solving the main research question / sub-questions.

As mentioned earlier, a number of publications as well as ongoing research have been provided by the first supervisor, those studies are utilized as the starting points of snowballing literature review:

1) Network exchange patterns in online communities (Faraj, S & Johnson, S, 2011)
2) Towards dynamic visualization for understanding evolution of digital communication networks (Trier, M, 2008)
3) The success and sustainability of online communities: a social analysis approach (Gabor, M, 2012)
4) The structural characteristics and conversational nature of enterprise social network (Bjerkenas, S, 2015)

Those studies have been reviewed intensively as the starting points of snowballing literature review. Backwards review (step 2) and forwards review (step 3) are then carried out iteratively. A literature found in these steps is considered relevant to our research based on whether its title, abstract or contents contain the keywords illustrated in Table 1.1, regarding one or more sub-questions elaborated previously. It is then reviewed by its abstract and conclusion to determine whether it is to be investigated in a more holistic manner. In such a case, it is then reviewed in depth in order to extract findings that could provide insights for this research from a theoretical perspective.

Keyword	SQ1	SQ2	SQ3	SQ4
Social network	x	x	x	x
Online community	x	x	x	x
Social network analysis	x	x	x	x
Longitudinal	x		x	
Time span			x	
Time window			x	
Time interval			x	
Segmentation			x	
Fragment			x	
Network metric				x
Network creation	x	x		
Unstructured data		x		
Relational reference		x		

Table 1.1: Keyword/Sub-question metrics

1.2.2 Experimentation

In this phase, the objective is to conduct experiments, in order to verify the applicability of findings from literature review on the obtained data set of an online community. This phase is designated to answer sub-research questions 1 to 3 from practical perspective, and it is carried out in four main steps illustrated in the following sections. A general approach illustrated by Leek, Collado-Torres and Reich (2013) when dealing with quantitative data is applied regarding data pre-processing. As for Network creation and Adaptation of longitudinal analysis in online communities, these two steps are utilized due to the exploratory nature of this research.

Research context

Experimentation begins with introduction to the research context. This includes detailed illustrations of the background and fundamental characteristics of the online community where the data sets of this research are obtained.

Data pre-processing

Social network analysis in general involves statistical analysis of quantitative and/or qualitative data, which usually includes data transfer between different parties as well as pre-processing of data sets. This way, researchers and statisticians are able to conduct analytical activities on a cleaner version of obtained data sets. We plan to apply the data processing guideline illustrated by Leek et al. (2013), to pre-process our data set for the following stages. Scripts used for data pre-processing are programmed in Python, for data merging and other data manipulation processes, Access database as well as Excel have been used.

As a result, a tidy data set is created as the first step towards the network analysis. This initial data set is vital for our research, as many cases involving statistical analysis have demonstrated that 80% of data analysis is spent on the process of preparing and cleaning the data (King et al., 2010).

Network creation

In order to properly apply social network analysis techniques, network structures have to be constructed beforehand (Hanneman & Riddle, 2005). Based on the tidy data set from the previous stage, a clean version of the data set is ready to be used as basis of the network creation.

In this stage of the research, we investigate the applicability of findings from literature review regarding network creation on the basis of the created tidy data set, through a series of experiments. The network structures created from methods suggested in literature are to be

validated against manually created networks through content analysis. Furthermore, alternatives can be proposed based on the results of network validation. Additionally, important network structure properties will be taken into consideration and evaluation.

The implementation of the edge-lists is done by using various pseudo network creation methods programmed in Python. The network validation is then performed between these network structures (edge-lists) by applying the quadratic assignment procedure function in UCINET software tool, to reveal the correlation between pseudo networks and the baseline network. In addition to that, the basic network metrics (e.g. density, average path length etc.) of these pseudo networks and baseline network are compared. Ultimately, the goal is to identify the best pseudo network creation method available, in terms of its accuracy in contrast to the baseline network (reality) , based on findings of existing literature as well as from a practical point of view

For the network creation, each step of the process is to be documented in detail. As a result, an edge-list is created from the optimal pseudo network creation method, containing important network properties and is ready to be analyzed in the following stages. These processes are dedicated to answering sub-question 1 and 2 from a practical perspective.

Adaptation of longitudinal analysis in online communities

One of the limitations of social network analysis is that most of the results are derived from static network structures, that have been created from accumulative data of online communities (Trier, 2008). Such results could be misleading, since the results may not present the dynamic nature of social relationships (Emirbayer, 1997).

In our research, we focus on the area of longitudinal development of online community's structures. We investigate the applicability of various methods and techniques from the longitudinal perspective, based on findings of our literature review, in order to select the most appropriate methods as well as their parameters to formulate a suitable manner for solving the issue mentioned above. As a result, the edge-list can be prepared for social network analysis from a longitudinal perspective. This phase is designated to answer sub-question 3 from an empirical point of view.

1.2.3 Social network analysis

Otte and Rousseau (2002) stated that "*Social network analysis (SNA) is a strategy for investigating social structures through the use of network and graph theories*". It enables the researchers in this field to investigate the structural characteristics of online communities, that are created based on the interaction patterns accumulated over a period of time of its participants (Trier, 2008).

Within the context of our research, we rely on social network analysis techniques to investigate the longitudinal development of an online community. As elaborated in the section of literature review, we utilize appropriate social network analysis techniques solely based on the results of literature review and our understanding of the theoretical background of this subject. Such methods or techniques are to be adopted carefully in order to analyze the quantitative data set we have at hand.

To investigate the longitudinal development of an online community, network analysis has to be carried out. This can be achieved by illustrating the descriptive statistics and analyzing structural characteristics plus its related measureable metrics of the edge-list created previously. The goal of this stage is to answer sub-question 4, and to present results of the analytical work. Based on these results, we draw our conclusions for this research.

For measuring the network's evolution, a set of metrics for Edge-Ratio analysis (Helms & Majdan, 2015), which is specifically designed for this purpose, are utilized and calculated. Other basic SNA measurements (network metrics and centralities, see section 2.7 and 2.8) are calculated by a self-made program developed on Python software and its SNA library NetworkX. For visualization of the network's evolution, Gephi software tool and an additional plug-in Circular Layout are applied.

1.3 Threat to validity

Howison, Wiggins and Crowston (2010) have identified a number of validity issues specifically towards applying social network analysis techniques for research of online communities. In order to maintain the quality of this research, we plan to cautiously identify these threats and to deploy countermeasures against them in order to retain the integrity of this study.

Construct validity

As elaborated by Straub, Boudreau and Gefen (2004), construct validity means "the extent to which a given test/instrumentation is an effective measure of a theoretical construct". In this research, as we study the longitudinal development of an online community, this implies that the data set we have is accumulated over time. Additional verifications are to be performed in the data pre-processing stage as well as later stages to check the consistency of the data stability with respect to the construct of interest (Howison et al., 2010).

External validity

To defend the external validity in the network creation phase, the subjects for online community structural creation are carefully grouped based on their categories of interest, in such a way that the results of network creation can be generalized to a certain extent towards other online communities with similar features and characteristics.

Reliability

The reliability entails whether the operations of the research can be repeated with the same results, so that biases and errors in a study can be minimized. To secure the reliability of research, we plan to establish proper documentation along each step of the way, so that a research operation executed with the same settings would produce the same results and findings. Additionally, a code book will be created according to the guideline illustrated by Leek et al.(2013).

Ethical issue

Last but not least, we take issues regarding research ethics into serious consideration, since the data set may contain sensitive or confidential information of those online community members. Any data that may reveal anonymity of the community's members or may be damaging the confidentiality terms will be excluded from this research.

1.4 Scientific & social relevance

The conclusions of this study might help different kinds of stakeholders in online communities. Owners and operators of these networks will be better able to understand which are the key elements in their networks, and how to monitor their evolution and with it the dynamism, health of these key elements and their surrounding community.

Users of these networks will receive a better service, on precondition that the owners of the networks they use intervene timely and appropriately: growing, – maybe even preening it where necessary – and nourishing the social network for growth and richness of content dependent on the status of monitoring parameters. Broader society, finally, might benefit too, through a more transparent cooperation between government and its citizens, and a more efficient and service-oriented economy.

Scientifically, this research attempts to provide a new way of analyzing longitudinal development in an online community from a network's perspective, by using various social network analysis techniques. Furthermore, experiments have been carried out in this research to offer possible solutions on solving practical issues caused by an imperfect context (i.e. unstructured data) or the longitudinal nature of an online community (i.e. cumulative data), which in turn, can be seen as a helpful guideline to other scholars who are also interested in this field of study.

1.5 Research planning

In the following sections, the research planning and the main deliverables of this thesis are illustrated (Figure 1.1), by means of a Process-Deliverable Diagram (PDD) in order to visualize the research planning. On the left-hand side of the Process-Deliverable Diagram, all processes, data flows and control flows are displayed. Notations used here are based on the activity diagram in UML. The right-hand side are the deliverables, visualized by using the class diagram of UML. Activities are connected with dotted arrows to the produced deliverables (Weerd & Brinkkemper, 2008).

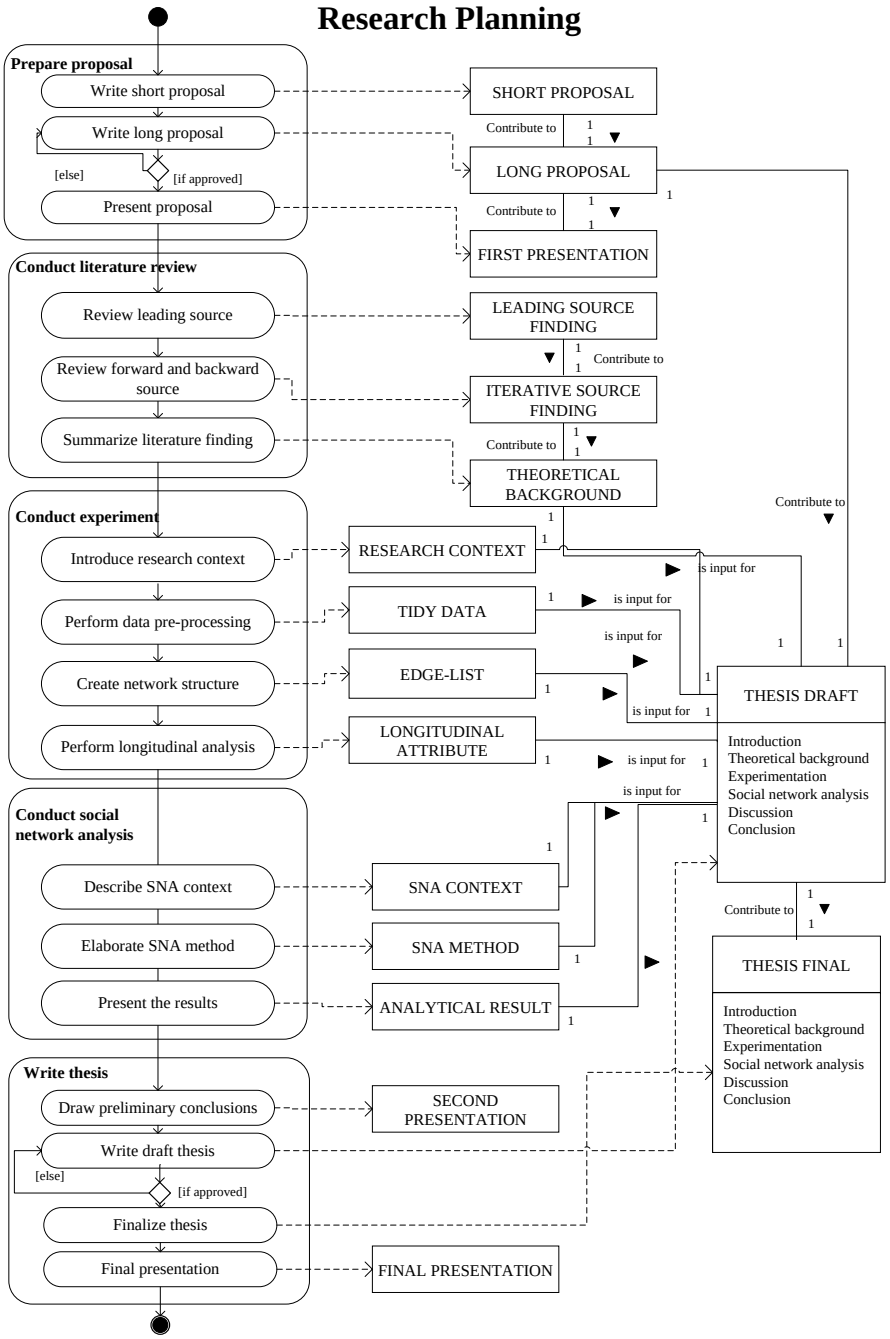


Figure 1.1: Process-Deliverable Diagram of the research planning

1.6 Document structure

In chapter 2, the theoretical background of this research is elaborated. Chapter 3, illustrated the experimentations in this research, regarding the applicability and accuracy of the findings of existing literature on network creation methods for unstructured data and data segmentation. Chapter 4 discusses the methods and results of longitudinal SNA. And finally in chapter 5 and 6, answers to the main and sub research questions are answered, and thereafter, the conclusions are drawn for this research project.

2 Theoretical background

In this chapter, we discuss the findings and results from literature review of aforementioned subjects in previous chapter, in order to investigate the theoretical background of this study. We first elaborate the concepts regarding online community, as a specific online [business] community is the study subject of this thesis. Then, applications on social network analysis in online communities are discussed; thirdly, the essential characteristics of social network structures are elaborated. Furthermore, the network creation and the longitudinal nature of the online community are depicted. Last but not least, the metrics and different types of centralities regarding social network analysis are illustrated.

2.1 Online community

Since the beginning of the age of the Internet, communities are no longer bounded by the limitations of physical world (Lee & Lee, 2008), with the term online/virtual community lacking consensus with regard to a generally accepted definition among scholars in this field of study (Porter, 2004), even though the term itself is seemingly self-explanatory and is relatively easy to be interpreted.

In 1997, Jones described this phenomenon of a "*virtual community as a new form of community*", emphasizing its hosting platform as computer mediated communication (CMC) with the following characteristics:

1) a minimum level of interactivity;
2) a variety of communicators;
3) a minimum level of sustained membership;
4) a virtual common public space where a significant portion of interactivity group CMC occur.

Preece & Maloney-Krichmar (2003) summarized a number of important characteristics that have been utilized by other scholars previously (Rheingold, 1993; Jones, 1997; Wellman 2000) in order to provide a clear overview of online communities, i.e. *"people with shared interest, experiences and/or needs, engaged in supportive and sociable relations, where they obtain important resources, develop strong interpersonal feelings of belonging and being wanted, and forge a sense of shared identity."*

The fundamental concept of an online community is to bring people with various backgrounds together regardless of their physical locations, to share information or to interact with other members over certain agreed topics of shared interests within the community (Hunter, 2002; Lee et. al., 2008), which can be further illustrated as a group of people who regularly interact online and share their common goals, ideas and values (Owston, 1998).

Since the rise of SNSs, the foundation of online communities has been rebalanced. While the traditional online communities (such as bulletin-board systems (BBS's) or Internet Relay Chat (IRC)) that are centered around and structured by category of *interest* continued to exist and prosper, SNSs have also gained tremendous recognition.

This new type of online communities is mainly organized around *people*, not interests (Boyd & Ellison, 2007). However, it is worth mentioning that this rebalancing process is being approached from both directions; the traditional online communities have adopted concepts regarding the egocentric element (such as "friendship") of popular SNSs, while SNSs have embraced topic centric and interaction oriented aspects from the traditional online communities. Most of the definitions for online community are an embodiment of the following aspects: who (people), how (activities or actions), why (motives) and where (platforms), although the emphasis may vary slightly.

We should accept the fact that the concept of online communities has fuzzy boundaries (Bruckman, 2005; Preece & Maloney-Krichmar, 2005). In the context of this research, we plan to utilize the definition elaborated by Ridings, Gefen, and Arinze (2002): "*groups of people with common interests and practices that communicate regularly and for some duration in an organized way over the Internet through a common location or mechanism.*" As indicated in the Ridings et al. paper, terms such as "*common location*" or "*mechanism*" refer to online or virtual "places" which facilitate communication.

There have been many attempts in creating taxonomies for modeling or classifying online communities. In the late nineties, Lazard and Preece (1998) created a classification schema for online communities based on four major categories: 1) by attributes, 2) by supporting software, 3) by relationships to physical communities and 4) by boundedness.

Stanoevska-Slabeva (2002) defined a more comprehensive classification mechanism that emphasized on the *types* of online communities: discussion or conversation communities; task and goal-oriented communities, virtual worlds and hybrid communities.

However, with the rise of SNSs and its friendship-oriented networks, defining the boundary for classifying online communities is just as problematic and fuzzy as providing a widely accepted definition for online communities. It might be a better idea to look deeper into the characteristics of fundamental properties of the online communities.

2.2 Social network analysis in online communities

The phrase "social network" is defined by Wasserman and Faust (1994), as a set of nodes that are bounded by the relations between them. It is a set of interaction patterns [relations] of involved parties that form a structural representation of the network with the following features:

- Actors and their actions are viewed as interdependent rather than independent, autonomous units,
- Relational ties between actors are channeled for transfer or "flow" of resources,
- Network models focusing on individuals view the network structural environment as providing opportunities for, or constraints on individual action,
- Network models conceptualize structure as lasting patterns of relations among actors.

According to Wasserman et al. (1994), Social Network Analysis (SNA) is not a formal theory, but rather a research perspective to approach structural issues regarding social and behavioral science, by formalizing social properties and processes, and providing consistent definitions that allow for testable models of social concept such as "group" and "social role" (Howison et al., 2010).

Otte et al. (2002) also argued that SNA is a powerful strategy for investigating social structures in the field of information science, with extension towards information systems; in this sense, SNA would be better described as *"a set of mathematical techniques for analyzing networks"* (Howison et al., 2010).

Traditionally, SNA relied on questionnaires, interviews, observations and studies of archival records as its primary data collection mechanism, for creating the social networks (Wasserman et al.,1994). These particular ways of data collection and conducting SNA research have accomplished much in the past, including the introduction of the famous *"small world effect"* term/hypothesis (Travers & Milgram, 1969), a concept often referred to as the "*six degrees of separation*" principle.

With the development over the past two decades regarding the field of information science and information systems, the number of studies on SNA research have also skyrocketed (Knoke & Yang, 2008). SNA techniques in their core fundamentally rely on quantitative data, in order to create networks/graphs for analysis; a requirement which could easily be fulfilled by the large quantities of user-created contents and system generated digital traces of the online communities. In this sense, SNA and studies of online communities seem to match naturally, by using the enormous amount of available data sources provided by the online communities (Agarwal, Gupta & Kraut, 2008; Howison et al., 2010).

Scholars have quickly recognized the advantages of the match between SNA and online communities, and have begun utilizing those advantages in various domains. Examples can be given such as identification of the key participants of online communities in terms of added value in knowledge sharing (Berger, Klier, Klier & Richter, 2014) or to distinguish bottlenecks in the knowledge sharing processes (Helms & Buijsrogge, 2006). Social network analysis can also help in defining the social roles of people participating in online communities (Gleave & Welser, 2009). Furthermore, it enables researchers to investigate the social capital of online communities (e.g. Lee et al., 2008; Huffaker, 2010).

Gloor, Laubacher, Zhao and Dynes (2004) presented a research that focused on the analysis and visualization of a consulting practice, based on its digital communication exchange patterns (i.e. e-mail exchange) within this organization over one year of time with 200 participants in total. The objective of this study is to investigate the dynamics and evolution between individuals of this particular social network, by means of measuring and visualizing the generic and widely accepted network metrics and centralities in SNA (see section 2.7 and 2.8) using the software tool UCINET (Borgatti, Everett & Freeman, 1992) in a periodic manner (i.e. time windows). The results of this study can be presented in two different modes based on 30 days time span: 1) history mode and 2) no history mode. In case of history mode, the calculations of the network metrics and centralities are visualized in a cumulative manner in graphs, e.g. every new graph generated will include of the current time window as well as the decayed time windows previously, in this sense, this mode can also be interpreted as a cumulative mode. In no history mode, the those measurements are visualized only on basis of the current time window.

Another interesting research conducted by Kossinets and Watts (2006) which focused on investigating the influencing factors of evolution on a much larger community in contrast to the previous one (a large university, 43553 participants) based on a similar setting (i.e. e-mail exchange) accumulated over 355 days. In this research, the results revealed that the network's evolution is dominated by a combination of both standard SNA measurements (e.g. average degree, average clustering coefficient etc.) and two sets of proposed metrics, i.e. 1) cyclic closure: *"the empirical probability that two previously unconnected individuals who are distance apart in the network will initiate a new tie."* and 2) focal closure: *"the empirical probability that two strangers who share an interaction focus (in the present case, a class) will form a new tie."*. Those measurements were calculated and presented on basis of three different time intervals, namely, 30, 60 and 90 days over the entirety of their data set.

Furthermore, a study presented by McKerlich, Ives and McGreal (2013). In this research, the context is the longitudinal network data (210 days) obtained from an online community of the University of California, with its user base as the enrolled students of the university (1899 users & 59835 messages). This community was aimed to facility social activities among students in order to help them enlarge their circles of friends. Unlike e-mail communication mentioned previously, this online community allows users to create their own profiles containing personal details, and offers search functions ob basis of the information provided by those profiles, it also provides additional insights such as visit counts on the user's profile

(i.e. an indication of popularity). The objective of this research is to investigate the network structure and its evolution, in terms of examining the pattern of users' behavior and social interaction of the given research context. In this study, the researchers applied multiple standard SNA metrics such as node degree, average length path, reciprocity as well as a proposed measure "acquaintances" on basis of out-degree. The results were calculated and presented on basis of three different time intervals, i.e. two-week lifespan, three-week lifespan and six-week lifespan.

2.3 Structural characteristics of online communities

In this section, we discuss a number of important structural characteristics and factors regarding social networks and online communities, that are relevant in the given context of our study. Those structural characteristics will influence the foundation of how SNA metrics are calculated, and is therefore important to be elaborated in details.

Actors and relations

As elaborated by Wasserman et al. (1994) and Hanneman et al. (2005) actors or nodes are partial elements that constitute a network. In the context of online communities, actors are usually referred to as individuals within the community; they could also be referred to as corporate, organizations or other collective social units depending on the given research context. A basic example of a social network is displayed in Figure 2.1, A, B and C represent the actors of the network and E1, E2, E3 represent the relations among actors.

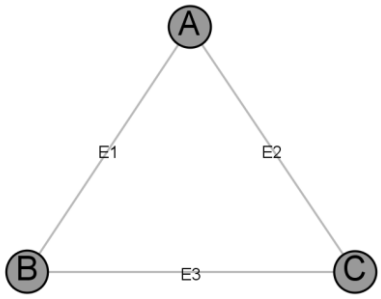


Figure 2.1: Example of a network

Next to the actors in the network, are the relations [edges] between the actors. There may be multiple relations involved between the same pair of actors: e.g. person A is a friend of person B (*friendship tie*), and at the same time, person A replies a message to person B in an online forum (*activity tie*). Borgatti, Mehra, Brass & Labianca, (2009) conducted a holistic review over the existing typologies of relations studied in SNA, and have identified four major types of relations: 1) relations based on similarities, 2) social relations, 3) interactions and 4) flows (Table 2.1). However, the vast majority of SNA studies in this field still utilize a single kind of relation per study (Howison et al., 2010) with very few exceptions (Kazienko, Musial, & Kajdanowicz, 2008).

Similarities			Social Relations				Interactions	Flows
Location	Membership	Attribute	Kinship	Others	Affective	Cognitive	e.g., Talked to Advice to Helped etc.	e.g., Information Beliefs Personnel Resources etc.
e.g., Same spatial and temporal space	e.g., Same clubs Same events etc.	e.g., Same gender Same attitude etc.	e.g., Mother of Siblings of	e.g., Friend of Boss of etc.	e.g., Likes Hates etc.	e.g., Knows Knows about etc.		

Table 2.1: A topology of ties studied in social network analysis

Directional network and nondirectional network

As illustrated by Hanneman et al. (2005, p55), a social network may be a symmetric (nondirectional) network or asymmetric (directional) network, i.e. if a relation between two actors in a network is mutual, it would be referred as a "*bound tie*" and is therefore nondirectional.

A visualized example is given in Figure 3.1. where A and B share a mutual relation E1, and so do A and C (E2) as well as B and C (E3). The relations in such networks can be easily implicated by using binary representations i.e. 1 or 0, which means that the relation between them either exists or it does not.

Asymmetric (directional) networks on the other hand, imply that the relations between actors are either reciprocated or not. The example shown in Figure 2.2 demonstrated this feature. The relations (A, B) and (A, C) only indicate that there are edges built from A to B and from A to C (presented in arrowed edges), in a source to target oriented fashion, but not vice versa.

There is arguably a natural fitting between different types of networks and their applicability to the context of various research domains. If we adopt the topologies elaborated by Borgatti et al. (2009), similarity and social relations would fit well with nondirectional networks; e.g. friendship, partnership and collaboration oriented relations are often mutual among involved individuals or entities, and are therefore symmetric (e.g. Otte et al., 2002).

Interaction and flow oriented relations are often directional; e.g. e-mail exchange, post & reply-to and passing information along and so on, are often initiated from one to another, and are therefore asymmetric (e.g. McLure Wasko & Faraj, 2005; Trier, 2008). Additionally, the implementation of directional and nondirectional networks will also influence the outcomes of social network analysis techniques.

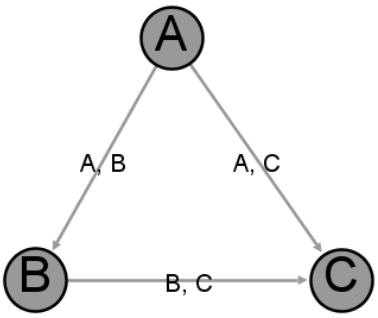


Figure 2.2: Example of a directional network

Networks and weights

The weight of a relation is often referred to as the *strength* of a tie. Similar to the concepts discussed in the previous section, weight of a relation could also be presented in two different forms, i.e. dichotomous or valued (Petróczi, Nepusz & Bazsó, 2006).

In case of dichotomous weight, the weight of a relation between two actors either exists or it does not. Valued weight on the other hand would carry a depicted value (in number or other assigned measurements) based on the tie-strength component. Many scholars argue that constructing networks with weighted relations would provide additional insights regarding the true structure of social networks (Toivonen et al., 2007; Howison et al., 2010).

Empirical studies have shown that it is difficult to assign appropriate values to certain types of relations such as friendship (Petróczi et al., 2006), and it is even more complicated to evaluate whether the assigned value does indeed represent the true value of the relation in a quantitative manner (e.g. how to compare and evaluate the strength of ties for friendships of 1000 people?).

Such difficulties seem to be less problematic in other types of relations, such as interaction or flow oriented relations. For instance, a commonly used weight measure for relations is based on the frequency of interactions among pairs (Figure 2.3). The example graph on the left hand side shows that A approached B 3 times (visualized as a thicker arrowed line) and approached C for 1 time, and B approached C for 1 time. In the matrix representation, the exact same results are displayed. Technically speaking, if the strength of a tie decreases to 0, it would have the same implication as not having a tie at all.

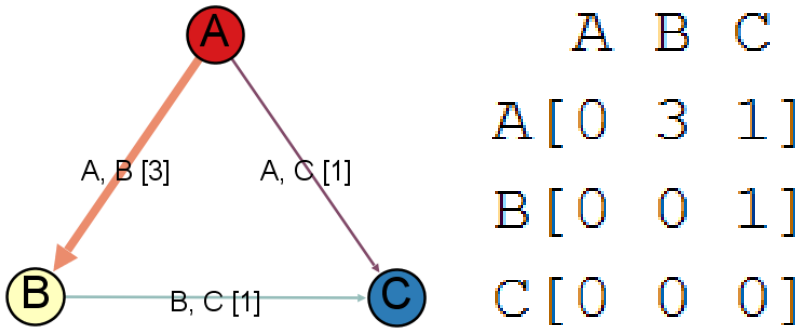


Figure 2.3: An example of directional and weighted network in graph and matrix forms

Other methods have been introduced recently, regarding the weight assignment procedure in the construction of network. Howison et al. (2010) suggested that the volume of information flow (e.g. number of texts in the message) could be considered as an additional factor to identify the intensity of weight. Further, Wiggins, Howison, & Crowston (2008) approached this issue from a different angle, i.e. a time based decay, which implies that more recent interaction would weigh more in contrast to older interactions.

Petrovčič et al. (2012), finally, developed a fascinating method that applies the time elapsed between messages of a thread (in an online forum structure), in order to determine the weight of an edge. This provides a new perspective to SNA in online communities with limited information about relational references. Unfortunately, the general applicability of their weight assigning method towards other research contexts is still quite limited.

2.4 Network creation in online communities

As discussed in the previous section, building a network in a nutshell is nothing more than connecting [relations/edges] the dots [actors/nodes]. The network can be constructed with nondirectional or directional relations; the relations within the network can either carry a weight or not. From a helicopter point of view, utilizing data sources for constructing network structures for social network analysis can be grouped by two distinguishable features: 1) structured data and 2) unstructured data.

2.4.1 Network creation from structured data

With traditional data collection methods for SNA, such as interviews, surveys, observations or studies of archival records, the researchers are able to identify the relation between a pair of actors based on an explicit indication, e.g. person A indicates the existence of a friendship tie with person B in the survey (Howison et al., 2010).

Such explicit indications can also be found in studies regarding online communities; e.g. in the context of an organization, the email addresses could be utilized as identification mechanism and a directional relation is created between the pair, if person A sends an email to person B (Trier, 2008). A similar case can be illustrated in the context of social media, a nondirectional relation is created if person A and person B share a friendship status on Facebook (Ang, 2011).

In this way, networks are created on basis of repetition of aforementioned processes over the available data source of the research, i.e. establishing relations between actors when the [explicit indication] condition is met (Fisher, Smith & Welser, 2006). In the context of this research, we refer to this type of data set as *structured data*.

2.4.2 Network creation from unstructured data

Convenient, structured, data sets are however not always available. Data retrieved from other types of online communities such as online forums (e.g. BBS) do not always have explicit or strong relational references regarding relations among actors (Sack, 2000). This poses a challenge to scholars who are interested in utilizing this enormous amount of quantitative data generated by such online communities (Petrovčič et al., 2012). In our research, this type of data set is referred to as *unstructured data*.

There have been a number of attempts towards solving the issues mentioned above with regard to unstructured data extracted from online communities. Berger et al. (2014) conducted a research to investigate the role and impact of key users in knowledge-intensive online communities regarding internal information and knowledge sharing, by using qualitative text analysis and standard SNA techniques / measurements (e.g. degree centrality, closeness centrality etc.). In this particular case, the data was obtained from a multinational corporation that used Yammer.com, a cloud based Enterprise Social Network (ESN) platform. To tackle the issue illustrated previously (i.e. without explicit relational references), they elaborated a solution, which proposes to create directional relations between all actors on the basis of grouping attributes such as group_id or thread_id, in case of relations that are not explicitly indicated (e.g. message sender_id is available, but the receiver_id is missing).

Another approach developed by Toral, Mart ńez-Torres and Barrero (2010), in the context of an online community for open source projects (i.e. an online forum). The research objective is to analyze the behavior of this online community, in terms of knowledge sharing, improving the underlying projects as well as the interactive collaboration among its participants, by applying SNA techniques to identify the members that uphold knowledge / information broker's role. Besides the grouping attribute, this network creation approach focuses on an additional attribute to the network creation method, namely a timestamp. With the help of a timestamp on each post, the sequence of posts can be pre-arranged in a chronological order, i.e. from the oldest to the newest.

From there, directional edges can be built from the latter to all previous posters within the same thread. To a certain extent, this approach can be seen as "reply-to preceding posts" while submitting a message in a thread. The essence of this approach as elaborated by Toral et al. (2010) is that *"in contrast to a reply to a single message, it is more cognitively complex to reply to a threaded discussion, because the ebb and flow of earlier postings must be taken into account to develop a coherent answer. That is the reason why an author posting to a thread will be tied to all the authors who have previously posted to the same thread when constructing the social network."*

In similar settings as the previous one (i.e. online forums), Faraj et al. (2011) conducted a research to measure the network interaction patterns of long-duration online communities from theoretical and practical perspectives, by identifying how individuals behave in network

level, in terms of direct reciprocity (i.e. direct interaction between a pair of actors), indirect reciprocity (i.e. indirect interaction between a pair of actors via a third actor) and preferential attachment (i.e. a concentration of communication). They illustrated yet another alternative network creation method. The posters within a thread are still organized in a chronological order based on the timestamp; the network is created on the basis of inbound links, every poster that replies in the thread will result in a directional edge created from the latter to its immediately prior poster within the same thread. This method represents inbound links as an expression of relationship formation in online communities.

Petrovčič et al. (2012) conducted a research specifically targeting this issue regarding unstructured data, in the context of a forum for student union at the university of Ljubljana. As many online communities (especially online forum oriented platforms) do not always provide data with explicit relational reference. In case of emerging discussions, it is difficult to identify who is replying to whom, in a conversation oriented structure. This research utilized the network creation method mentioned earlier (Toral et al., 2010), and proposed an algorithm to adjust edge weight properties of the pseudo network, by modifying edge weight according to a set of parameters that can be summarized as: the weight of an edge between a pair of actors depends on the number of messages between this pair and the time elapsed between this pair.

The resulting networks were then, compared against the baseline network which represented the reality. The findings of this research suggested that with appropriate parameter settings of time (i.e. time elapse between the pair) and space (i.e. number of messages between the pair), this approach can produce decent pseudo networks that are comparable to the reality in terms of correlation coefficients between them. However, the researchers of this study have pointed out that this research is *"a starting point for the development of a standardized methodology for studying social networks in online communities where only limited direct information about communication ties is available."*, and it has limitations regarding generalizability towards other online communities with different characteristics.

With this in mind, only three of the network creation methods discussed above are utilized and are codified as *method 1(Berger et al., 2014), 2 (Toral et al., 2010)* and *3 (Faraj et al., 2011)* in the context of this research. Visualized network creation methods are shown in Figure 2.4, and the details are described in Table 2.2. It is essential to notice that those network creation methods are constructed on presumptions of pre-existing literature, without further validation, and are therefore pseudo methods.

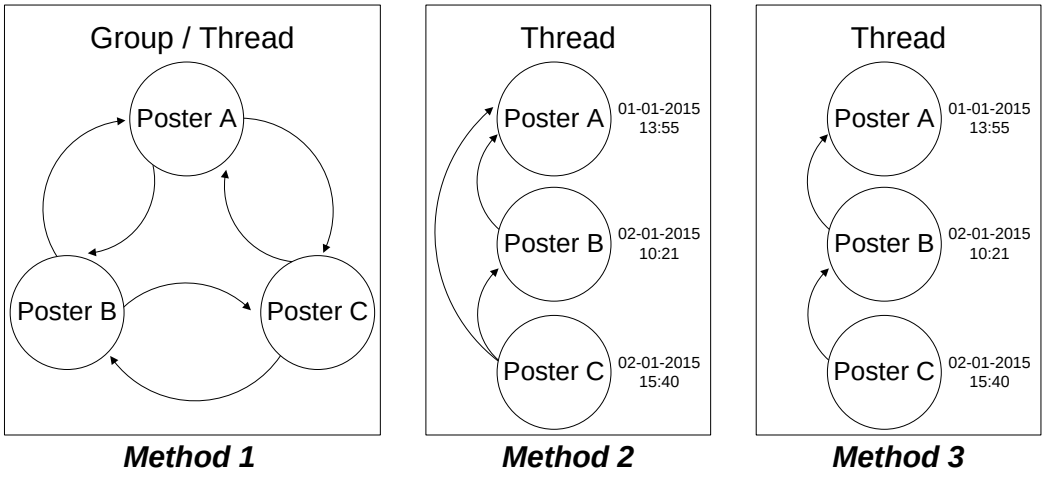


Figure 2.4: Visualization of three pseudo network creation methods

Method	Sender	Grouping attribute	Resulting edge(s)
1	A	group_id or thread_id	A → all nodes within the boundary set by grouping attribute
2	A	thread_id and timestamp	A → all nodes in thread_id and prior to the timestamp
3	A	thread_id and timestamp	A → closest node in thread_id and prior to the timestamp

Table 2.2: Details of three pseudo network creation methods

2.5 Longitudinal nature of online community

Traditionally, SNA data was/is collected over a short time, e.g. the researcher takes a week to conduct interviews and thereafter creates the network on basis of the results, i.e. only representing a single snapshot over that particular period of time -- in this sense, the network created from such a data set can be illustrated as a cross-sectional (i.e. static) network.

With the emerging interests in online communities, SNA in such contexts relies on data accumulated over a much longer period of time, i.e. the longitudinal nature of online communities, with time also being an important characteristic to represent the dynamic nature of online communities. For instance, the interactions among online forum participants are most likely to occur over time, which implies that without proper assistance, conclusions might be drawn based on misleading results (e.g. identification of central actors who had left the community a long time ago; Wiggins et al., 2008).

To investigate the longitudinal nature of networks created from aggregated data of online communities, additional methods would be required to incorporate with SNA techniques

(Christley, & Madey, 2007). In the following sections, we will discuss a few symbolic approaches based on the results of literature review in this regard.

2.5.1 Actor centric

Snijders (1996) introduced an actor-oriented method for investigating the dynamic nature of a social network. In his paper, the focus is on users and their actions as the driving force of network evolution, in such a way that the ambition or preference of a user would impact others within the network via the outgoing relations towards them.

A major advantage of this method is that it offers great insights with regards to the static and dynamic characteristics of the actors in the network. However, a constraint is also present, namely the requirement for continual and intensive observation of the same group of actors.

This constraint is arguably less problematic in a certain context, e.g. a closed network of a commercial organization, where employees do not leave the organization permanently on a frequent basis (i.e. a static network in terms of user base). Public online communities on the other hand, are often unable to fulfill this requirement as the users are allowed to leave at any given time as they see fit (Dignum & Eeden, 2005; Varik & Oostendorp, 2013). Therefore, the applicability of this method would be rather situational depending on the research context.

2.5.2 Time centric

Another approach to resolve this issue is to look at it from a time perspective. This a commonly applied method when investigating the longitudinal nature of networks. It involves constructing networks for consecutive time periods, and therefore create a time based series of network structures, or by segmenting the whole network into network "snapshots" in a chronological order. In such way, the SNA measurements can be calculated and presented on a periodic basis, and that could provide additional insights regarding the changes of social positions among online participants over time (i.e. temporal analysis; Christley et al., 2007; Howison et al., 2010) in contrast to a holistic global analysis (i.e. cross-sectional analysis). Moreover, this approach can also be utilized for visualization of the online community's evolution (i.e. cumulatively or dynamically; Moody, McFarland & Bender-DeMoll, 2005), or to demonstrate the visual temporal communication patterns of various types of collaborative knowledge networks (Gloor, Laubacher, Zhao & Dynes, 2004).

Such "snapshots" of network fragments are extracted on the basis of continuously non-overlapping time intervals continuously over the entire data set of the community, e.g. segment_1 is extracted between date x and y, segment_2 is extracted between date y and z and so forth. This approach offers in-depth insights regarding statistics and measurements of network snapshots over their respective time interval, and enables the researchers to analyze

the development and characteristics of online communities in terms of trends and evolution of the network, by revealing the changes of interaction patterns over time (Falkowski, Barth & Spiliopoulou, 2008) or the positional change of core members at the center of the community (Howison, Inoue & Crowston, 2006). Surprisingly, there is an alarming issue in the existing literature that have applied this method for data segmentation, i.e. the duration of time interval sizes used in those studies are often not well rationalized, or justified; the choices were made based on convenient colander units (e.g. day/month etc.) or an arbitrary division mechanism (Panzarasa, Opsahl & Carley, 2009; Howison et al., 2010).

Yao, Zhou, Han, Xu & Lü (2011) conducted research on data sets retrieved from Flickr and Epinions regarding the static and dynamic characteristics of online communities. In this research, the goal is to investigate network structures created on basis of different relations (i.e. social relations and user interactions) of those online communities, in terms of comparing standard SNA measurements between the networks created based on social relations and user interactions. The findings of this research indicated that several SNA measurements between these two types of networks have several measurements in common, however, showed significant difference in degree correlation (i.e. it reflects the frequency of nodes with similar degree that are connected to each other), which can be further illustrated as a key distinction between them. In their study, two division measures were used to analyze the data from a longitudinal perspective, i.e. 27 snapshots of network fragments were created on the data set with 104 days time span from Flickr (1 snapshot covers 3.85 days), and 31 snapshots were created on basis of 31 months time span from Epinions (1 snapshot covers a month). However, these choices were not explicitly elaborated nor are they scientifically sound. Such arbitrary decisions may lead to inconsistent or unstable construct as discussed in chapter 2.3, Threat to validity.

There are three interesting observations found on basis of the studies elaborated above, which could provide useful guidelines, with regard to longitudinal SNA when using a time centric approach for this research:

1) The time interval for creating the network snapshots is set on either a constant value (x number of days) or a calendar unit that may vary slightly (e.g. a month),

2) The network snapshots are created on the principle of a consecutive and non-overlapping manner, over the entirety of the data set available, based on the value of time interval mentioned in the first point.

3) The duration of a time interval is determined on a situational basis depending on the given research context, and is often not well rationalized. However, the choices do have certain commonalities overall, i.e. the duration of a time interval must be long enough, so that each of the network snapshot will contain sufficient number of nodes and edges for SNA techniques to produce sensible results, and the duration should not be too long, otherwise, it will diminish the quality of the results from a longitudinal perspective.

2.5.3 Event centric

A third method is elaborated by Trier (2008), which is an event-based method not only for longitudinal analysis, but rather a comprehensive set of techniques and tools for visualization and analysis of dynamic networks i.e. for analysis of dynamic networks in general. The fundamental difference of this method is that it *"is disaggregating relationships into ordered series of timed events, and explicit recognition of variety of event and actor attributes"* (Trier, 2008). This method emphasizes the evolutionary aspect of networks, and takes external events as its driving force. In this way, the longitudinal nature of networks can be monitored and analyzed in a holistic manner.

Nevertheless, it would demand certain traceable information to be available regarding the external events that occur and impact the network, as this method is created on an event-driven basis. This can be done via utilizing event indicators such as hash-tags on Twitter, through prior knowledge of the community (e.g. sudden increase in login rate of members and the number of interactions among them on a basketball fan forum, is caused by a basketball game day), or by conducting intensive content analysis.

2.6 Metrics for describing structures of online communities

This section focus on non-standard network metrics that have been used to describe network structures of online communities. There are many forms of network structures, in terms of relations (e.g. social relations or user interactions) or technical settings (e.g. social network sites, online forums etc.). In the context of this study, we are interested in one type of online community, namely the online forums. Online forums have a specific hierarchical structure. On the higher level, the main or sub-categories are usually separated by the topic of interests (e.g. sports, business etc.) or by the distinctions between their functions (e.g. technical support, discussion, announcement etc.). On lower level, messages are grouped by threads that have been posted in one of the main or sub-categories. In the following sections, we discuss a few studies that have emphasized the special metrics for measuring this type of online community.

Helms et al. (2015) conducted a research on the same data set of this thesis project (details of the research context is discussed in section 3.1), to investigate the user interaction patterns of this online forum from a helicopter's point of view, regarding its longitudinal development by examining the interactions among different groups of users. For this sole purpose, they have developed an approach called *"edge-ratio analysis"*, with a set of metrics that measure the interactions between old and new users (in relative terms) on a periodic basis. The data extracted from this online community had eight sub-categories grouped by the topic of interest, each sub-category was segmented into eight periods (i.e. 8 network snapshots), with each segment presenting a calendar month of network data. The results of this research

revealed the importance of interactions among old users and a constant stream of new users in terms of a healthy and sustainable growth of an online community.

This set of metrics used in edge-ratio analysis is described below:

1) Number of new users: number of users posting for the first time in a network snapshot.
2) Number of old users: number of users posting in a network snapshot, but have also posted in an earlier network snapshot.
3) New user's percentagewise: relative amount of new users in comparison to the total number of active users in a network snapshot.
4) Edges between new users percentagewise: relative amount of edges between new users in comparison to the total number of edges in a network snapshot.
5) Edges between old users percentagewise: relative amount of edges between old users in comparison to the total number of edges in a network snapshot.
6) Edges between old and new users percentagewise: relative amount of edges between old and new users in comparison to the total number of edges in a network snapshot.
7) Impact of new users: ratio of "edges between new users" and "new users".

Adamic, Zhang, Bakshy, Ackerman and Arbor (2008) performed another research on a large and diverse online community, namely Yahoo Answers. The community is a question-answer oriented online forum, and holds a hierarchical structure. This research focuses to investigate the knowledge sharing activities by grouping the data based on the characteristics of its contents and the interaction patterns among its users. The objective is to provide a systematic approach to predict the quality of user interactions. One of the important metrics that used for this particular purpose, is the combination of thread length (reply/thread ratio) and post length (how verbose the answers are). The thread length in this context is illustrated as a important metric to represent the foundation of this online forum's structure, as this metric is used to create the network for SNA (e.g. in and out degrees). The post length is utilized to examine the quality of the contents in terms of knowledge sharing among the users. However, despite the longitudinal nature of this online community, the results were presented in a cross-sectional manner.

2.7 Network metrics

In this section, we present definitions of a set of standard network metrics regarding SNA in online communities in the context of this research. The taxonomies applied in the following section are mostly based on the works of Hanneman et al. (2005).

Network density

In the context of a weighted network, density can be defined as the sum of relations divided by the number of possible relations (Hanneman et al., 2005). In contrast to nondirectional networks, directional networks, with the same number of actors in theory can have twice as many relations. The density of a network may offer insights regarding the speed at which

information diffuses among actors within the network, with further implications that may help in identifying actors with high a level of social capitals (Hanneman et al., 2005).

Network diameter

As defined by Hanneman et al. (2005), "*the diameter of a network is the largest geodesic distance in the (connected) network.*". The diameter of a network implicates how big the network is, not in the sense of counting the number of actors and relations, but in the sense of how far the information needs to travel between two farthest actors on each side the network.

Clustering coefficient

The clustering coefficient of a network measures the degree to which actors in the network tend to cluster together (Hanneman et al., 2005), and it quantifies how well connected are the neighboring actors of an actor are within the network (Soffer & Vázquez, 2005). There are three versions of this measure: global clustering coefficient, local clustering coefficient and average clustering coefficient. Global clustering coefficient measures the degree of clustering in the local neighborhood of a given node (Watts & Strogatz, 1998). Local clustering coefficient refers to the tendency of nodes to cluster together on the network level (Luce & Perry, 1949). At last the average clustering coefficient is measured by taking the local clustering coefficient mean on top of all nodes in the network (Watts et al., 1998).

Average path length

The average path length is defined as the mean of all shortest paths between any pair of nodes in a network, where the shortest path is referred to as the geodesic distance between two nodes in a network (Hanneman et al., 2005). It is an essential concept in SNA, as many important centrality calculations rely on the results of it (see section Centralities). Finding the shortest paths in an unweighted network setting is rather simple by using the definition mentioned above. However, in case of weighted network settings, the shortest path cannot be calculated simply based on the number of ties between a pair of nodes, the weight attribute has to be taken into consideration as well.

Reciprocity

In SNA, reciprocity illustrates the tendency of relations between actors in a directional network setting to be reciprocated (Hanneman et al., 2005). This measure only makes sense in a directional network setting, whereas in the context of nondirectional networks, all existing relations among actors are mutually presented, and is therefore reciprocated. As elaborated by Borgatti, Everett, and Freeman (2002), reciprocity in the directional network setting can have three different versions, i.e. dyad-oriented, arc-oriented and hybrid. The hybrid version of reciprocity is based on the combination of the first two. Dyad-oriented reciprocity calculates the ratio between the number of adjacent nodes connected by a tie and the number of adjacent nodes connected by a reciprocated tie; whereas arc-oriented reciprocity presents the ratio between the reciprocated ties and the total number of ties in the network.

2.8 Centralities

Centrality is an important concept in this field of research. It helps researchers to identify the most central node, which in turn could improve the rate of disseminating information in the network or to prevent a network from breaking apart (Newman, 2010). In this section we discuss different types of centralities, their characteristics and applications with regards to SNA.

Degree centrality

In a nondirectional network, the centrality can be calculated on basis of the number of ties a node has, but in a directional setting, the degree centrality can be separated into two categories: 1) in-degree and 2) out-degree (Figure 2.5). The in or out degree can then be measured by calculating the sum of incoming or outgoing ties of any node within the network. (Hanneman et al., 2005). In-degree could provide preliminary implications regarding popularity of nodes, e.g. a node with large number of incoming ties could be interpreted as a popular individual in the network. Out-degree on the other hand, is useful in case of identifying influencing individuals within the network (Hanneman et al., 2005).

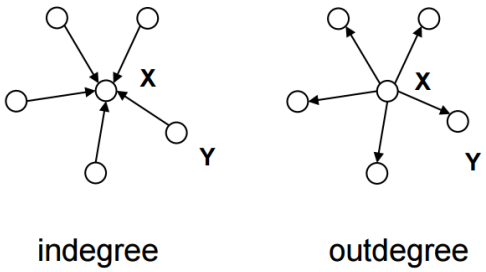


Figure 2.5: Degree centrality

Betweenness centrality

As elaborated by Freeman (1977), the betweenness centrality of a node, is described as the frequency with which it appears on the shortest path between pairs of nodes in the network, i.e. how many pairs of actors would have to go through a particular actor to reach one another in the shortest path. Nodes with high betweenness centrality can be argued to have more power or influence than others in the same network, which is often referred as brokers due to their central positions within the network (Figure 2.6; Hanneman et al., 2005).

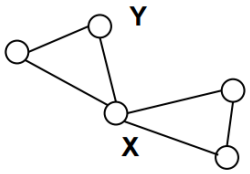


Figure 2.6: Betweenness centrality

Closeness centrality

The closeness centrality of a node is defined as the inverse of the sum of distances (Sabidussi, 1966). In other words, it is calculated on the length of the average shortest path between a node and all nodes in the network. What if it is not so important to have many direct friends (degree centrality), or "between" others (betweenness centrality), but still be in the middle of things, or not too far from the center (Mascolo, 2011)? Arguably, nodes that carry higher scores in terms of closeness centrality can be viewed as central actors with a network, due to the fact that they are closer to other actors in the network, and are therefore able to disseminate information faster in comparison to nodes with lower score of closeness centrality (Berger et al., 2014).

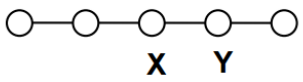


Figure 2.7: Closeness centrality

Eigenvector centrality

Unlike other centralities discussed in previous sections, eigenvector centrality does not derive the structural characteristics directly from a node standpoint, but is rather determined by the centrality of its neighboring nodes, i.e. *"someone who has powerful friends"* (Bonacich, 1972). In this sense, nodes with high eigenvector centrality are regarded as central actors by having popular neighboring nodes (Bonacich & Lloyd, 2001).

3 Experimentation

In this chapter, the objective is to apply findings of literature review in practice, to the data set at hand. In the first section, the research context is introduced. The second section focuses on the pre-data processing of the data set obtained from the research context. The last two sections involve various experimentations regarding social network creation processes as well as adaptation of longitudinal analysis in online communities, from a practical point of view.

3.1 Research context

The *Hallo! community* is an online forum of the initiative of the Dutch Chamber of Commerce (Kamer van koophandel), with the slogan "*Share your business knowledge here*" (Figure 3.1). It is a Dutch language free and public online forum for all entrepreneurs (mostly Dutch oriented) for knowledge sharing, exchanging experiences, asking questions and providing answers to each other regarding any business related activities (Hallo! Community, 2015).

Like many other online forums, people can view existing topics on Hallo! forum freely without registration; however, in order to submit a thread or a post, registration is required. In the following sections, we will discuss the structural nature of this online community as well as the data collection for this research. As elaborated in a secondary source (Majdán, 2012), user activities were closely monitored by forum moderators. In cases such as the posting of inappropriate content, or spamming, user created content was removed by the forum moderators.

From a technological perspective, the forum does not support functionalities such as "reply-to" or "quote". Although users may utilize such styles when posting a message, those digital traces will not be recorded systematically by the forum. As these relational references are missing from the data set, this poses a challenge in creating networks/graph structures from the forum data.

Figure 3.1: Screenshot of the home page of Hallo! community

3.1.1 Forum structure

Over the years, the Hallo! community forum has evolved over time -- its categories and sub-categories of interest having been extended as well as further refined in comparison to its original form in 2009. In all those years though, the main hierarchical structure of Hallo! community remained intact. The forum is constructed on a three-layered hierarchical structure; utilizing the taxonomy illustrated by Stanoevska Slabeva (2002), we describe Hallo! community as a *Topic-oriented Discussion Community*.

The first layer in the forum's hierarchical structure is the main category, which consists of 6 categories. Underneath this main category, we find the sub-categories with each main category having 6 sub-categories, except one. Threads can only be posted by registered users in one of the sub-categories. Those main categories as well as sub-categories are interest oriented, e.g. "Communication & Marketing", "Finance" and so forth.

Registered users are able to post replies in any threads, and from the forum's perspective, a thread itself is also considered not only a post, but the first one. An illustration of the forum structure is displayed in Figure 3.2.

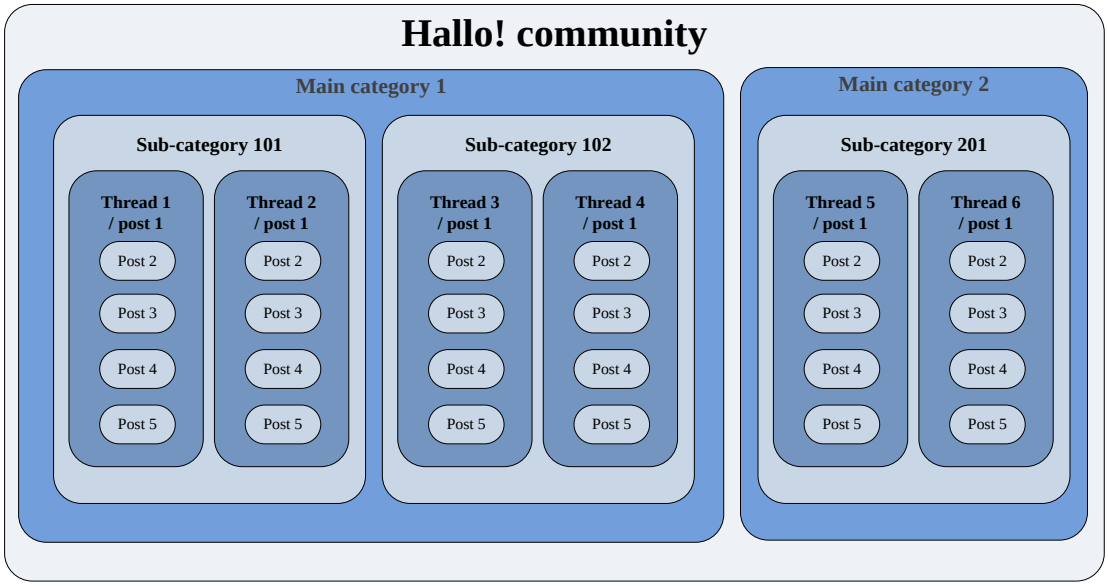


Figure 3.2: An example structure of Hallo! community forum

3.1.2 Data collection and basic descriptive statistics

In order to investigate the longitudinal development of Hallo! community, we collected approximately two years of data for this community. The files were retrieved directly from Hallo! community's database into two separate excel spreadsheets. The data origins are a mixture of system-generated data as well as user created contents (Howison et al., 2010).

When new users register on Hallo! community, some of their personal information is made available to the community hosts. As mentioned in sub-chapter 2.3 Thread to validity, we take ethical issues in our research very seriously, any sensitive or confidential information regarding user anonymity or which may damage the confidentiality terms is excluded from this research.

According to Majdán (2012), Hallo! community was officially launched in March 2009, though the forum was online two months earlier, with tests regarding its technical functionalities being carried out. As of 25-01-2011, there were almost 36000 registered users on Hallo! community with 7662 users having contributed to the community by posting at least a thread or a reply to a thread in this community; 12776 threads were posted, and those threads received 45682 replies to them across all 35 sub-categories. Mode detailed descriptive statistics is shown in Table 3.1.

Number of registered users	35972
Number of active users	7662 (21.3%)
Number of main categories	6
Number of sub categories	35
Total number of threads	12776
Total number of posts	45682
Average reply rate per thread	3.58 replies
Average activity per active user	7.63 threads/replies
Longest thread (maximal number of replies of a thread)	571 replies

Table 3.1 Descriptive statistics of Hallo! community

3.2 The data set and data pre-processing

The sole purpose of data pre-processing is to create a "clean" version of the data sets we have at hand for further use in this research. As mentioned in previous section, the data extracted from Hallo! community contains approximate two years of system-generated records as well as user created content; which implies that the quantity of the data sets can be considered relatively large: 20 separate data sheets stored in two excel files, containing over 100 columns and a few hundred thousands of records in total. However, only some of the data is required to build the social networks, for investigating the longitudinal development of Hallo! community. In the following sections, we discuss the procedures that have been performed in order to create the tidy-data set for this research.

3.2.1 Raw data

The sets were first examined closely, to verify whether they can be classified as raw data. The verification mechanism used for this process is based on the data processing guideline elaborated by Leek et al. (2013) and is shown below:

1) Ran no software on the data,
2) Did not manipulate any of the numbers in the data,
3) Did not remove any data from the data set,
4) Did not summarize the data in any way.

Based on our extensive observation by using the criteria illustrated above, there is no significant trace, suggesting any of the data records have been deliberately manipulated, removed, or summarized, and can therefore be classified as raw data.

The raw data set is stored in the format of excel files. As mentioned earlier, these files were exported directly from Hallo! community's database. As a result, the data from the original

files possesses a relational database structure, i.e. a table/column/row structure, and followed the principles illustrated in the data processing guideline (Leek et al.,2013; Table 3.2). There is no further need to reformat their structures.

Principle	Description
1	Each variable you measure should be in one column,
2	Each different observation of the variable should be in a different row,
3	There should be one table for each kind of variable,
4	If you have multiple tables, they should include a column in the table that allows them to be linked.

Table 3.2: General principles of a tidy data set

3.2.2 Preliminary data processing

The goal in this stage is to create a raw tidy data set for further examining regarding the data quality and the contents of the data.

Selection of columns

As mentioned in previous sections, the data sets at hand consist of over a hundred columns and hundreds of thousands of records (though not all of them are needed to create the network structures for this research). As the next step, we identified a set of indicators that would provide enough information on the basis of our research into the pseudo-network creation methods discussed in the literature review. These indicators are listed below:

1) Anonymous user identifier
2) Thread identifier
3) Category identifier
4) Links between aforementioned identifiers
5) Chronological indicator

The *anonymous user* identifiers represent a set of unique nodes in the context of this research. *Thread identifiers* are the most important indications of user interactions, as the interactions [edges] of users [nodes] will be created upon and grouped by thread identifiers. *Category identifiers* are used as a higher level of grouping attribute, to specify how users are grouped by their category of interest, and therefore illustrates the forming of an online community. *Links* between different identifiers are needed for technical purpose, so that different data retrieved from various data sheets can be merged as a single file for the network creation process in a later stage. Last but not least, the *chronological indicators* are needed as an assumption for network creation based on unstructured data.

As a result, a list of selected columns from various data sheets has been created. Table 3.3 below contains all details regarding the selected data columns from the raw data sets,

duplications in attribute names are indications of links between different files. An additional representation of their relationships is illustrated in Figure 3.3, in the form of an ER-diagram. Based on the results of our analysis, "PostauthorID" in file Thread is equivalent to "UserID" in file User and "UserID" in file Post. Attributes that have been used to establish links between different files are shown below, the text before dot indicates the file name, and text after dot indicates the attribute name:

- Category.ThreadID = Thread.ThreadID
- Thread.PostauthorID = User.UserID
- Thread.threadID = Post.ThreadID
- Post.UserID = User.UserID

File name	Attribute name	Data type	Attribute description
User	UserID	Continuous	System generated unique identifier of an user
	Username	Categorical	Forum alias of an user
Thread	ThreadID	Continuous	System generated unique identifier of a thread
	PostauthorID	Continuous	System generated unique identifier of an user who initiated this thread
	TotalReplies	Continuous	The total number of replies within a thread
Post	ThreadID	Continuous	System generated unique identifier of a thread
	UserID	Continuous	System generated unique identifier of an user
	Subject	Categorical	Subject of a post
	Body	Categorical	Content of a post
	PostDate	Date/time	Timestamp of a post
Category	ThreadID	Continuous	System generated unique identifier of a thread
	CategoryMain	Categorical	Main categories of the forum
	CategorySub	Categorical	Sub-categories of the forum
	Cate_ID	Missing	A codified ID of a sub-category

Table 3.3: Overview of selected data columns

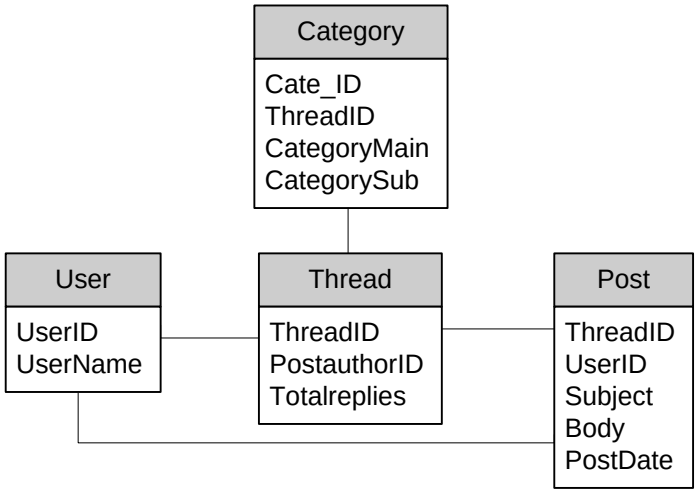


Figure 3.3: ER-diagram of the selected data sources

Filtering, filling and merging

Based on the previously created data sheets, we now are able to perform preliminary data pre-processing. The main objectives here are:

1) to eliminate non-contributing data based on known factors,
2) to extend existing data sheets in a way that could further enhance the network creation process,
3) to merge separate data sheets into one single file as a preliminary version of the tidy-data set, which can be used in later stages.

As discussed in section 3.1.2, Hallo! community was officially launched on March 2009, therefore threads posted before the date 01-03-2009 were omitted on basis of the attribute "Postdate"; i.e. if the post date of the first post of a thread is before 01-03-2009, it will be omitted. Secondly, on the thread level, attribute "TotalReplies" indicates the number of replies a thread contains. Threads with zero reply were removed, due to the fact that a single node alone, without any sign of interaction, cannot be used for network creation. Thirdly, threads with replies are double checked. In case of threads with no other participants than the thread starter himself, such threads will then be removed from the context of this study as well, since spinning around oneself does not provide much value to the community. Finally, on the category level, a thread that is not associated with any category, is treated as system error and is therefore removed; as according to the technical constraint, a user cannot post a thread without indicating a sub-category provided by Hallo! community.

The data sheet category does not contain its own unique identifiers. Therefore such identifiers needed to be assigned to all unique combinations between the main and sub-categories. As discussed previously, each main category contains 6 sub-categories (with one exception). This constitutes 35 unique combinations; values from 101 to 106 and 601-605 are assigned to each of the unique combination among sub-categories respectively. In this way, the string values of two data columns in category data sheet are combined into a single categorical value, which will enhance the analysis process in the following stages.

Next, the modified data sheets are merged into a single sheet. Such action can be clarified by using the relational database term "Equi-join", which implies that all join conditions must be met among all data sources (e.g. Figure 3.4.). The description of the final attributes are listed in Table 3.4.

```
SELECT *
FROM thread trd
,    post   pst
WHERE trd.threadid = pst.threadid
;
```

Figure 3.4: Example of Equi-join

Attribute name	Data type	Attribute description
ThreadID	Continuous	System generated unique identifier of a thread
UserID	Continuous	System generated unique identifier of an user
PostDate	Date/time	Timestamp of a post
Subject	Categorical	Subject of a post
Body	Categorical	Content of a post
Username	Categorical	Forum alias of an user
PostauthorID	Continuous	System generated unique identifier of an user who initiated this thread
Cate_ID	Missing	A codified ID of a sub-category
TotalReplies	Continuous	The total number of replies within a thread

Table 3.4: Description of data attributes

3.2.3 Final data processing

In this stage, the objective is to perform additional data cleaning processes as well as to select a number of active sub-categories as the subject communities for network creation and SNA. Moreover, the data quality is to be verified according to the construct validity criterion, and research ethics, as discussed in sub-section 1.4 Threat to validity.

Category selection

Based on our observation of the previously created preliminary tidy data set, and its related descriptive statistics, an interesting pattern was observed: the number of messages among sub-categories is not evenly distributed. In some cases such as sub-category 306 "Internet, online marketing & sales" and sub-category 101 "Start-up", the number of messages posted in these sub-categories are extremely high (7647 and 6354 respectively), in contrast to sub-categories such as 206 "Resignation", 505 "Pension", 506 "Bankruptcy" (43, 52 and 29 respectively).

This observation poses an intriguing question: are these minor fractions worthy to be included in the context of this research? If so, then to what extent should they be included? As discussed earlier, Hallo! community was launched officially on 01-03-2009, and the last post of the data set is on 25-01-2011, which means that the preliminary tidy data set contains 695 days of messages. We argue that only the active sub-categories should be included in this research. The term active in this context is defined as following:

> *"A sub-category is only considered active, if there is at least one message posted in it per day on average, during the entirety of the data set (695 days)".*

The reason of this choice is that, across all the sub-categories of given research context, many of the sub-categories have very few threads and posts in them, over almost two years of time. There is not much added value to investigate sub-categories with low post rate, as the low

post rate will result in oversimplified network snapshots from a longitudinal perspective. With this definition, we select 15 sub-categories that satisfied the minimum requirement. The visualized presentation is shown in Figure 3.5. In this figure, the sub-categories are illustrated in a descending order on basis of total number of messages per sub-category. The red horizontal dash line indicates the minimal requirement of 695 messages on basis of 695 days, and the black vertical dash line indicates the borderline of whether a sub-category is selected for this research.

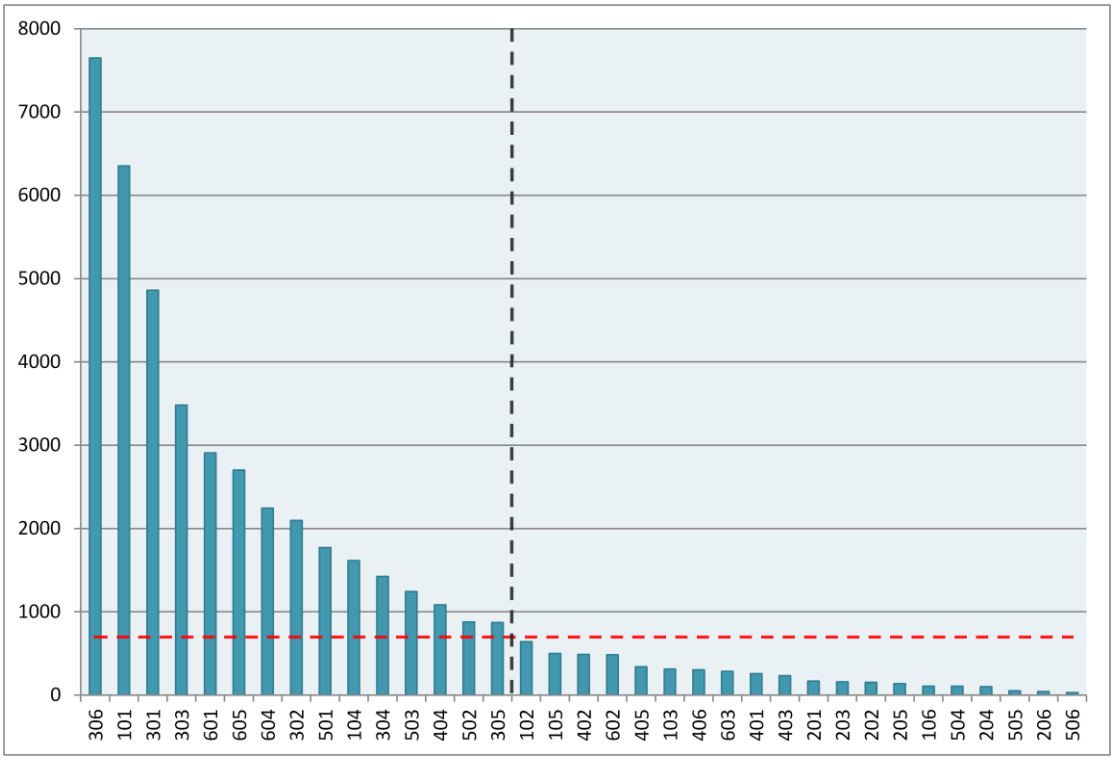


Figure 3.5: Total number of messages per sub-category on basis of 695 days (descending order by total number of messages)

Content control & Data quality

Next, the final round of pre-data processing is performed regarding content control as well as data quality control. The average reply rate of the whole Hallo! community is 3.58 replies per thread. However, there are 10 threads with at least over 100 replies in each of them. This symptom is rather counterintuitive in this context. Therefore we take a closer look at them to investigate whether those threads have violated the research ethics (e.g. might reveal the true identity of users) as mentioned earlier.

As a result of our observation, 8 of these threads contained information that may be considered as confidential and may reveal the identity of users (e.g. “Introduce yourself” , “Post your Linkedin account here” etc.). We decided to remove those threads and their corresponding replies from the data set.

Furthermore, we have identified 476 threads with bizarre errors, i.e. the first post (the thread

itself) has a post date later than the first reply, which is obviously not logical. The observed values of post dates on those error threads are mostly on 23-10-2009 and 26-10-2009 (299 occurrences), the remainder are scattered over to other dates throughout the years, till 21-01-2011.

The cause of these errors remains unknown. However their relative proportion (8.9% of the total number of threads), is rather significant. If such errors are not treated correctly, it may cause further damage to our study in terms of research integrity. Fortunately, there is an alternative way to retrieve the original post date of a thread, stored in the **Thread** file mentioned previously. All threads with post dates later than their respective first replies have been corrected retroactively.

As a result, the final version of tidy data set which contains details of 40908 messages, is now ready for network creation as well as for investigating the longitudinal nature of Hallo! community. Details regarding message counts and distribution between 15 selected sub-categories are listed in Table 3.5.

Category	Sub-category	Code	Number of threads	Number of replies
Company operations	Start-up	101	843	5743
	Telecom & ICT	104	215	1457
Communication & Marketing	Networking	301	663	3836
	Acquisition	302	279	1883
	Promotion	303	353	3190
	Product, price and services	304	188	1296
	Competition and market	305	132	778
	Internet, online marketing & sales	306	774	6006
Legal issues	Laws & regulations	404	140	978
Finance	Administration	501	241	1581
	Billing & collection	502	107	804
	Taxation	503	186	1109
Others	Hallo!	601	397	2528
	Offers & wanted	604	357	1933
	Others	605	436	2475
Total number of threads			5311	\
Total number of replies			\	35597

Table 3.5: Overview of selected sub-categories

3.3 Network creation

As discussed in section 2.4, Structural creation of online community, two types of data structures were identified on SNA in online communities. In the context of this research, the tidy data set we have at hand falls into the latter category, i.e. unstructured data. This tidy data set does not contain explicit indicators regarding relational references (e.g. sender_id and receiver_id). Available indicators for network creation are solely based on the grouping attributes and timestamps, which implies that pseudo network creation methods are to be applied.

However, in order to validate the applicability and goodness of fitting of existing pseudo network creation methods, we plan to investigate these matters comprehensively, by means of conducting content analysis on the data set, proposing an alternative network creation method as well as evaluating the goodness of fitting among networks created from various pseudo network creation methods.

3.3.1 Content analysis

As discussed in section 2.4.2, unstructured data extracted from online communities do not provide relational references between users in terms of their interaction patterns, which is a crucial element to construct social networks. Therefore, pseudo network creation methods were applied to simulate the networks based on the unstructured data for SNA. Furthermore, existing literature indicates that networks built upon unstructured data are often not validated in the sense of their accuracy to reality. Therefore, in this research, we conduct a series of content analysis on the data set at hand, in order to obtain sufficient amount of traces that could grant us insights about how users have interacted in reality. The main objective here is to produce evidence regarding the interactions among users of Hallo! community, by answering the following questions:

1) To whom is a reply meant for?
2) Can we identify this recipient by using existing pseudo network creation methods?
3) How much does a thread starter get involved after he has posted a thread initially?
4) To what extent can the pseudo network creation method be applied?

The goal is to analyze the data for gathering traces that could provide answers to these questions posed above. Traces can be hidden inside the body of a message, e.g. an indication such as @person_X or quoting a sub-section of a message from another person, but as mentioned in the beginning of this chapter, such indications are not recorded systematically by the forum, and therefore the traces must be obtained manually via content analysis on the data set.

Conventional content analysis and direct content analysis are the approaches utilized in this case. As elaborated by Hsieh and Shannon (2005), "*Conventional content analysis is generally used with a study design whose aim is to describe a phenomenon*" and "*Sometimes, existing theory or prior research exists about a phenomenon that is incomplete or would benefit from further description*". As illustrated in their paper, conventional content analysis begins with observation, codes are then defined during data analysis and are derived from data. Direct content analysis on the other hand, begins with reviewing existing literature, and the codes are defined before and during data analysis.

For this section of our research, we aim to study the interactions between users among various sub-categories of the community, based on existing literature as well as observations of our own. Therefore, both approaches are applicable in the given context. The final version of our tidy data set contains 5311 threads and 35597 replies, it would be impractical to analyze all of them. Therefore we will randomly select a number of samples to represent the whole data set.

The content analysis begins with a frequency analysis, to investigate the relative proportions of threads in terms of number of replies. This analysis is performed by using the statistical software tool IBM SPSS. As shown in Appendix A, 83.4% of the threads have ended within 10 replies, and only 16.6% threads have more than 10 replies.

This indicates that attention should be paid to threads with 1-10 replies, as the pseudo network creation methods rely on thread as their primary grouping attribute. However, threads with more than 10 replies must not be overlooked. Else, it would be impossible to provide a comprehensive overview regarding the community.

Samples

Conducting content analysis manually, is a very time consuming activity. Therefore, a few sets of samples have to be selected. In order to analyze the contents comprehensively for investigating the user interaction patterns, the samples have to be selected from a wide range of threads in terms of number of replies. In this way, the selected samples can be considered to represent the whole community. Additionally, as mentioned in previous section, the overwhelming majority of the threads in this data set have less than 10 replies to them. It is important to gain insights regarding whether the interaction pattern varies between short and long threads. The sample selection is done via three attempts, based on random selection in a non overlapping manner. This means that threads and their respective replies selected in each attempt, will not reappear in other attempts.

First, 50 threads per number of replies between 1 to 10 and their respective replies were randomly selected; in the first attempt, 500 threads and 2750 respective replies across 15 sub-categories were selected. Secondly, to analyze the differences between sub-categories, 25 threads with 1 to 10 replies were randomly selected from top 6 sub-categories in terms of number of messages, in order to investigate whether interaction patterns between sub-categories are significantly different. 150 threads and their 648 respective replies were selected. Finally, to look at threads with more than 10 replies, 43 threads were selected

randomly among 15 sub-categories as well as their 932 respective replies. These three sample sets are referred to as sample set 1, 2 and 3 in the remainder of this research. With three sample sets combined, **693 threads** (13.1% in terms of total threads count) and **4330 of their respective replies** (12.2% in terms of total replies count) were made available for content analysis. An overview of the samples is displayed in Table 3.6.

Sample / sub-category	Number of replies	Number of threads	Number of replies
Set 1	1	50	50
	2	50	100
	3	50	150
	4	50	200
	5	50	250
	6	50	300
	7	50	350
	8	50	400
	9	50	450
	10	50	500
Set 2 - 101	1-10	25	127
- 301	1-10	25	111
- 303	1-10	25	107
- 306	1-10	25	120
- 601	1-10	25	94
- 605	1-10	25	89
Set 3	More than 10	43	932
Grand total		693	4330
Grand total percentagewise		13.1%	12.2%

Table 3.6: Samples for content analysis

Code

Three sets of codes are defined before and during data analysis, on the basis of edge type, thread starter indicator and network creation method. Edge type is associated with each post, as it provides an indicator of how a user interacts (e.g. replies to user X or replies to all above etc.) with other users in the same thread. Thread starter indicator is used to indicate whether a reply in a thread is posted by the thread starter. Additionally, from a technical perspective, it is easier to keep track of who the thread initiator is (i.e. the person who posts the first message), which will offer additional insights regarding the interactions between the thread initiator and other participants within the same thread . The details of coding are shown in Table 3.7 and Table 3.8.

Code	Edge Type	Code Source
1	Create edges from this node (actor) to all nodes within a thread. This type of edge can be seen as a broadcasting message, which only applies to a thread starter.	Literature
2	Create edges from this node to all nodes above (except itself) chronologically. This type of edge can be interpreted as "reply to all" in a conversational context.	Literature
3	Create an edge from this node to its immediate node above.	Literature
4	Create an edge from this node to the first node within a thread.	Observation
5	Others. This includes all other type of edges that cannot be identified by using the codes illustrated above.	Observation

Table 3.7: Details regarding edge type

Code	TS_ind (thread starter indicator)	Code Source
12	Indicates that the thread starting node reappeared again and replied to all nodes above chronologically, within a thread.	Observation
13	Indicates that the thread starting node reappeared again and replied to its immediate neighboring node above chronologically, within a thread.	Observation
15	Indicates that the thread starting node reappeared again and replied to one or more nodes above, but it does not follow the reply pattern of the previously defined thread starter indicator code.	Observation

Table 3.8: Details regarding thread starter indicator

Content analysis on sample set 1

During content analysis, edge type 1 was initially assigned to all thread starters (i.e. the first post of a thread), which can be interpreted as a broadcast of the thread. In rare occasions, the first reply of a thread (i.e. second post) have also been posted by the thread starter as an addition to the thread itself (e.g. more contents, or other matters). In such cases, edge type 1 have been assigned to this type of replies as well. Additionally, in the context of our data set, only thread starters are able to create broadcasting edges towards other participants. However, due to the fact that this type of edge presents a broadcasting nature which contradicts the remainder replies within a thread (Howison et al., 2010), edge type 1 has been excluded from this frequency distribution analysis.

Based on the results of content analysis, it is clear that a large majority of replies (> 78.6% on average) are intended towards the thread starter. This kind of interaction pattern among participants represents a Q&A style oriented online community (Adamic et al., 2008). The percentage of this interaction pattern for threads with between 1 to 10 replies varied between the minimal percentage of 71.9% (with 9 replies) and maximal percentage 100% (with 1 reply).

Another observation indicates that as a thread goes longer, the more diversity appears from a structural perspective. This is shown as [the number of] different edge types (edge type 3 and 5) beginning to rise. However, this does not change the fact that most participants within a thread

would reply directly to the thread starter. The details of this frequency distribution analysis are shown in Figure 3.6.

totalreplies			Frequency	Percent	Valid Percent	Cumulative Percent
1	Valid	4	50	100.0	100.0	100.0
2	Valid	2	22	22.0	22.0	22.0
		4	78	78.0	78.0	100.0
		Total	100	100.0	100.0	
3	Valid	2	28	18.7	18.7	18.7
		3	2	1.3	1.3	20.0
		4	119	79.3	79.3	99.3
		5	1	.7	.7	100.0
		Total	150	100.0	100.0	
4	Valid	2	39	19.5	19.5	19.5
		3	8	4.0	4.0	23.5
		4	149	74.5	74.5	98.0
		5	4	2.0	2.0	100.0
		Total	200	100.0	100.0	
5	Valid	2	41	16.4	16.4	16.4
		3	12	4.8	4.8	21.2
		4	191	76.4	76.4	97.6
		5	6	2.4	2.4	100.0
		Total	250	100.0	100.0	
6	Valid	2	41	13.7	13.7	13.7
		3	25	8.3	8.3	22.0
		4	225	75.0	75.0	97.0
		5	9	3.0	3.0	100.0
		Total	300	100.0	100.0	
7	Valid	2	43	12.3	12.3	12.3
		3	28	8.0	8.0	20.3
		4	263	75.1	75.1	95.4
		5	16	4.6	4.6	100.0
		Total	350	100.0	100.0	
8	Valid	2	38	9.5	9.5	9.5
		3	35	8.8	8.8	18.3
		4	308	77.0	77.0	95.3
		5	19	4.8	4.8	100.0
		Total	400	100.0	100.0	
9	Valid	2	48	10.7	10.7	10.7
		3	64	14.3	14.3	24.9
		4	323	71.9	71.9	96.9
		5	14	3.1	3.1	100.0
		Total	449	100.0	100.0	
10	Valid	2	50	10.0	10.0	10.0
		3	33	6.6	6.6	16.6
		4	394	79.0	79.0	95.6
		5	22	4.4	4.4	100.0
		Total	499	100.0	100.0	

Figure 3.6: SPSS output on frequency distribution of edge types per number of replies

Content analysis on sample set 2

Based on the observation, a large majority of replies (73.8% on average) are intended towards the thread starter. This is slightly lower than the result of the previous sample set. The percentage of this interaction pattern [reply to thread starter] between sub-categories varied between the minimal percentage of 70.3% (sub-category 601) and maximal percentage of 78.3% (sub-category 101). This minor reduction (4.8%) in percentage is obviously caused by the rising of other edge types, as they constitute a larger portion of the edge type distribution as a result by excluding number of replies as a factor. The details are illustrated in Figure 3.7.

cate_id			Frequency	Percent	Valid Percent	Cumulative Percent
101	Valid	2	20	13.2	13.2	13.2
		3	12	7.9	7.9	21.1
		4	119	78.3	78.3	99.3
		5	1	.7	.7	100.0
		Total	152	100.0	100.0	
301	Valid	2	15	11.1	11.1	11.1
		3	12	8.9	8.9	20.0
		4	101	74.8	74.8	94.8
		5	7	5.2	5.2	100.0
		Total	135	100.0	100.0	
303	Valid	2	16	12.2	12.2	12.2
		3	14	10.7	10.7	22.9
		4	97	74.0	74.0	96.9
		5	4	3.1	3.1	100.0
		Total	131	100.0	100.0	
306	Valid	2	14	9.7	9.7	9.7
		3	15	10.4	10.4	20.1
		4	105	72.9	72.9	93.1
		5	10	6.9	6.9	100.0
		Total	144	100.0	100.0	
601	Valid	2	17	14.4	14.4	14.4
		3	14	11.9	11.9	26.3
		4	83	70.3	70.3	96.6
		5	4	3.4	3.4	100.0
		Total	118	100.0	100.0	
605	Valid	2	15	13.3	13.3	13.3
		3	8	7.1	7.1	20.4
		4	82	72.6	72.6	92.9
		5	8	7.1	7.1	100.0
		Total	113	100.0	100.0	

Figure 3.7: SPSS output on frequency distribution of edge types per sub-category

Content analysis on sample set 3

Lastly, we investigate the frequency distribution of edge types on the basis of our third sample set, namely threads with more than 10 replies. As reported previously, in this set of data there are 932 replies to 43 threads. However, one of the threads contained a first reply by the thread starter himself, which is considered a broadcasting message as well (since reply to oneself does not make much sense), and is therefore removed from the analysis.

Unlike the findings of the two sample sets before, the percentage of interaction patterns between thread starter and other thread participants dropped below 60% (58.1%), which is rather significant compared to previous results. This indicates that, as the number of replies increases, the interactions among participants will slowly lose the form of Q&A oriented style, as the number of replies addressed to thread starters decreases rather significantly. The exact edge type distribution is displayed in Figure 3.8.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2	173	18.6	18.6	18.6
	3	132	14.2	14.2	32.8
	4	541	58.1	58.1	90.9
	5	85	9.1	9.1	100.0
	Total	931	100.0	100.0	

Figure 3.8: SPSS output on frequency distribution of edge types of threads with more replies

Involvement of the thread starters

In this section, the emphasis is put towards the first sample set, i.e. 500 threads and 2750 respective replies, as this set is well-quantified and represents over 80% of the thread length in terms of thread count of the selected sub-categories. Next, we look at the involvement of thread starters respectively. We calculate the percentage with which thread starters are involved in later conversations in their own thread as well as the percentage they hold in different edge types.

In short, this includes investigation of edge types that do not belong to edge type 1 (towards to all & broadcasting) and edge type 4 (towards the thread starter), on the basis of number of replies and the respective involvement of the thread starters. Threads with just 1 reply have been excluded from this table, due to the fact that the thread starters did not reappear in their own threads.

In this case, we see a rather steady decline percentage wise, in terms of the involvement of thread starters with a few exceptions (3, 5 and 7). This implies that the structure of the interactions does change as the threads become longer. The percentage of thread starters' involvement dropped from 95.5% with 2 replies down to 66.7% with 10 replies. Details are shown in Table 3.9.

Total replies per thread	Total count of edge type 2, 3 and 5	Total count of edge type 2, 3 and 5 from thread starters	Percentage of thread starters' involvement
2	22	21	95.5%
3	31	30	96.8%
4	51	38	74.5%
5	59	50	84.7%
6	75	55	73.3%
7	87	70	80.5%
8	92	67	72.8%
9	126	86	68.3%
10	105	70	66.7%

Table 3.9: Summary of thread starters' involvement

Additionally, details regarding how the thread starters are involved in each specific edge type and their distributions have been calculated; a summary is presented in Table 3.10. Figure 3.10 visualized the comparison between total number of replies and thread starters' involvement, in term of edge types.

Surprisingly, the involvement rate of thread starters is not evenly distributed among these three edge types (Figure 3.9). The results indicate that thread starters have a significantly higher participation rate in edge type 2 (average 94.2%) than edge type 3 (average 45.6%) and edge type 5 (average 53.1%). This means that, if thread starters reappear in their own threads, it is most likely that they are going to reply to all other participants who had replied earlier.

A similar statement cannot be made with regards to the other two edge types. Furthermore, the relative proportion (54%) of edge type 2 is also much larger than edge type 3 (14%) and 5 (32%).

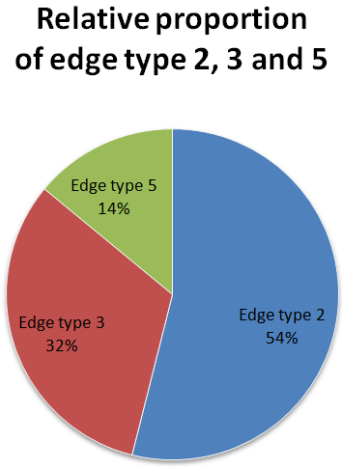


Figure 3.9: Relative proportion of edge type 2, 3 and 5

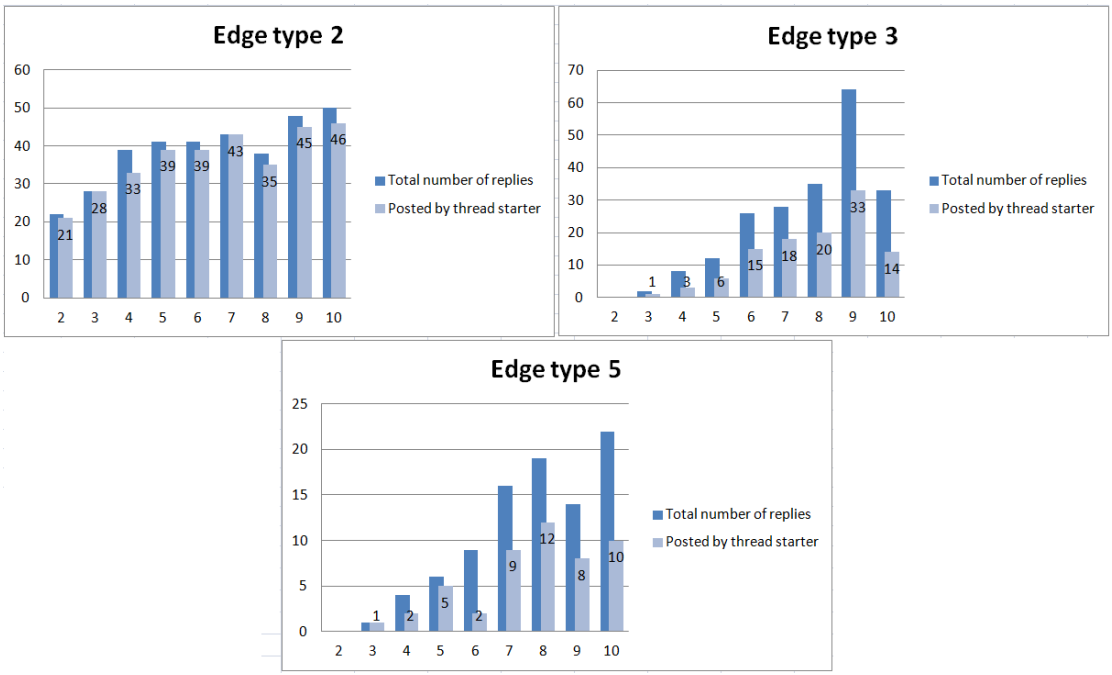


Figure 3.10: Involvement of thread starters per edge type

Total replies per thread	Edge type	Total edge count	Total edge count of thread starters	Percentage of thread starters' involvement
2	**2**	22	21	95.5%
3	**2**	28	28	100%
	3	2	1	50%
	5	1	1	100%
4	**2**	39	33	84.6%
	3	8	3	37.5%
	5	4	2	50%
5	**2**	41	39	95.1%
	3	12	6	50%
	5	6	5	83.3%
6	**2**	41	39	95.1%
	3	26	15	57.7%
	5	9	2	22.2%
7	**2**	43	43	100%
	3	28	18	64.3%
	5	16	9	56.3%
8	**2**	38	35	92.1%
	3	35	20	57.1%
	5	19	12	63.2%
9	**2**	48	45	93.8%
	3	64	33	51.6%
	5	14	8	57.1%
10	**2**	50	46	92%
	3	33	14	42.4%
	5	22	10	45.5%

Table 3.10 Edge type distribution per number of replies and thread starters' involvement

3.3.2 Alternative network creation method

Based on the results and observations of content analysis elaborated in previous sections, it is clear that, in the context of this research, those pseudo network creation methods derived from literature review are insufficient in terms of goodness of fitting or pose significant drawbacks, i.e. are unable to predict the interactions among the users of Hallo! community. With this in mind, we propose an alternative pseudo network creation method based on the results and observations discussed in previous sections.

The results of content analysis suggest the vast majority of replies within the same thread, are addressed toward thread starters. And when thread starters reappear in their own thread, (s)he

is most likely going to post a reply addressing every other participant who previously has replied within the discussion thread. Based on these two factors, we propose an alternative pseudo network creation method with the following rules:

1) The initial post (the thread itself) does not build any edge towards others due to its broadcasting nature,
2) A reply posted by a participant other than the thread starter himself will result in an edge from this participant to the thread starter,
3) If a thread starter replies in its own thread, and this is not the first reply of the thread, then this reply will result in edges being built from this [thread starter] replies' re-entering position to all thread participants who had replied previously.

A visualized version of this proposed method is displayed in Figure 3.11, and is named Method 4 in the context of this research.

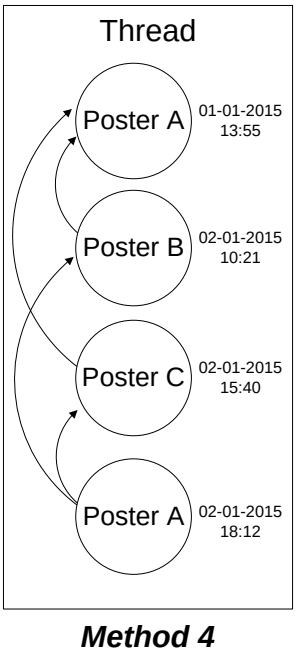


Figure 3.11: Pseudo network creation method 4

3.3.3 Evaluation of pseudo network creation methods

In order to evaluate the pseudo network creation methods retrieved from literature and from our own proposal, we plan to approach this issue by using a three-step process: 1) to test the goodness of fitting, 2) to study the network metrics and its visualizations and 3) by using quadratic assignment procedure.

Pseudo network creation methods were implemented in the Python programming language, based on the definitions of pseudo network creation methods (Table 3.11), and additionally

verified on a network data subset by means of the open source Gephi network analysis package. By executing the network creation scripts on the data set, network structures were created and ready to be analyzed.

Code	Pseudo network creation methods	Source
1	All nodes are interconnected (Berger, Klier, & Richter, 2014)	Literature
2	A node builds edges to all nodes above it in an chronological order (Toral, Mart ńez-Torres, & Barrero, 2010)	Literature
3	A node builds an edge to its immediate neighboring node above it (Faraj & Johnson, 2011).	Literature
4	If a node is not the thread starting node, then this node builds an edge to the thread starting node. If a node is the thread starting node that reappeared in the same thread, then this node builds edges to all other nodes above it in an chronological order.	Observation
5	Others. Which means that one or more replies within this thread do not 100% match any of the network creation methods elaborated above.	Observation

Table 3.11: Definitions of pseudo network creation methods

Goodness of fitting

First, we evaluate the network creation methods elaborated previously, by reporting their respective fittings on threads and replies from the content analysis on the first sample set. A method code is only assigned on the basis of threads, which indicates all replies in the same thread must qualify to apply one of the existing methods: if one reply does not fulfill the conditions, then the whole thread would have a code 5 assigned to it. Code 5 in this sense could be interpreted as the baseline of reality.

This evaluation is solely based on the network creation code assigned to a thread in comparison to the network creation methods illustrated previously. The results were assembled in 10 separate worksheets in Excel, and were then examined on basis of the code. Each thread regardless of its number of replies will only have one network creation method code assigned to it. The outcomes of this process resulted in 10 Excel worksheets and each containing 50 randomly selected threads of sample set 1 from threads with 1 reply to 10 replies (500 threads in total). The results provide an overview regarding how many percent of the user interaction patterns can be detected by one of the network creation methods. The actual network creation method codes as well as their relative percentage in one of the categories (i.e. number of replies) are displayed in Figure 3.12.

Due to oversimplified network structures from threads with 1 or 2 replies, almost all methods (code 1, 2, 3 and 4) were applicable. As the number of replies increases (e.g. starting at 3 replies), the overlapping cases occurred earlier, have consequently vanished. This indicates that the prediction rate (e.g. 50% at 7 replies; Figure 3.12) of pseudo network creation methods decrease as the number of replies in threads increase. Furthermore, the results shown below imply that the network creation methods are unable to provide perfect detections as the

threads grow longer. For instance, only 42% of the threads with 10 replies can be detected by method 4. Higher percentage of occurrences in code 5 illustrates that in reality, the user interactions are much more complicated. However, this overview does not provide any insights on the approximate between pseudo networks and baseline network, therefore it is essential to conduct addition analysis for investigating the networks characteristics in-depth.

totalreplies			Frequency	Percent	Valid Percent	Cumulative Percent
1	Valid	1234	50	100.0	100.0	100.0
2	Valid	4	27	54.0	54.0	54.0
		1234	23	46.0	46.0	100.0
		Total	50	100.0	100.0	
3	Valid	4	48	96.0	96.0	96.0
		5	2	4.0	4.0	100.0
		Total	50	100.0	100.0	
4	Valid	2	1	2.0	2.0	2.0
		4	39	78.0	78.0	80.0
		5	10	20.0	20.0	100.0
		Total	50	100.0	100.0	
5	Valid	4	37	74.0	74.0	74.0
		5	13	26.0	26.0	100.0
		Total	50	100.0	100.0	
6	Valid	4	30	60.0	60.0	60.0
		5	20	40.0	40.0	100.0
		Total	50	100.0	100.0	
7	Valid	4	25	50.0	50.0	50.0
		5	25	50.0	50.0	100.0
		Total	50	100.0	100.0	
8	Valid	4	18	36.0	36.0	36.0
		5	32	64.0	64.0	100.0
		Total	50	100.0	100.0	
9	Valid	4	19	38.0	38.0	38.0
		5	31	62.0	62.0	100.0
		Total	50	100.0	100.0	
10	Valid	4	21	42.0	42.0	42.0
		5	29	58.0	58.0	100.0
		Total	50	100.0	100.0	

Figure 3.12: Goodness of fitting analysis on pseudo network creation methods

Network metrics and visualization

The previous analysis only provides a preliminary overview regarding the prediction rate of existing methods. In order to gain additional insights, more sophisticated network evaluation techniques have to be applied. Simulations of real network structures are to be created manually on the basis of results and observations of content analysis conducted in previous sections (i.e. sample set 1, 2 and 3). These manually created networks can then be used as baselines for comparison between pseudo network creation methods and the reality.

A list of basic descriptive network metrics has been selected and calculated on networks created from all sample sets. One network is visualized as a showcase by using Gephi software. Last but not least, quadratic assignment procedures (QAP) are to be performed on various sample sets in UCINET software, to calculate the correlations between baseline network and pseudo networks respectively.

As illustrated by Hanneman et al. (2005), QAP can be used to test significance of associations between different networks constructed by the same set of actors [nodes]. The example elaborated in Hanneman et al.'s case was based on the information exchange network and money exchange network of the same set of actors. The correlation of these two networks could have multiple implications, e.g. positive correlation may indicate that information

exchange would lead to monetary exchange; negative correlation could imply a complementary relation, i.e. "*money flows in one direction and information flows in the other*" (Hanneman et al., 2005); or the two sets of networks did not have anything to do with each other, i.e. no correlation.

QAP in the context of this research, can be extended to calculate the correlations between two or more networks constructed on basis of the same set of actors, not in the sense of their relations (e.g. information & money) but rather from the perspective of structural creation (e.g. baseline network vs. pseudo networks).

Sample set 1

The first set of networks are implemented on the sample set 1 from content analysis. By evaluating the descriptive network metrics, method 4 seems to be the best fitting in comparison with the baseline network.

As highlighted in Table 3.12, the network diameter and network density between the two networks are exactly the same, whereas other metrics are also within close approximation. Networks created from method 1 and 2 are somewhat off the mark, especially in terms of edge counts, average degree and average weighted degree.

	Manually	Method 1	Method 2	Method 3	Method 4
Number of nodes	1046	1046	1046	1046	**1046**
Number of edges	2852	11112	6772	2445	**2863**
Average degree	2.727	10.623	6.474	2.337	**2.737**
Avg. weighted degree	3.221	18.889	9.439	2.55	**3.648**
Network diameter	10	6	9	18	**10**
Graph density	0.003	0.01	0.006	0.002	**0.003**

Table 3.12: Descriptive network metrics (sample set 1)

Next, the visualizations of 5 networks created on sample set 1 are presented in Figure 3.13. The visualization algorithm applied (in Gephi) was "Forced Atlas 2" (Jacomy, Venturini, Heymann, & Bastian, 2014). This algorithm provides a number of options that enable the user to choose how distances between nodes are presented by manipulating key parameters. The algorithm takes degree as well as weighted degree into account regarding node size and edge weight for visualization purposes.

In Figure 3.13, denser paths between nodes indicate that the edges carry higher weights in contrast to paths with a smaller density value. The positions of nodes among all 5 visualized graphs are set on fixed positions for illustration purpose, as each node on the given position in one graph, would represent the same node in other graphs. This is done by applying the following steps:

1) Import all network structures (edge lists) that need to be put to contrast into Gephi software with directional setting;

2) Export all edge details generated by Gephi software among imported network structures;
3) Choose one of the networks as basis and execute "Forced Atlas 2" algorithm with desired settings (regarding nodes and edges);
4) Wait until the graph is generated, and stabilized (arbitrary), then stop the algorithm;
5) Take a snapshot of the generated network graph;
6) Remove all records in edge tab, and import one of the exported edge details back into Gephi, and refresh the graph review button;
7) Repeat step 1-6 till snapshots have been taken on all desired network graphs.

The visualizations of those network graphs indicate similar results as discussed in previous section. Method 4 possess a higher likelihood in comparison to the baseline network than any other network graphs, especially in contrast to method 1 and 2; whereas method 1 and 2 are way overflowed in terms of edge counts.

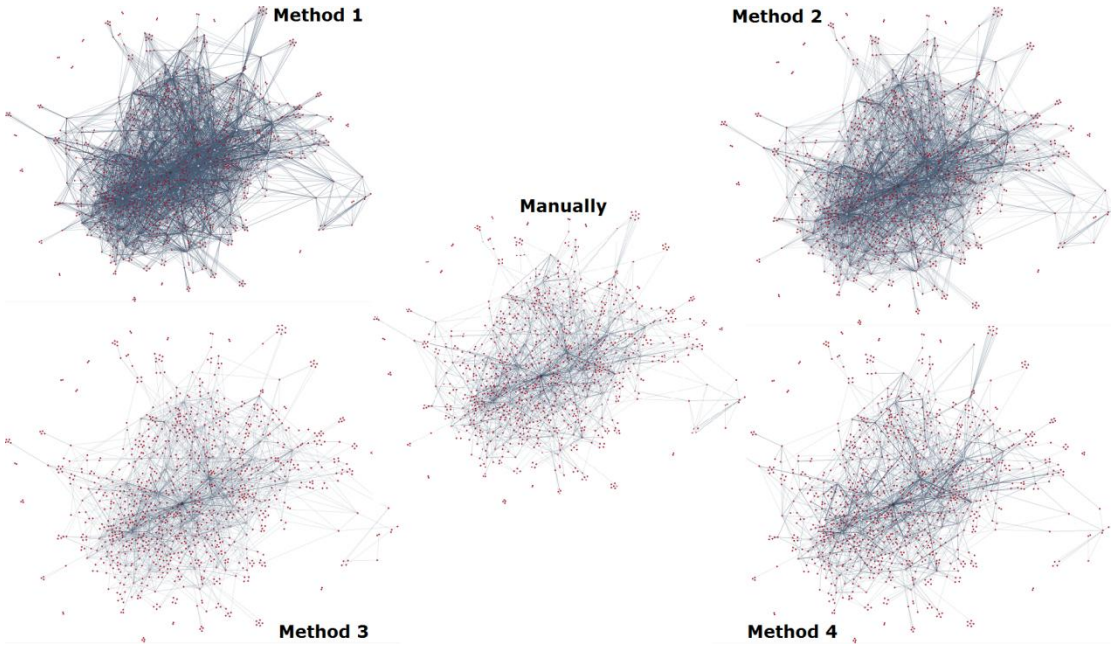


Figure 3.13: Visualization of various networks from content analysis sample set 1

To examine the approximate between pseudo networks and baseline network, correlation analysis between these networks have been conducted by applying QAP . In Table 3.13 below, the first column presents which of the two networks are put into correlation analysis. The second column indicates the Pearson's correlation coefficient of two networks, values here are measured between 0 and 1, where 0 indicates no correlation and 1 indicates a perfect correlation, it is the most important measure for comparing two sets of networks with the same actors. The following columns depict their corresponding significance, average and standard deviations.

The results of QAP on networks created from sample set 1 show that the baseline network and method 4 have the highest correlation coefficient, i.e. 0.8879, in comparison to baseline vs. method 1, 2, 3 (0.691, 0.6968 and 0.5416 respectively). This indicates that the proposed

method is more suitable for network creation, in contrast to existing pseudo network creation methods, within the context of threads with 1-10 replies.

The pseudo networks have also been compared as shown in row number 6 – 11. Method 1 and 2 have the highest correlation coefficient (0.8656) among any other combinations of pseudo networks.

Method	Observed value	Significance	Average	Standard deviation
Manually, method 1	0.6910	0.0002	-0.0000	0.0015
Manually, method 2	0.6968	0.0002	0.0000	0.0014
Manually, method 3	0.5416	0.0002	0.0000	0.0012
Manually, method 4	**0.8879**	**0.0002**	**0.0000**	**0.0012**
Method 1, 2	0.8656	0.0002	0.0000	0.0020
Method 1, 3	0.6410	0.0002	-0.0000	0.0015
Method 1, 4	0.6954	0.0002	-0.0000	0.0015
Method 2, 3	0.6778	0.0002	-0.0000	0.0014
Method 2, 4	0.7074	0.0002	-0.0000	0.0013
Method 3, 4	0.5096	0.0002	0.0000	0.0012

Table 3.13: QAP correlations of sample set 1

Sample set 2

The sample set 2 is then analyzed based on descriptive network metrics. As discussed previously, this sample set contains 25 randomly selected threads (with 1 to 10 replies) and their respective replies from top 6 sub-categories in terms of total number of messages. By analyzing the data extracted from various sub-categories therefore, the diversity of interactions can be revealed across sub-categories. Despite the similarities of the user interaction patterns across 6 selected sub-categories, it is still interesting to compare the pseudo networks and the baseline network created from the data of sample 2, in terms of network metrics as well as correlation coefficients.

As for descriptive network metrics, method 4 remained the most suitable pseudo network across 6 selected sub-categories in contrast to pseudo networks created based on existing literature, in terms of comparative measures, with a very few exceptions (e.g. highlighted in red, Table 3.14) regarding the metric of network diameters. Details of the results of descriptive network metrics of sample set 2 is attached in Appendix B.

Sub category 101, Start-ups					
	Manually	**Method 1**	**Method 2**	**Method 3**	**Method 4**
Number of nodes	105	105	105	105	**105**
Number of edges	170	692	410	138	**162**
Average degree	1.619	6.59	3.905	1.314	**1.543**
Avg. weighted degree	1.943	11.029	5.505	1.4	**1.924**
Network diameter	10	7	11	20	**8**
Graph density	0.016	0.063	0.038	0.013	**0.015**

Table 3.14: Descriptive network metrics (sample set 2)

In the same setting, method 4 upholds the highest QAP correlation coefficient between 0.8488 on sub-category 101 and 0.9236 on sub-category 303, among 6 selected sub-categories. This implies that pseudo networks created from method 4 are an overall better choice than other pseudo network creation methods from literature in this setting. Such a result provides inferential evidence, which suggests that method 4 could be further used in network creation for other sub-categories. An example of QAP results on sub-category 101 is displayed in Table 3.15, and the full details can be found in Appendix C.

Sub category 101, Start-ups				
Method	**Observed value**	**Significance**	**Average**	**Standard deviation**
Manual, method 1	0.7265	0.0002	-0.0002	0.0135
Manual, method 2	0.6580	0.0002	-0.0001	0.0134
Manual, method 3	0.4883	0.0002	-0.0004	0.0133
Manual, method 4	**0.8488**	**0.0002**	**0.0001**	**0.0140**

Table 3.15: QAP correlations of sample set 2

Sample set 3

As discussed previously, an additional QAP is performed on the sample set 3, in order to obtain an overview of the whole community, in terms of reply counts. The results of content analysis on sample set 3 indicate that as number of replies on a thread increases, other interaction patterns would emerge (e.g. the number of occurrences on edge types). This poses a challenge to the proposed method, as method 4 builds the network on basis of interactions between the thread starter and other participants within a thread.

As expected, the descriptive metrics on pseudo networks no longer provide a clear choice as they did in the two sections above. The comparative margins between the baseline network and pseudo networks are significantly larger than the ones discussed previously. The results are illustrated in Table 3.17.

	Manually	Method 1	Method 2	Method 3	Method 4
Number of nodes	455	455	455	455	455
Number of edges	2709	12832	7109	822	**900**
Average degree	5.954	28.202	15.624	1.807	**1.978**
Avg. weighted degree	8.899	66.888	33.433	1.996	**3.943**
Network diameter	9	4	6	23	**9**
Graph density	0.013	0.062	0.034	0.004	**0.004**

Table 3.16: Descriptive network metrics (sample set 3)

The last option is to conduct QAP analysis, by putting the baseline network in comparison to pseudo networks in terms of their correlation coefficient. In this particular setting, method 2 obtained a higher correlation (0.6645) with the baseline network; method 1 on the other hand, has a slightly lower correlation (0.6494). Both methods are significantly better in terms of correlation with the baseline than method 3 and 4. The QAP results are shown in Table 3.17

Method	Observed value	Significance	Average	Standard deviation
Manually, method 1	0.6494	0.0002	0.0001	0.0060
Manually, method 2	**0.6645**	**0.0002**	**0.0001**	**0.0056**
Manually, method 3	0.4200	0.0002	0.0001	0.0045
Manually, method 4	0.4910	0.0002	-0.0000	0.0033

Table 3.17: QAP correlations of sample set 3

3.3.4 Summary

As elaborated at the beginning of this section, unstructured data retrieved from online communities, especially from online forums, do not always provide explicit relational references between users in terms of their interaction patterns. Pseudo network creation methods were applied to simulate the networks based on the unstructured data for SNA. Furthermore, existing literature indicates that networks built upon unstructured data are often not verified in the sense of their accuracy to reality. With this in mind, an extensive content analysis was carried out on three randomly selected sample sets containing various essential features of Hallo! community, in terms of interest variety [sub-categories] and topic popularity [number of replies]. The sample sets are well quantified, and represent a decent margin of the population [13.1% thread wise / 12.6% reply wise].

The result of this content analysis is three-fold:

1) short threads (up until 10 replies) have a rather clear and consistent interaction pattern between thread starter and other participants within the same thread;

2) interaction patterns between different sub-categories on the basis of short threads (up till 10 replies) are quite similar; and

3) this pattern decays in longer threads (more than 10 replies), as the thread starters would become relatively less involved in their own threads, combined with a rise in other types of unpredictable interactions between participants.

As a result, a baseline network which represents the reality was created manually based on content analysis, so that pseudo networks can be evaluated properly. Additionally, an alternative pseudo network creation method was proposed based on these results and observations. This method has then been evaluated amongst other pseudo network creation methods derived from literature review, against the baseline network mentioned earlier.

The results indicate that the proposed method is able to create a pseudo network that reflects the reality across all sub-categories that have been selected for this research. However, this method contains a severe limitation with regards to the fitting to threads with more than 10 replies.

Therefore, a decision has been made to apply two pseudo network creation methods based on the choice of the lesser of two evils. For threads with up to 10 replies, the proposed method is used; for threads with more than 10 replies, pseudo method 1 is used. This way, pseudo networks created from this mixed method will supply more reliable network structures for SNA in the next chapter.

3.4 Adaptation of longitudinal analysis in online communities

As it has been introduced in sub section 3.1.1, Hallo! community upholds a hierarchical structure, the fundamental unit of analysis that constructs the network structure is a post [or message]. Posts are grouped by threads, with the initial post (earliest post in a thread) considered the thread starter. Threads are then grouped by sub categories in terms of interests and ultimately situated under one of the main categories of interest.

As posts are grouped by threads, and within each thread, there is a chronological sequence for each post (timestamp); the goal of this section is to investigate proportional attributes as well as their distributions. In this way, when conducting a longitudinal analysis regarding online communities, continuous threads could maintain their status quo, preventing the threads from breaking into multiple fragments (or at least to do so to a certain extent).

The time centric approach is seemingly the most suitable choice in the context of this research, and is utilized based on the objective illustrated previously. The motive for this choice is two-fold:

1) the actor centric approach elaborated by Snijders (1996) is well suited to analyze actors' evolvement, however, it lacks the ability of monitoring the longitudinal development of open online communities such as *Hallo!*, and

2) the event centric approach focuses on external events and their impacts on the community as the driving factor of this approach, yet indications regarding external events are barely available in the context of this research (the only known external event in this case, is the official launch date of Hallo! community).

As mentioned earlier, each post has its own timestamp, which indicates its position within the thread. The duration of a thread can then be calculated on basis of the posts' timestamps. In this context, the duration of a thread is defined as following:

thread duration = last post date of a thread – initial post date of a thread

For illustration purposes, the precise time elements (e.g. hours, minutes & seconds) of the timestamps have been left out of the calculation. Threads that ended on the same calendar day are considered as threads with 0 day of duration.

3.4.1 Duration analysis

As introduced previously, in section 3.1.1 and 3.4, the foundation of Hallo! community is based on a hierarchical online forum structure, which implies that the basis for constructing a social network structure relies on the group attributes such as sub-category and thread. Thread as the lowest level of grouping attribute, is constructed from cumulative data over time (i.e. replies). Therefore it is essential to analyze the duration of threads in order to select appropriate time intervals for data segmentation to conduct a longitudinal SNA. If the duration of threads is overlooked, it may result in fragmented network snapshots, as the replies are separated from its grouping attribute by different time based data segmentation boundaries, i.e. time intervals.

The first idea is to look at the percentage of unbroken threads in terms of periodic intervals, which could be used as time intervals for data segmentation, in order to create network snapshots as discussed in the literature review. We utilize a fixed length approach, i.e. each time interval has the same length in terms of calendar days. Details regarding thread / reply distribution over the entire duration of 15 sub-categories were presented previously in Table 3.5. An unbroken thread is defined as following:

> *A thread is considered unbroken, if the post date of this thread and the post date of the last reply in this thread are completely covered by the pre-defined periodic time interval.*

Table 3.18 shows the percentage of unbroken threads regarding each of the individual sub category on basis of various time intervals. On the bottom of this table, average values across sub-categories were calculated and displayed. The results in general, indicate that the vast majority of threads can be covered by periodic time intervals of 60 and 90 days (78% and 81% coverage). Additionally, a visualized example is demonstrated in Figure 3.14 on sub-category 302. In this figure, Y-axis illustrates the thread identifiers (only shown partially, due to the length of the figure) in an ascending order, whereas X-axis indicates the duration of the final data set, i.e. 01-03-2009 to 25-01-2011 (is displayed till 19-02 for the last interval).

The vertical lines in between is based on 90 days time interval. Each of the horizontal line in blue represents an unique thread, and its length is determined by the thread duration defined earlier. However, the exact distribution of replies cannot be revealed this way, e.g. the number or percentage of replies to threads that have been intercepted by the different time frame borders.

Sub category id	% of unbroken threads per 7 days	% of unbroken threads per 14 days	% of unbroken threads per 30 days	% of unbroken threads per 60 days	% of unbroken threads per 90 days	% of unbroken threads per 120 days
101	65.1%	73%	81%	85.4%	86.5%	88.7%
104	49.3%.	55.8%	63.3%	68.4%	73.5%	76.7%
301	49.5%	57.2%	66.5%	73.3%	77.4%	81.4%
302	54.5%	62%	65.2%	74.2%	78.1%	78.9%
303	49.3%	57.8%	65.7%	73.7%	76.2%	80.7%
304	54.3%	61.7%	68.6%	70.7%	77.1%	77.1%
305	60.6%	67.4%	69.7%	77.3%	78%	82.6%
306	55.3%	62.3%	71.3%	77%	78.7%	80.6%
404	68.6%	73.6%	83.6%	87.1%	87.1%	90.7%
501	54.5%	57.7%	68.5%	74.3%	77.6%	83.8%
502	63.6%	72%	74.8%	82.2%	87.9%	85%
503	64%	72.6%	82.3%	86.6%	91.4%	90.9%
601	62%	72%	79.1%	86.4%	90.2%	90.9%
604	54.1%	58.8%	67.8%	72.8%	76.2%	78.4%
605	58.9%	66.3%	71.6%	78%	81.7%	86.5%
Average	**57%**	**64.3%**	**72.1%**	**78%**	**81%**	**83.7%**

Table 3.18: Percentage of unbroken threads per sub category on basis of various time intervals

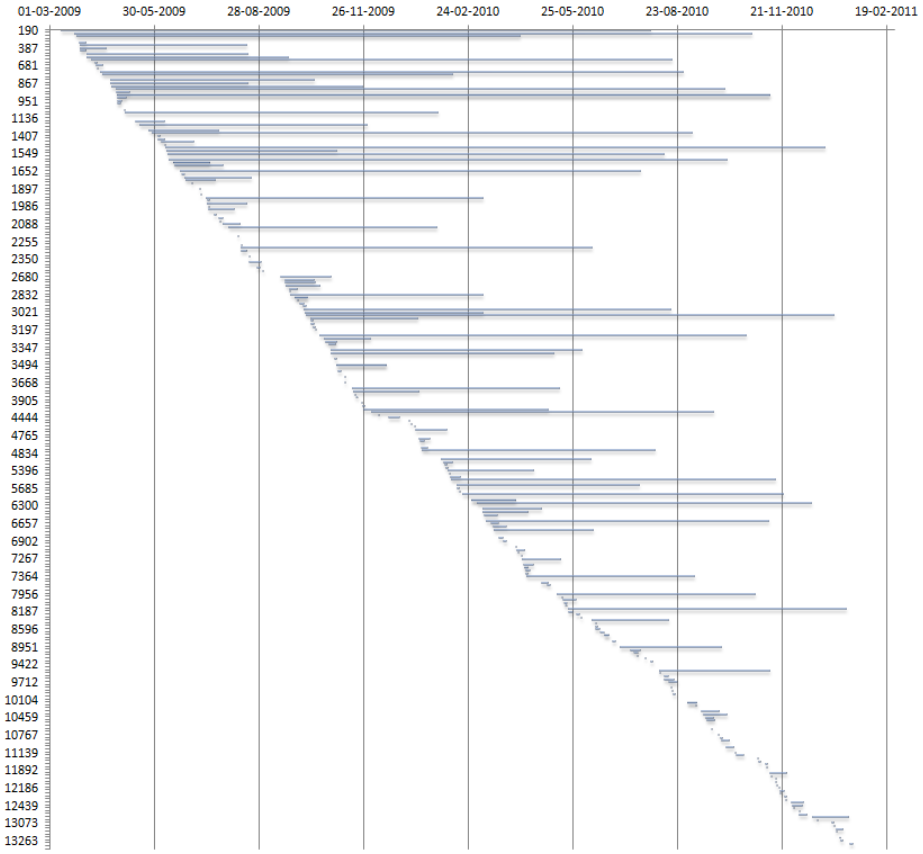


Figure 3.14: Visualization of thread duration, sub-category 302

In this section, the focus is towards a different perspective, namely, we look at the average duration of threads, on the basis of number of replies. As shown in Table 3.19, without considering the factor posed by sub-categories, the general trend is as one would normally expect: threads with more replies are more likely to have a longer average duration.

Threads with 1 to 10 replies have average duration between the minimal value 14.7 days (on 2 replies) and the maximal value 51.5 days (on 8 replies); threads with more than 10 replies have an average duration of 85.5 days. However, this trend is not 100% consistent, e.g. threads with 1 reply have a higher average duration (16.3 days) than threads with 2 replies (14.7 days); such inconsistencies have been highlighted in the table.

Number of replies	Average thread duration in days
1	**16.3**
2	14.7
3	26.3
4	27
5	**37.1**
6	33.4
7	37.5
8	**51.5**
9	38.9
10	48.2
more than 10	85.5

Table 3.19: Duration analysis on basis of number of replies

The factor posed by sub-categories has been included into the duration analysis. The average thread duration per number of replies is calculated within each sub-category. The full details are reported in Appendix D. We apply Pearson's correlation coefficient to analyze whether there is a correlation between the baseline average thread duration illustrated in Table 3.18 and each individual sub-category. The objective is to investigate whether duration of threads varies between sub-categories and the baseline duration illustrated in Table 3.19, in terms of correlation coefficient. As the results indicate in SPSS output (Table 3.120), 10 out of 15 sub-categories have positive correlations with the baseline average thread duration, which implies that one third of the sub-categories have a different pattern in terms of average thread durations per number of replies.

		baseline	g101	g104	g301	g302	g304	g305	g306
baseline	Pearson Correlation	1	.766**	.856**	.625*	.823**	.805**	.409	.869**
	Sig. (2-tailed)		.006	.001	.040	.002	.003	.212	.001
	N	11	11	11	11	11	11	11	11
		g404	g501	g502	g503	g601	g604	g605	
baseline	Pearson Correlation	.348	.862**	.678*	.348	.838**	.698*	.319	
	Sig. (2-tailed)	.295	.001	.022	.294	.001	.017	.339	
	N	11	11	11	11	11	11	11	

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Table 3.20: SPSS output on Pearson's correlation coefficient

In addition to that, the deviations between sub-categories and the baseline average thread duration have been put into contrast (Table 3.21; NoR refers to number of replies). The deviation is calculated on the values in each sub-category minus its corresponding cell in the baseline illustrated in Table 3.18 (except row "Avg"). The values calculated deviation per sub-category is then calculated in average per sub-category.

The results of Pearson's correlation coefficient and average deviation among different sub-categories seem to be contradictory, e.g. there is no correlation between baseline average thread duration and sub-category 605 ($r = 0.319$, $n = 11$, $p = .339$), but the average deviation of sub-category 605 indicates otherwise, i.e. -2.8 days. Such results indicate that the general trend illustrated in Table 3.18 (more replies lead to a longer average thread duration) does not occur in every sub-category, and therefore the results shown in this section are inconclusive.

NoR	101	104	301	302	303	304	305	306	404	501	502	503	601	604	605
1	-7.8	11.1	7.3	12.1	9.6	-8.9	31.9	-5.4	-1.6	13.1	-13.4	-12.4	-4.7	-4.5	-1.4
2	-7.7	1.8	9.0	2.6	0.0	22.1	-9.7	2.2	-13.6	-12.5	-13.5	1.1	-10.4	5.1	4.3
3	-15.7	25.7	2.6	1.1	-8.0	-24.2	-21.7	21.2	-18.8	4.1	-22.9	-21.8	-13.4	20.4	4.5
4	-13.4	-7.2	7.5	22.4	9.6	7.8	-3.9	-13.1	12.5	-17.7	-6.6	-24.4	-19.0	36.1	24.8
5	3.6	-19.0	5.2	44.6	-14.1	-8.2	-30.3	0.2	9.3	-2.4	30.0	-21.7	-29.7	17.7	-8.2
6	-9.1	48.2	-2.9	35.5	-14.2	48.8	-2.6	1.7	-19.6	-3.9	-26.0	47.1	-16.3	-3.8	-10.9
7	-10.7	13.0	23.3	39.7	-21.3	13.0	3.3	7.5	-20.2	-18.9	-28.2	-32.1	-28.1	-9.9	36.7
8	-34.3	79.5	-0.4	-13.0	3.2	44.4	-34.3	7.8	29.9	-1.6	-49.1	58.3	3.2	16.0	-46.7
9	-6.6	21.9	-29.5	14.3	-27.7	15.9	43.9	-1.1	-32.5	33.5	-26.1	-36.7	-10.4	81.6	11.3
10	-37.2	38.4	3.9	24.2	27.6	-1.2	217.8	-20.1	-46.4	-1.6	-43.2	-9.2	-9.1	165.6	-15.1
10+	-27.2	59.3	10.9	-5.6	11.9	31.2	11.5	0.8	-54.4	19.7	5.2	-55.2	-31.7	75.5	-29.7
Avg	-15.1	24.8	3.4	16.2	-2.1	12.8	18.7	0.15	-14.1	1.1	-17.6	-9.7	-15.4	36.4	-2.8

Table 3.21: Deviation between each sub-category and the baseline average thread duration

3.4.2 Post distribution analysis

In this section, the time between posts is analyzed among the selected sub-categories. Reply distribution in this sense is a relative term, e.g. it measures the relative distance (illustrated in days) between replies of a thread and the post date of the thread. The results cannot indicate absolute values, due to the fact that posting in online forums is a continuous process. However, to a certain extent, the results of post distribution analysis could provide evidence to support the choices researchers make with regards to determining the appropriate time intervals (e.g. how many replies will absolutely be "cut-off" to a different time segment), in order to create network snapshots for longitudinal SNA.

For this analysis eight categories of measures have been applied, a post in the data set is assigned to one of the eight categories based on its relative distance to the initial post (i.e. the thread itself), where the post was posted in. Each of the categories is created on two time boundaries, and is illustrated as follows:

1) Between 1 to 7 days
2) Between 8 to 14 days
3) Between 15 to 21 days
4) Between 22 to 28 days
5) Between 29 to 60 days
6) Between 61 to 90 days
7) Between 91 to 120 days
8) 120 days and more

The choices for these categories are made, based on arbitrary decisions; the first 4 categories are based on 7 days interval, the next 3 categories are based on approximately 30 days interval and the final category would summarize the remainder. The sum of all replies is 35597, which is the total number of replies among all 15 sub-categories, the remaining 5311 posts are the initial posts, which are also known as the threads.

As illustrated in Figure 3.15 and Table 3.22, the overwhelming majority of replies (75.21%) had been posted within the first 7 days once threads were initiated, with only a minority (7.41%) of replies that have been posted in the far distance, i.e. over 120 days and later. The remainders were posted between 8 to 120 days across 6 different time categories. Such results indicate that, a certain percentage of replies could be either included or excluded depending on our decision, regarding data segmentation for longitudinal SNA.

Additionally, an analysis is performed to investigate the last post distribution among different categories. The objective here is to analyze how much impact do last posts have, in terms of distance between the first and last posts. Presumably, if the last posts are frequently posted in distant times (e.g. over 120 days), the contribution of those posts would become quite questionable. The results illustrated there are 8826 replies that have been posted after the first

7 days (in a relative term), in which 1742 (19.7%) are the last posts. This means that 3569 threads (67.2%) have ended within the first 7 days, and 1742 threads (32.8%) ended after the first 7 days, based on the total number of threads (5311 threads) in this research context.

The last post distribution per category is visualized in Figure 3.16. From this visualized representation, the last post distribution is quite similar to the post distribution displayed in Figure 3.15. Furthermore, there are 563 last posts that have been posted in the category of over 120 days, which is about 32.3% out of the 1742 last posts that have been posted after the first 7 days.

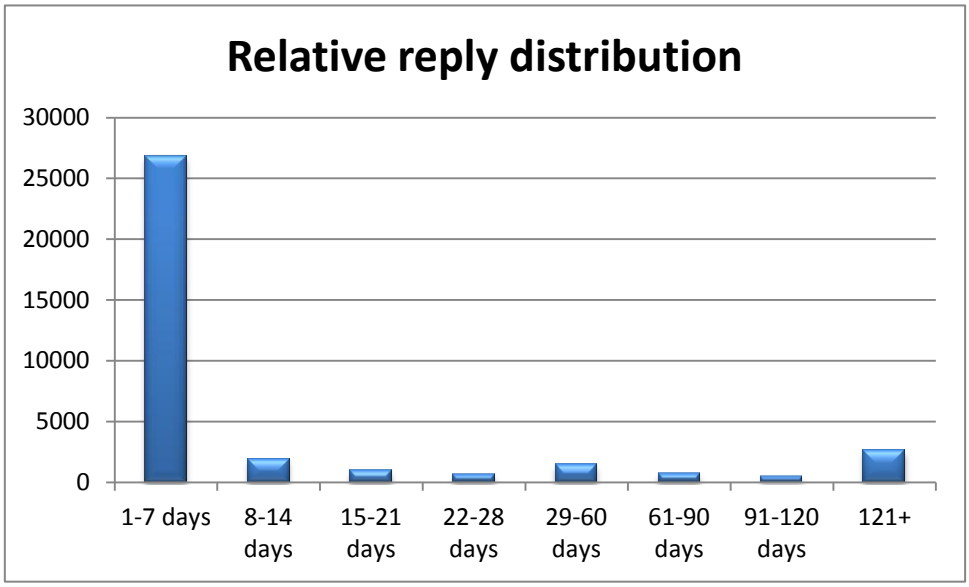


Figure 3.15: Relative reply distribution (whole)

Number of replies	1-7 days	8-14 days	15-21 days	22-28 days	29-60 days	61-90 days	91-120 days	121+
1	697	24	19	14	24	16	10	77
2	1290	55	25	19	51	21	26	62
3	1487	39	22	21	55	20	2	93
4	1810	63	33	23	29	16	5	73
5	1744	65	50	38	81	38	18	93
6	1835	106	40	29	64	17	13	81
7	1854	111	41	37	72	26	20	102
8	1561	80	50	15	71	16	14	79
9	1646	86	29	19	42	43	12	43
10	1269	70	46	28	62	28	5	44
10 or more	11578	1232	633	413	907	466	325	1889
Grand total	26771	1931	988	656	1458	707	450	2636

Table 3.22: Post distribution per number of replies

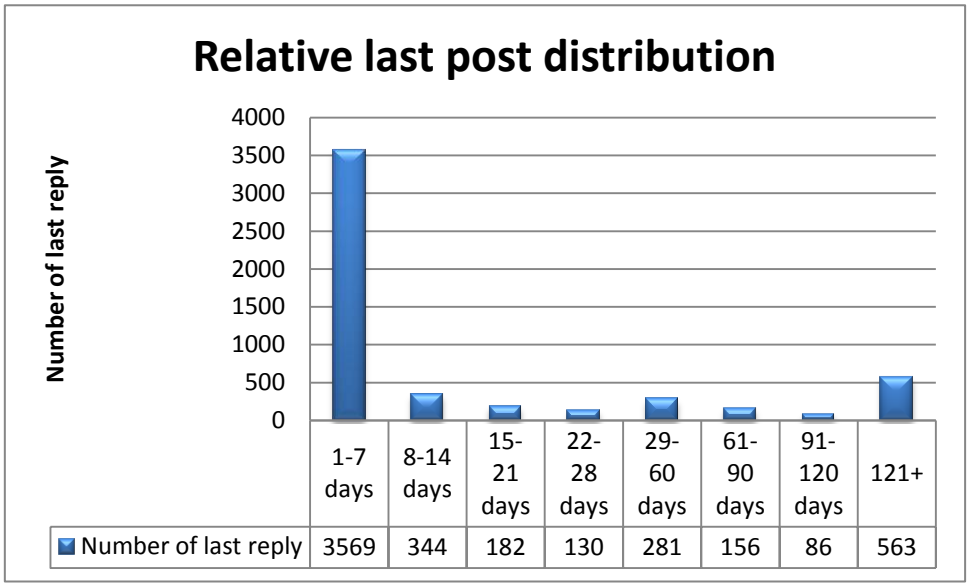


Figure 3.16: Relative last post distribution (whole)

Additionally, to investigate whether such trend occurs in every sub-category, the reply distribution among all sub-categories has been calculated. In Figure 3.17, 4 examples of sub-categories and their reply distributions are illustrated. Based on the results, the general trend of reply distribution remained in a similar range amongst all sub-categories, and maintained acceptable deviations (between +10% and -10%; Table 3.23). Which further strengthens the point stated in previous section, that the overwhelming majority of replies in threads are located close to the thread initiation dates (i.e. within the first 7 days).

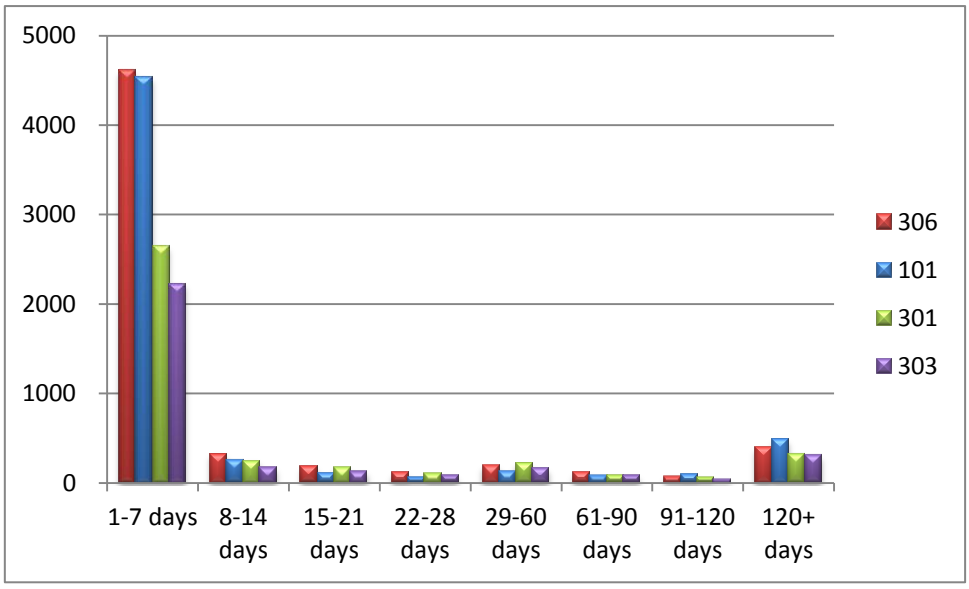


Figure 3.17: Relative reply distribution (4 show cases)

Category	1-7 days	8-14 days	15-21 days	22-28 days	29-60 days	61-90 days	91-120 days	120+ days
101	78.83%	4.41%	1.92%	1.06%	2.30%	1.43%	1.65%	8.41%
104	70.01%	4.87%	2.68%	1.44%	5.42%	6.18%	1.10%	8.30%
301	68.77%	6.44%	4.41%	2.74%	5.68%	2.09%	1.54%	8.34%
302	76.37%	3.82%	3.66%	2.44%	3.66%	1.38%	0.48%	8.18%
303	69.72%	5.30%	3.89%	2.54%	5.14%	2.51%	1.35%	9.56%
304	69.98%	6.79%	2.39%	1.54%	7.02%	1.31%	0.85%	10.11%
305	78.41%	5.40%	2.57%	1.67%	1.29%	1.41%	2.44%	6.81%
306	76.77%	5.41%	2.95%	1.93%	3.23%	2.01%	1.17%	6.53%
404	83.95%	5.11%	2.97%	0.72%	3.89%	0.51%	0.20%	2.66%
501	73.06%	6.96%	3.35%	2.21%	3.29%	2.72%	0.89%	7.53%
502	79.35%	5.22%	1.87%	1.87%	4.23%	2.24%	0.37%	4.85%
503	79.35%	4.87%	2.61%	1.80%	5.14%	1.62%	1.71%	2.89%
601	80.34%	5.66%	1.94%	2.18%	3.88%	0.51%	1.15%	4.35%
604	68.08%	7.14%	1.97%	1.19%	6.16%	2.48%	1.81%	11.17%
605	78.99%	5.13%	1.45%	1.54%	4.16%	2.22%	1.05%	5.45%

Table 3.23: Relative reply distribution percentage wise per sub-category

3.4.3 Data segmentation

As discussed previously, we chose to adopt a time centric approach for longitudinal SNA in this research. Therefore, it is essential to elaborate the specifics regarding the research context and the data at hand. In this section, the objective is to select or provide appropriate data segmentation methods in order to prepare the network structures for the analysis from a longitudinal perspective. In order to conduct a comprehensive analysis for data segmentation, the following sections are focused on three essential aspects: first of all, the results in previous sections indicate that 60 days and 90 days time intervals yielded better outcomes overall, in terms of coverage of unbroken threads, these two time intervals are utilized for further investigation to provide an overview regarding how many replies and edges. Secondly, an issue of time based decay with regard to relations between actors is addressed. The main concern here is to provide a sound solution in order to determine whether a reply can be considered relevant to the research context. Last but not least, a technical solution is introduced to solve the problem of fragmented network snapshots in terms of edge relocation.

Data segmentation on the basis of 60 and 90 days interval

As elaborated previously, Hallo! community possesses a hierarchical structure, and the networks created from this online community are fundamentally, based on thread level. This basically means that the interactions between users are grouped by threads. Therefore, it is important to keep replies jointly in the same network snapshot in order to retain their natural form. However, as demonstrated in previous sections, it is highly unlikely that all threads and their respective replies can be segmented perfectly on basis of static time intervals.

Nevertheless, it is still essential to recognize how many replies will be cut-off from the main groups based on static time intervals. To a certain level, this can be interpreted as the extent to which the network would be fragmented by the data segmentation process for longitudinal SNA.

The results illustrated in previous sections indicate that there are various static time intervals available for creating network snapshots. The goal here, is to choose the most appropriate time intervals for creating network snapshots. The time intervals must satisfy three requirements in order to produce sensible results.

- First, the duration of a time interval should be long enough, so that the network snapshots created over a period of time would be not be oversimplified. For instance, in the context of this research, it does not make much sense to investigate a network snapshot that consists of only a few actors and very limited number of relations.

- Secondly, a time interval should be able to provide a reasonable coverage to prevent the network from being too fragmented.

- Finally, an additional factor has to be taken into consideration, as the number of messages per sub-category varies drastically across the 15 selected sub-categories, this implies that the time interval should be suitable and applicable towards all sub-categories.

In the section Duration analysis, the percentage of unbroken threads based on 60 and 90 days had achieved very high coverage in general (78% and 81%). Furthermore, the overwhelming number of replies of threads were posted within the first 7 days after the thread initiation. Obviously, there is an unavoidable risk of separating replies from the network snapshot where their threads are situated in. The question here is whether the margin of replies that are being broken off from their threads is acceptable. To further illustrate this matter, an additional measure has been taken into consideration, namely, the number of edges that are being isolated from the main groups.

The results of data segmentation on basis of 60 and 90 days are shown in Table 3.24. Reply wise, the number of replies that are isolated from the main groups on basis of 60 and 90 days is quite high (16.88% and 14.12%); however, edge wise, these numbers are increased significantly due to the fact that networks created from threads with more than 10 replies use a different pseudo network creation method. As a result, 41.11% and 35.8% of the edges on basis of 60 and 90 days respectively are cut off from the main network snapshots where the threads are situated in. Such results suggest that additional techniques would be required to further process the edge list created in the previous section, in such a way the fragmented network could be regrouped again based on the segmentation time intervals.

Number of replies	60 days reply wise	60 days edge wise	90 days reply wise	90 days edge wise
Up to 10	1955	1712	1631	1426
More than 10	4048	119072	3397	103747
Total	6003	120784	5028	105173
Percentage	16.86%	41.11%	14.12%	35.8%

Table 3.24: Fragmented replies and edges on basis of 60 and 90 days

Time based decay

Another very important issue which should not be overlooked is whether a reply to a thread is relevant in this research context, e.g. replies are cumulated over time which can be seen as a continuous progress, and it involves a factor of time based decay. For instance, if a reply is posted 200 days later since the last message, it is arguably less relevant in contrast to replies that are stacked together. According to McKerlich, Ives and McGreal (2013), many SNA studies have assumed that a relationship never decays once established, which in reality is not the case; a time decay factor should be taken into consideration. It is therefore, essential to address this issue in a SNA research.

We adopt this approach in our research by further processing the data set with a time based method. We argue that the time span between each message (including threads and replies) is considered as a qualification measure to determine whether a reply is relevant to the research context. A variable is set on 60 days, e.g. if the gap between two messages (whether it is between the thread and its first reply or between two replies in the same thread) is larger than 60 consecutive days, then the later replies would be excluded from the context. The main difference between this approach and setting up a static cap (e.g. 60 days) is that this approach allows the duration of a thread to be prolonged every time a new reply is posted in the thread; in such way that more contents could be included for this research, while still addressing the issue regarding time based decay.

As shown in Table 3.21 previously, in relative terms, the number of replies that were covered by a 60 days interval constitutes a significant portion of the total number of replies (89.4%, i.e. the sum of number of replies between 1-60 days divide by the total number of replies). In this case, the coverage can be further extended on basis of 60 consecutive days, i.e. replies are considered relevant as long as the gap between two messages is smaller than 60 days. The reason of this choice is that 60 days time interval for data segmentation yielded a decent coverage in terms of unbroken threads, and by extending this coverage for another 60 days whenever a new reply is submitted to the thread within this boundary, we could utilize the network data as much as possible, while taking the issue of time based decay into consideration. This is a situational decision, however, as mentioned by McKerlich et al. (2013), such choices are indeed situational, and dependent on the research context.

Based on the approach elaborated above, the results are displayed in Table 3.25. The results are grouped by two categories i.e. threads up to 10 replies and threads with more than 10 replies. In total, there are 2759 replies that are to be excluded from this research;

percentagewise, it is about 7.75% in terms of number of replies. Based on the results, the edge list created from the new data set will exclude those replies for the final longitudinal SNA.

Number of replies	60 consecutive days
Up to 10	995
More than 10	1764
Total	2759
Percentage	7.75%

Table 3.25: Number of replies excluded on basis of a gap of 60 consecutive days

Regroup the edges

From the data set in previous section, a new edge list is re-created from the mixed method elaborated in section 3.3.4 Summary. However, this still does not solve the issue regarding fragmented network structures on basis of static time intervals. As mentioned earlier, this online community upholds a hierarchical construct, and is build around the grouping attribute thread. We recognize the importance of threads and their contribution to the community, and therefore, we propose a technical solution to provide an answer to the aforementioned issue.

As emphasized previously, this online community is based on threads and their replies. In this sense, replies of a thread can be interpreted as a part of the thread, and should therefore be grouped by the thread. Based on this logic, when the network snapshots are created for longitudinal analysis, replies should not be separated or isolated from the thread. This means that if the thread is situated in one of the selected time periods, its respective replies should also be covered by the same time period. In this way, the foundation of network structure will be able to retain its natural form, instead of being fragmented into several pieces.

This objective can be achieved by relocating the edges created from replies that are separated or isolated from the thread by the static time interval. For instance, if the post date of a thread is situated in time interval “01-03-2009 00:00” and “01-06-2009 00:00”, the edges created from replies that are outside of this time interval will be relocated into this time interval. From a technical perspective, this feature is implemented by modifying the post date of the edges to the end date of the time interval minus one minute, i.e. “31-05-2009 23:59” in this given example. Details of edge relocation and descriptive statistics of the final edge list is illustrated in Table 3.26.

Number of unique nodes	5252
Number of replies	32838
Number of edges	252002
Number of edges relocated on 60 days interval	82903 (32.9%)
Number of edges relocated on 90 days interval	67753 (26.9%)

Table 3.26: Descriptive statistics of the final edge list

3.4.4 Summary

In section 3.4, we first argued about which of the longitudinal SNA approaches is suitable to be adopted for this research. The findings of literature review suggest that a time centric approach would be the most appropriate option in the given context. Whereas the actor centric approach lacks the applicability towards open online communities with unstable user base, and there is insufficient prior knowledge with regards to external events, which is essential for the event centric approach to be effective.

From thread duration analysis, the results indicate that static time intervals on basis of 60 and 90 days yield decent overall coverage in terms of unbroken threads for network creation. Furthermore, the reply distribution analysis offered additional insights regarding the relative distribution of replies in their respective threads. However, as posting in online forums is considered as a continuous process, there is an unavoidable risk of defragmenting the network structures when applying static time intervals for creating network snapshots.

To resolve this issue, we proposed a technical solution by relocating isolated edges to the network snapshots of where their respective threads are situated in. This could prevent the network structures from getting fragmented, and further retain the natural form of the networks. As a result, two edge lists are created on basis of 60 and 90 days time interval for the longitudinal SNA in the next chapter.

4 Social network analysis

In the previous chapter, we have illustrated two major elements of SNA in online communities, the network creation processes and adaptation of longitudinal SNA approaches. In this chapter, the focus is towards the final analytical work of SNA from a longitudinal perspective. There are three objectives in the following sections:

1) We identify the applicability of important network metrics, centralities and the metrics of edge-ratio analysis with an extension towards all network structures created from previous sections.

2) The network snapshots are investigated from a longitudinal perspective based on the previously discussed time intervals of 60 and 90 days. The results are contrasted with each other in order to determine the best practice for data segmentation with regards to longitudinal SNA.

3) Visualization of longitudinal development in online communities

With the results from longitudinal SNA, we will finally be able to answer the main research question as well as sub research questions, from theoretical as well as practical perspectives.

4.1 SNA context

As summarized in section 3.3.4, the final edge list for this research is created by using a mixed method (Figure 4.1) based on the results of experimentation as well as observation of literature review, i.e. in case of threads with up to 10 replies, Method 4 will be used, and in case of threads with more than 10 replies, Method 2 will be used. Additionally, the final edge list has been separated into two files in preparation of longitudinal SNA, on basis of the technical solution discussed in section 3.4.4; i.e. edges are regrouped on 60 days and 90 days time intervals. The goal is to relocate edges that have been isolated from their respective threads due to data segmentation processes.

The descriptive statistics of the final SNA context is displayed in Table 4.1. It is worth mentioning that the fourth column in the table represents the unique number of nodes that have been active in the respective sub-category indicated in the first column, and it is possible that these nodes have appeared in other sub-categories during the entirety of the obtained data set. The number of unique nodes per sub-category thus does not represent an absolute value across all sub-categories, and should therefore be considered as a relative value within each individual sub-category.

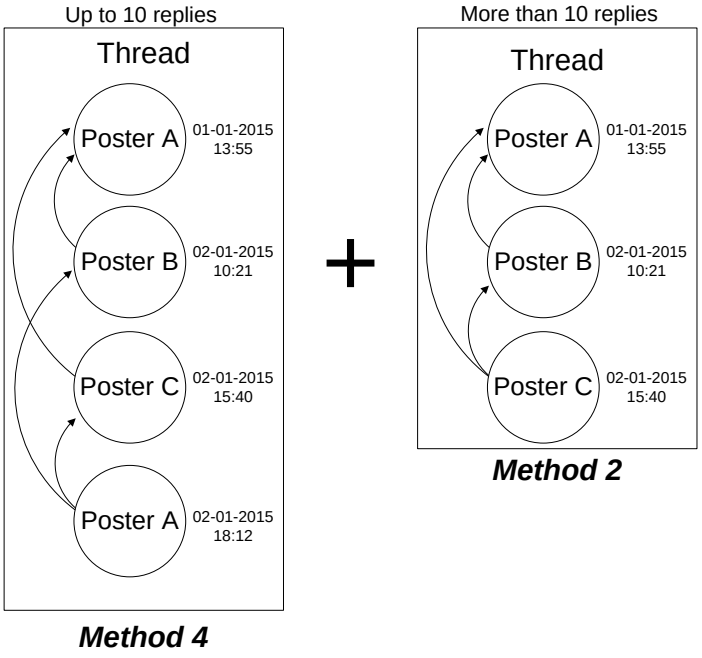


Figure 4.1: Mixed pseudo network creation method

Sub-category ID	Number of threads	Number of replies	Number of unique nodes	Number of edges
101	843	5430	1848	77148
104	215	1323	534	6607
301	663	3496	1191	21774
302	279	1708	681	8788
303	353	2864	984	22819
304	188	1157	558	6331
305	132	697	325	2695
306	774	5568	1339	37775
404	140	949	334	9420
501	241	1452	593	6633
502	107	766	381	6214
503	186	1048	385	4516
601	397	2403	647	17195
604	357	1668	830	12571
605	436	2309	816	11516

Table 4.1: Descriptive statistics of SNA context

4.2 SNA method

In this section, the focus is to elaborate the SNA methods used to conduct longitudinal SNA. The first objective is to describe the selection of network metrics as well as various centrality measures discussed in section 2.6 and 2.7. We utilize those widely applied network metrics and centrality measures (Hanneman et al., 2005; Wasserman et al., 1994) for this research, and implemented them by using Python software with NetworkX library (NetworkX, 2015).

Details regarding technical implementation, such as definition and coding of the Python program is attached in Appendix F. The selected network metrics and centrality measures are displayed in Table 4.2 and 4.3. Those network metrics and centralities are calculated based on the two edge lists illustrated in previous section (60 days and 90 days time intervals). The SNA measures are calculated on the basis of a 60 days interval as well as a 90 days interval respectively. As a result, two sets of records are made available for the final longitudinal SNA.

Set 1: 12 records per sub-category on 60 days interval * 15 sub-categories = 180 records
Set 2: 8 records per sub-category on 90 days interval * 15 sub-categories = 120 records

Moreover, the metrics proposed by Helms et al. (2015) for edge-ratio analysis are utilized to measure the interactions between old and new users. The objective is to investigate the user interaction patterns of this online forum from a network perspective, regarding its longitudinal development by examining the interactions among different groups of users (i.e. old and new users in relative terms). Those edge-ratio metrics are calculated based on the two edge lists as mentioned in previous section (i.e. 60 and 90 days time intervals). The definitions of edge-ratio analysis metrics are displayed in Table 4.4.

Metric	Description
Network density	The ratio between the number of edges in the network and the number of potential edges in the network
Network diameter	The number of hops between the two furthest nodes in the network
Average degree	The average number of edges that are connected to a node
Average weighted degree	The average sum of weights per node in the network
Average clustering coefficient	The number of closed triplets in a network divided by the total number of triplets, generalized for weighted networks
Average length path	The average path length from a node to all other nodes in the network, generalized for weighted networks
Reciprocity	The proportion of ties in the network that are reciprocated

Table 4.2: Network metrics

Centrality	Description
Degree centrality	The number of edges incident on the node
Closeness centrality	The sum of distances from a node to all other nodes in the network, generalized for weighted networks
Betweenness centrality	How often a node appears on the shortest path between other nodes in the network, generalized for weighted networks
Eigenvector centrality	The centrality of a node measured by the connectedness by its neighboring nodes

Table 4.3: Centralities

Edge-ratio metrics	Description
Number of new users	Number of users posting for the first time in a network snapshot.
Number of old users	Number of users posting in a network snapshot, but have also posted in an earlier network snapshot.
New users percentagewise	Relative amount of new users in comparison to the total number of active users in a network snapshot.
Edges between new users percentagewise	Relative amount of edges between new users in comparison to the total number of edges in a network snapshot.
Edges between old users percentagewise	Relative amount of edges between old users in comparison to the total number of edges in a network snapshot.
Edges between old and new users percentagewise	Relative amount of edges between old and new users in comparison to the total number of edges in a network snapshot.
Impact of new users	Ratio of 'edges between new users %' and 'new users%'.
Post/user	Number of posts per user

Table 4.4: Metrics of edge-ratio analysis

4.2.1 Edge-Ratio analysis and SNA measurements

As elaborated by Helms et al. (2015), measuring the development of an online community can be achieved by examining the growth pattern and interaction patterns between users. The growth of an online community can be described as *"having more users joining than leaving the community"*, which is a significant indication of the development of an online community. This can be measured by calculating the number of new users that have appeared in a time segment and the number of old users who have participated in one or more prior time segments, as well as the impact caused by new users.

Additionally, the interaction pattern between users (both old and new) is a direct reflection of relations in an online forum. By measuring the interaction patterns between old, new and a mixture of both type of users, the results could reveal the trend of longitudinal development of the subject community in this research.

Moreover, the standard SNA measurements are calculated in a similar manner, i.e. based on previously illustrated 60 days and 90 days time intervals respectively. The goal is to perform correlation analysis between the edge-ratio metrics and standard SNA measurements, to gain additional insights regarding the relationships between the periodic edge-ratio metrics and their respective metrics and centralities of the network snapshots. For this research, correlation coefficient with p value > .05 will be considered as significant.

4.2.2 Visualization of longitudinal development in online communities

Last but not least, we propose a method for visualizing the longitudinal development of the Hallo! community by using the open source Gephi software tool. The idea is to use the Circular Layout plug-in of Gephi software, to create a circle of users ordered by their first appearance (i.e. the first time when a user submit a message) in a particular sub-category, i.e. the oldest user is positioned at 12 o'clock whereas the newest user is slightly positioned to the left of 12 o'clock in a descending order. Users are then grouped by desired time intervals with regards to the choice of data segmentation, with each of the data segments should have an unique color assigned to it in order to differentiate the time period.

This way, the interactions between old and new users can be visualized in a timed manner, which could provide graphical insights into of the longitudinal development of an online community. A stepwise procedure of this method is described as follows:

1) Import the edge list containing the entire network of its full duration into Gephi
2) Export the node table in panel "Data Table"

3) Edit the exported node table and order the nodes by their registration date
4) Assign appropriate data segmentation values to the nodes retroactively
5) Import the edge lists containing each individual network snapshots into Gephi
6) Export the edge lists containing each individual network snapshots
7) Import the edited node table back into Gephi
8) Choose the ordering value by registration date for option “Order Nodes by (decreasing)” on Layout panel
9) Run Circular Layout
10) Choose “Size/Weight” for option Nodes on Ranking panel, and select degree as rank parameter, and then set the min and max size for the nodes and click on “Apply”
11) Choose the data segmentation value for option Nodes on Partition panel, set a desired color for each data segment, and click on apply
12) Remove all edges from the window where the whole network is situated in
13) Import an individual network snapshot starting from the first data segment
14) Take a snapshot of the generated graph in Preview panel
15) Repeat step 12-14 till all data segments have been visualized and captured

4.3 Results

This section presents the analytical results of the longitudinal SNA methods outlined in section 4.2. In the first sub section, the results of a complete edge-ratio analysis and its metrics are presented on basis of both 60 days and 90 days time interval. The edge-ratio metrics are reported in separate tables for each of the time segmentation method (i.e. 60 and 90 days). Furthermore, a set of line charts are used to provide graphical displays for the demonstrations of how the sub-categories of Hallo! community have evolved over time, based on the results of edge-ratio analysis. In the second sub section, a series of Pearson's correlation analysis is performed, to investigate whether there are relations between the key metrics of edge-ratio analysis and standard SNA measurements. In the last sub section, an attempt is made to provide a visualized representation between interactions of new and old users in relative terms, by using the Circular Layout plug-in of Gephi software.

4.3.1 Edge-Ratio analysis

We begin the edge-ratio analysis by introducing the overall activities across 15 selected sub-categories, in terms of average number of posts per user. In Table 4.5, the average posts per user of each sub-category is displayed. Sub-category 306 "Internet, online marketing & sales" have achieved highest average posts per user (3.4 posts/user) whereas sub-category 502 "Billing & collection" scored the lowest (1.73 posts/user). For illustration purposes regarding edge-ratio analysis, in the following sections we elaborate the results in a within-case analysis focusing on the sub-category 306. Additionally, in order to obtain an overview regarding the longitudinal development of Hallo! community, a cross-case analysis is performed to calculate the average edge-ratio metrics across 15 selected sub-categories. The full analytical results of all 15 sub-categories are attached in Appendix G (tables) and Appendix H (line charts) for both data segmentation methods on basis of 60 days and 90 days time intervals.

Sub-category	**101**	**104**	**301**	**302**	**303**	**304**	**305**	**306**
Post/user	2.55	2.06	2.44	2.11	2.32	1.83	1.88	3.40
Sub-category	**404**	**501**	**502**	**503**	**601**	**604**	**605**	/
Post/user	2.38	2.13	1.73	2.33	3.19	1.87	2.33	/

Table 4.5: Average number of posts per user in 15 sub-categories

As illustrated earlier, according to Helms et al. (2015), the longitudinal development of an online community can be measured by analyzing the growth pattern and interaction pattern among the users. This is achieved by conducting edge-ratio analysis. The edge-ratio metrics have been calculated for sub-category 306, on basis of 60 days time interval (Table 4.6) and 90 days time interval (Table 4.7).

Sub-category 306	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10	P_11	P_12
Number of new users	180	107	125	117	115	146	117	91	67	117	120	42
Number of old users		36	52	63	56	69	71	69	53	69	73	48
New users' %		74.83%	70.62%	65.00%	67.25%	67.91%	62.23%	56.88%	55.83%	62.90%	62.18%	46.67%
Edges between new users %		28.06%	30.99%	27.21%	28.46%	31.27%	36.38%	11.10%	6.81%	10.72%	13.19%	3.59%
Edges between old users %		18.18%	21.47%	30.93%	16.99%	22.96%	15.89%	50.27%	54.24%	40.26%	33.35%	72.91%
Edges between old and new users %		53.75%	47.54%	41.85%	54.55%	45.77%	47.73%	38.64%	38.95%	49.02%	53.47%	23.51%
Impact of new users		0.38	0.44	0.42	0.42	0.46	0.58	0.20	0.12	0.17	0.21	0.08
Post/user	2.97	3.16	3.26	2.94	2.38	2.69	2.81	2.71	3.35	4.26	4.23	3.23

Table 4.6: Results of edge-ratio analysis, sub-category 306, 60 days time interval

Sub-category 306	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Number of new users	239	173	186	192	151	124	188	91
Number of old users		55	69	85	83	67	84	85
New users' %		75.88%	72.94%	69.31%	64.53%	64.92%	69.12%	51.70%
Edges between new users %		34.31%	29.24%	32.64%	37.12%	11.13%	15.22%	5.04%
Edges between old users %		18.15%	27.33%	15.73%	15.42%	45.77%	33.28%	52.83%
Edges between old and new users %		47.54%	43.42%	51.63%	47.47%	43.10%	51.51%	42.13%
Impact of new users		0.45	0.40	0.47	0.58	0.17	0.22	0.10
Post/user	3.15	3.56	2.96	2.74	3.05	3.40	4.54	3.78

Table 4.7: Results of edge-ratio analysis, sub-category 306, 90 days time interval

The tables displayed above show the edge-ratio metrics calculated for sub-category 306, based on 60 and 90 days time interval. Edge-ratio metrics are measured based on either on 12 periods (60 days per period), or 8 periods (90 days per period) respectively. The certain cells are left empty intentionally, in the first period (i.e. P_1) for both tables. This is due to the fact that during the first period, all users are considered as "new" users, therefore the metrics involving "old" users result in 0%, whereas metrics for "new" users result in 100%, is pointless to displayed either 0% or 100 % of those metrics.

In terms of growth pattern, the number of new user is calculated per period (i.e. P_1). Users that have remained active (e.g. reappeared in one or more later periods) are counted as old users in relative terms (shown in row "number of old users" in both tables). Moreover, the new users' % and their relative impact in each period are calculated. In terms of interaction pattern, edges between new users, old users and a mixture of both are presented in the respective period. Visualized representations for both data segmentation methods are displayed in Figure 4.2 and 4.3. These figures can be interpreted as a trend analysis of the edge-ratio metrics on basis of 60 days and 90 days time interval.

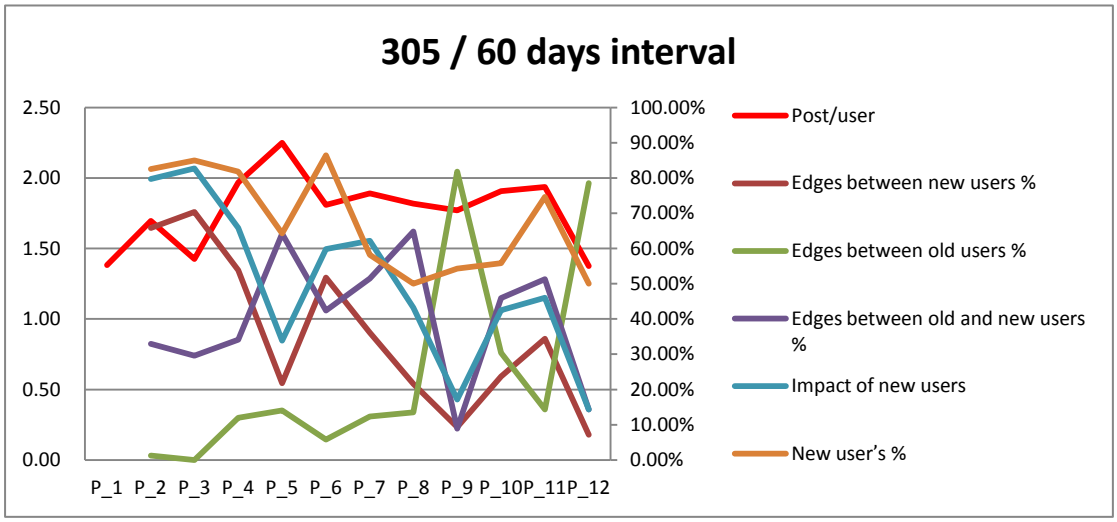


Figure 4.2: Line chart of edge-ratio metrics, sub-category 306, 60 days interval

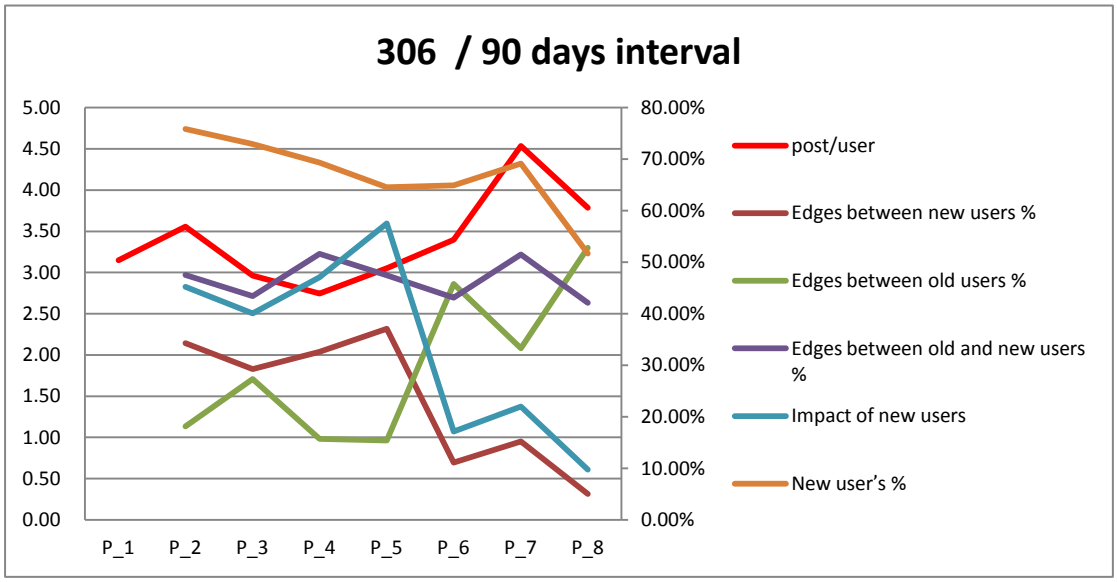


Figure 4.3: Line chart of edge-ratio metrics, sub-category 306, 90 days interval

In order to create an overview of the edge-ratio metrics for sub-category 306, the average values have been calculated (Table 4.8) for both 60 days and 90 days interval. In terms of average new and old users, there is a significant difference, which is to be expected. Over a period of 90 days, more new users would be included per period in contrast to a period of 60 days. Additionally, the average new users' % in terms of 60 days interval is slightly lower (62.94% vs. 66.92%) than 90 days interval. This is caused by the more frequently segmented data, e.g. users remained active in later periods are counted repeatedly, which indicates that the sum of old users in 12 periods (60 days interval) is higher than the sum of old users in 8 periods (90 days interval).

Overall, the interaction patterns between old, new and the mixture of both types of users indicate that the interactions between old and new users (44.98% on 60 days basis, 46.69% on 90 days basis) contribute the most to the longitudinal development in contrast to the

interactions between old or new users. Additionally, the interactions between old users contributes more than the interactions between new users (34.31% and 29.79% vs. 20.71% and 23.53%).

Sub-category 306	60 days interval (avg. over 12 periods)	90 days interval (avg. over 8 periods)
Avg. number of new users	112	168
Avg. number of old users	59.91	75.43
Avg. new users' %	62.94%	66.92%
Avg. edges between new users %	20.71%	23.53%
Avg. edges between old users %	34.31%	29.79%
Avg. edges between old and new users %	44.98%	46.69%
Avg. impact of new users	0.32	0.34

Table 4.8: Average edge-ratio metrics of sub-category 306

For the cross-case edge-ratio analysis, the average edge-ratio metrics have been calculated on basis of 60 and 90 days time interval, across all 15 sub-categories that have been selected for this research. The results are shown in Table 4.9. Due to variations in the number of users per sub-category, the average number of new and old users are significantly lower than the case (i.e. the most populated sub-category 306) illustrated in previous section.

The results of cross-case analysis indicate that the interactions between old and new users have remained as the main contributors (45.51% and 45.35%) toward the evolution of Hallo! community, in contrast to the other two types of user interactions, in both data segmentation methods. However, cross-case wise, the interactions between new users have increased slightly, in comparison to the results of sub-category 306. As a result, interactions between new users have become the second source of contribution towards the longitudinal development of Hallo! community.

15 sub-categories	60 days interval (avg. over 12 periods)	90 days interval (avg. over 8 periods)
Avg. number of new users	63.96	95.94
Avg. number of old users	30.76	37.78
Avg. new users' %	66.57%	71.15%
Avg. edges between new users %	28.23%	31.61%
Avg. edges between old users %	26.26%	23.04%
Avg. edges between old and new users %	45.51%	45.35%
Avg. impact of new users	0.41	0.43

Table 4.9: Average edge-ratio metrics of Hallo! community

4.3.2 Edge-ratio analysis and SNA measurements

In this sub section, the objective is to investigate whether there are correlations between the edge-ratio metrics and their respective SNA measurements from a longitudinal perspective. The Pearson's correlation coefficients have been calculated between 4 edge-ratio metrics and SNA measurements on basis of both data segmentation methods. The First periods from both data segmentation methods have been excluded from the calculation due to their nature deficiencies elaborated previously (i.e. all users are considered as new users in the first period). To improve the readability, we utilize the guidelines regarding the strength of correlation coefficients illustrated by Evans (1996) and denote strong correlations (.6 - 1) with "S", moderate correlations (.4 - .59) with "M" and weak correlations (.2 - .39) with "W", whereas positive correlations are indicated as (+) and (-) for negative correlations. The cells are left empty in case of no correlation or very weak correlation (0 - .19) between edge-ratio metrics and SNA measurements. The results are shown in Table 4.10 and Table 4.11.

N = 165	Edges between new users %	Edges between old users %	Edges between old and new users %	Impact of new users
Density				
Diameter				
Average clustering		W (+)	W (-)	
Average path				
Reciprocity				
Average weighted degree		W (+)		
Average degree				
Degree centrality				
Betweenness				
Closeness	W (-)	W (+)		W (-)
Eigenvector				

Table 4.10: Pearson's correlation coefficients between edge-ratio metrics and SNA measurements (on basis of 60 days interval)

N = 105	Edges between new users %	Edges between old users %	Edges between old and new users %	Impact of new users
Density				
Diameter				
Average clustering		W (+)	W (-)	
Average path				
Reciprocity				
Average weighted degree	W (-)	W (+)		W (-)
Average degree				
Degree centrality				
Betweenness				
Closeness		W (+)		
Eigenvector				

Table 4.11: Pearson's correlation coefficients between edge-ratio metrics and SNA measurements (on basis of 90 days interval)

Overall, there is no strong or moderate correlation between the 4 edge-ratio metrics and SNA measurements in both data segmentation methods. In some cases, weak correlations have been found. One possible explanation for this, is that the evolution of Hallo! community is established on basis of all three types of user interactions, i.e. interactions between new, old and a mixture of both types of users. A single type of user interaction alone is insufficient to describe the longitudinal development of this community.

4.3.3 Visualizing the evolvement of an online community

The stepwise procedure elaborated in section 4.2.2 was executed accordingly, in order to visualize a selected sub-category (306) regarding its longitudinal development. The users of sub-category 306 are ordered on the basis of their first appearance in this sub-category in a chronological order, i.e. the oldest user is positioned exactly at 12 o'clock on the circular layout, whereas the newest user is placed slightly to the left of 12 o'clock. 12 different unique colors have been assigned to data segments respectively, based on 90 days time interval. Figure 4.4 displayed the user distribution per data segment percentagewise, and on the right side of this figure, the exact color assigned to each corresponding data segment is marked. It is worth noticing that the percentage of data segment 8 is considerably lower in contrast to the other data segments, this is partially due to the fact that the last data segment officially ended on 25-01-2011 (the last posting date), whereas the time period of this segment ended on 18-02-2011.

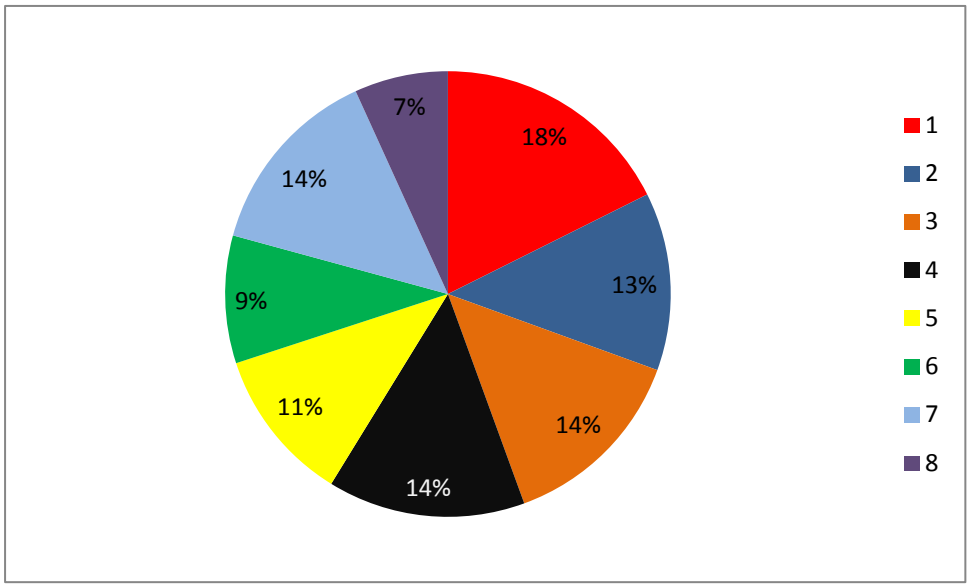


Figure 4.4: User distribution per data segment on basis of 60 days interval

Figure 4.5 demonstrates the longitudinal development of sub-category 306. Positions of the nodes are created according to the first appearances of users, and ordered chronologically (oldest to newest). Red nodes indicate that the users registration dates are situated in the first data segment (the first 90 days), whereas nodes colored in light blue are the users registered in the last data segment (the last 90 days) between 01-03-2009 and 18-02-2011. Each data segment has a numerical value assigned to it, those values are between 1 and 8, that further illustrated the data segmentation on 90 days interval. The size of a node is determined by its degree (in-degree + out-degree), higher degree will increase the size of the node.

This figure showed the interactions between users in sub-category 306 throughout the years. As time goes by, the progress continued to move clockwise toward the completion of a full circle with a duration of almost 2 years. Due to our technical solution for data segmentation, a certain number of edges from different time periods have been relocated to the network snapshots of where their respective threads were situated in, however, the first appearances of those users [nodes] remained unchanged. This can be spotted on various occasions, e.g. on segment 3, the black dots located at 6 & 7 o'clock of the circular layout are the users appeared in the fourth period (between 26-11-2009 and 24-02-2010), but are included in the visualization of the third period.

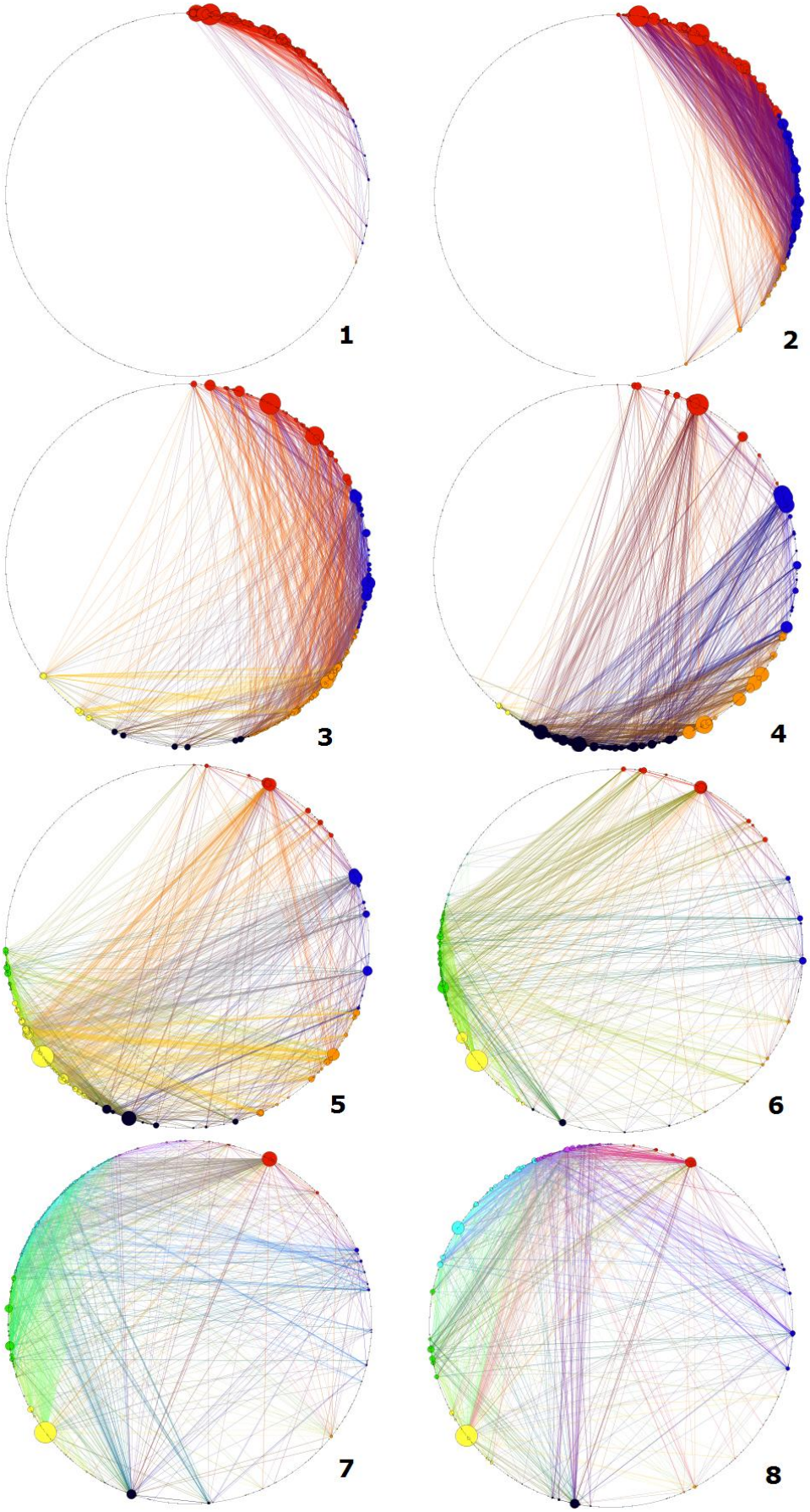


Figure 4.5: Longitudinal development of sub-category 306

5 Discussion

This chapter is dedicated to discussions regarding the findings and results in previous chapters, within the scope of the main research question and the sub-research questions stated in section 1.2. It concludes with addressing the limitations encountered during this research project.

Sub-question 1: *"What are the existing scientific methods for structure creation of online communities and social network analysis?"*

From a theoretical point of view, the existing literature provided 4 valuable methods for constructing social network structures. The applicability of those methods is situational in terms of the type [structured or unstructured] of available data extracted from an online community.

SNA of online communities is arguably a perfect match, whereas the data extracted from online communities could provide a huge and resourceful research context to scholars in this field of study. Based on the given context of an online community, various structural characteristics can be identified accordingly, in preparation for SNA. SNA techniques in combination with statistical analysis can then, be utilized to investigate those structural characteristics of the data from online communities, in order to contribute new ideas and insights to this ever-growing virtual society.

In this research we adopted a directional and weighted network setting based on our research context. Interaction based network structures were created by the self-created Python program, and then analyzed by using various SNA techniques based on NetworkX library in Python software and statistical analysis with the help of SPSS.

Sub-question 2: *What are the existing methods for network creation based on unstructured data?*

The most important factor in answering this sub-question, is whether a data set contains relational references between the actors. With explicit relational references between actors (i.e. *structured data*), the network creation method is simply straightforward: to connect the actors based on the relational references. A network created with explicit relational reference is much more reliable in contrast to networks created based on presumption. Without relational references (referring to *unstructured data*), researchers rely on *grouping attributes* from a data set in order to establish connections between actors. Such pseudo network creation methods often are based on presumption of the relationships without further elaboration.

In this research, the available data falls in the latter category, i.e. unstructured data without relational reference, and we have to rely on pseudo network creation methods for construction of network structures. To avoid the pitfall mentioned in previous paragraph (network creation

on the basis of presumptions), in the experimentation phase of this research, an extensive content analysis have been carried out. The goal is to validate whether the existing pseudo network creation methods are able to represent the reality, by comparing the network metrics and conducting quadratic assignment procedure analysis between the baseline network created from content analysis and networks created by pseudo methods. As a result, we identified certain deficiency, in terms of SNA measurements from the existing pseudo network creation methods in contrast to the baseline network, and proposed a mixed method based on results of our observation as well as from existing literature.

Sub-question 3: *What are the existing research perspectives regarding longitudinal analysis of online communities?*

From literature review, three symbolic approaches have been identified: 1) actor centric, 2) time centric and 3) event centric. The actor centric approach emphasizes the development of individual actors over time, in terms of their influences toward others in the same community. The time centric approach focuses on the study of network fragments created from non-overlapping time periods of the entire network; this way, the dynamic and longitudinal nature of online communities is taken into consideration. Finally, the event centric approach argues that the SNA of online communities should target the external events as the driving factor regarding the longitudinal development of online community.

In this research, we utilized the time centric approach due to the fact that this approach is more appropriate in the given context. The Hallo! community is an open online forum, which indicates that its user base is unstable in contrast to closed networks (e.g. internal forum of an organization), hence an actor-centric approach with its requirement for continuous monitoring of the same set of actors would not work well here. Further, as there is a lack of prior knowledge with regards to external events, we deem the event centric approach inapplicable for this research.

To circumvent arbitrary decisions on data segmentation for creating network snapshots, the characteristics of the hierarchical structure of this community have been investigated in depth. As a result, 60 days and 90 days time interval have been proven to be effective in terms overall coverage, and were therefore selected for data segmentation. The reason for including two different types of data segmentation, is to compare the two sets of network snapshots in practice, in order to determine the most appropriate data segmentation process for this research. However, based on the results of edge-ratio analysis, no major misalignment has been revealed between these two data segmentation methods. Additionally, a technical solution was proposed to protect network structures from undesired fragmentation. This is achieved by relocation of edges that are isolated from their respective data segments where the threads were initiated. This way, the natural form of the network structures can be

retained.

Sub-question 4: *What are the important metrics for measuring the development of online communities?*

A set of standard SNA metrics and centrality measures were selected from literature review. Those SNA measurements have been widely applied in this field of research, for measuring network characteristics. In the context of this study, edge-ratio metrics are a set of suitable measures in terms of monitoring the longitudinal development of this community. Aforementioned measurements have been implemented by using Access database, MS Excel, Python software and its NetworkX library, based on 60 days and 90 days network snapshots described previously.

The edge-ratio analysis offers two important analytical approaches regarding the longitudinal development of an online community: 1) growth pattern and 2) interaction pattern. By analyzing the growth pattern, the results indicate that over the duration of two years, there was a constant stream of new users in each time period. At the same time, a majority of old users would become inactive (e.g. stopped posting or left the community). Despite the high turnover ratio of new users in general, the impact of new users did not pair with it. This is further illustrated by analyzing the interaction patterns among new and old users. The results show that the interactions between new and old users surpassed the other two types of user interactions (i.e. interactions between new users and interactions between old users), as the driving force towards the longitudinal development of Hallo! community. This implies that the evolution of a community does not sole rely on "fresh blood", but also relies heavily on the senior users, i.e. more experienced users.

Moreover, a cross-case correlation analysis was performed to investigate whether there were relations between the edge-ratio metrics and their respective SNA measurements. However, there were a few cases of weak correlations found, that were statistically significant between the edge-ratio metrics and standard SNA measurements.

In addition to that, an attempt has been made to visualize the longitudinal development of Hallo! community. The Circular Layout plug-in in Gephi was used as basis for this visualization attempt. As a result, a series of graphical representations were created on the basis of 8 network snapshots based on data segmentation of 90 days interval. Each of the snapshots presents the interactions between users during that particular data segment based on a chronological order of first appearances of the users. It is an interesting perspective for monitoring the network's growth over time.

Practical implication

A considerable portion of time for this thesis project was dedicated to research the existing network creation methods based on unstructured data retrieved from an online community. In addition to that, three symbolic longitudinal SNA approaches have been evaluated, in order to determine which one is the most appropriate choice in the given context. A number of practical implications can be illustrated on basis of the results of extensive experimentation conducted during this research regarding these subjects.

1) To properly evaluate pseudo network creation methods for unstructured communication data, a baseline network that represents the reality is a crucial factor to success. Without advanced content analysis tools, an alternative approach can be applied to create a baseline network. This can be achieved by selecting samples that represent the reality and conduct content analysis manually, in order to obtain sufficient knowledge regarding the missing relational references between actors in the network.

2) By comparing SNA metrics and performing QAP between pseudo networks and the baseline network, researchers will be able to gain additional insights to identify the differences between pseudo networks and the baseline network. And make the decision thereafter for choosing the best options available, in order to create a reliable network structure on basis of unstructured data for SNA.

3) It is essential to obtain sufficient knowledge of the research context, in terms of its technical characteristics, before adopting an approach for longitudinal SNA. Some of the longitudinal SNA approaches may be less or not applicable towards all research contexts.

4) In this particular research context, a time centric approach is utilized. In order to determine the most appropriate time interval(s) for data segmentation, various features related to the online forums (e.g. thread duration, post distribution, last post distribution etc.) have been analyzed in-depth. In such a way, the choices of data segmentation methods could be justified scientifically to avoid arbitrary decisions as much as possible.

Limitation

As with any research in this field, we have also encountered limitations along the way. First of all, only 15 sub-categories were cherry picked from the 35 sub-categories of Hallo! community. This selection was made on the basis of a parameter of minimal one post per day on average during the entirety of the data set. The value of this parameter is however arbitrary, based on close (manual) observation of this data set.

Secondly, despite our best effort, the mixed pseudo network creation method proposed and implemented in this research, is still imperfect (i.e. measured by SNA metrics, centralities and QAP analysis) with regards to its accuracy in comparison to the reality. Although, this mixed pseudo network creation method is superior in contrast to other existing network creation methods for unstructured data, there is still room for improvement in the future.

Finally, in the data segmentation process, in order to protect the hierarchical network [snapshot] structure, a trade-off was made by using the proposed technical solution, which relocated a small portion of edges manually to their originating segment.

6 Conclusion & future work

This research has explored various state-of-the-art SNA techniques and statistical analysis in order to investigate the longitudinal [in-time] development of a public online community's hierarchical structure from a network's perspective. Experimentations were conducted with regards to the applicability and validities of existing literature in this field of research as well as revealing their practical implications.

Overall, the results of our work revealed a number of crucial issues related to utilizing cumulative and unstructured communication data, in researching the evolution of a particular type of online community, i.e. an online forum. It can be concluded that when applying SNA for researching communication data without explicit relational references (i.e. unstructured data) extracted from an online community regarding its longitudinal development, it is essential to validate the accuracy of the networks created from pseudo network creation methods against baseline networks (i.e. networks that represent the reality), by contrasting SNA metrics as well as performing QAP between these networks, before choosing pseudo network creation method(s). Moreover, it is important to evaluate the applicability of longitudinal SNA approaches based on the given research context. In addition to that, significant characteristics of the online community need to be analyzed, in order to justify the choices made with regard to appropriate time intervals for data segmentation scientifically.

By utilizing edge-ratio analysis and its related metrics, the growth pattern of the online community and interaction patterns among users can be identified clearly, for the longitudinal SNA in the given context. Additionally, the development of an online community can be monitored by calculating SNA measurements periodically, on the basis of data segments created from cautiously chosen time intervals. Furthermore, the results indicate that the evolvement of an online community does not solely rely on the turnover ratio of its membership; the participation rate of "old" users is just as important as the streams of new users. Finally, the graphical representations created by using Gephi Circular Layout plug-in of this research provided distinctive insights regarding the longitudinal development of an online community.

Future work

As mentioned previously, the context of this research is based on the data extracted from a public online forum. Despite the original intentions (i.e. sharing business experience among Dutch entrepreneurs) of the community owners, the data (83.3% of the threads ended within 10 replies) suggests that the forum was being used as a Q&A styled forum. For future work, it would be interesting to conduct similar research as described in this study, on other online communities with comparable characteristics.

Reference

Adamic, L. a, Zhang, J., Bakshy, E., Ackerman, M. S., & Arbor, A. (2008). Knowledge Sharing and Yahoo Answers : Everyone Knows Something. *International World Wide Web Conference Committee*, 665–674.

Agarwal, R., Gupta, A. K., & Kraut, R. (2008). Editorial Overview–The Interplay Between Digital and Social Networks. Information Systems Research, 19(3), 243-252.

Agichtein, E. (2008). Finding the Right Facts in the Crowd : Factoid Question Answering over Social Media Categories and Subject Descriptors. *Search*, 467–476.

Agichtein, E., Donato, D., Gionis, C., Mishne, G. & Castillo, C. (2008). Finding High-quality Content in Social Media. *International Conference on Web Search and Data Mining*, 183–194.

Backstrom, L., Dwork, C., & Kleinberg, J. (2007). Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. *Proceedings of the 16th International Conference on the World Wide Web (WWW), 54*, 181–190.

Barrat, A., Barth ́elemy, 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

Berger, K., Klier, J., Klier, M., & Richter, a. (2014). Who is Key?-Value Adding Users in Enterprise Social Networks, 1–16.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology.*

Bonacich, P., & Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, *23*, 191–201.

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). Ucinet for Windows: Software for Social Network Analysis. *Harvard Analytic Technologies, 2006*, SNA Analysis software.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science (New York, N.Y.), 323*(5916), 892–895.

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*, 210–230.

Cha, M., Haddai, H., Benevenuto, F., & Gummadi, K. P. (2010). Measuring User Influence in

Twitter : The Million Follower Fallacy. *International AAAI Conference on Weblogs and Social Media*, 10–17.

Christley, S. & Madey, G. (2007). Global and Temporal Analysis of Social Positions at SourceForge.net. *In The Third International Conference on Open Source Systems (OSS 2007), IFIP WG 2.13*. Limerick, Ireland.

Correa, T., Hinsley, A. W., & de Zúñiga, H. G. (2010). Who interacts on the Web?: The intersection of users' personality and social media use. *Computers in Human Behavior, 26*(2), 247–253.

Culnan, M. J., McHugh, P. J. & Zubillaga, J. I. (2010). How Large U.S. Companies Can Use Twitter and Other Social Media to Gain Business Value. *MIS Quarterly Executive, 9*(4), 243–259.

Dignum, V., & Eeden, P. (2005). Seducing, engaging and supporting communities at Achmea. *Knowledge Management and Management Learning, 9*, 125–141.

Eisenhardt, K. (1989). Building Theories from Case study research. *Academy of management review, 14*(4), 532-550.

Emirbayer, M., (1997). Manifesto for a Relational Sociology. *American Journal of Sociology 103*(1997)2, 281- 317.

Faasse, R. E. M., Helms, R. W., & Spruit, M. R. (2011). Web 2.0 in the CRM domain: defining Social CRM. *International Journal of Electronic Customer Relationship Management, 5*(1), 1–22.

Facebook. (2014). http://newsroom.fb.com/company-info/

Falkowski, T., Barth, A., & Spiliopoulou, M. (2008). Studying Community Dynamics with an Incremental Graph Mining Algorithm. In AMCIS 2008 Proceedings.

Faraj, S., & Johnson, S. L. (2011). Network Exchange Patterns in Online Communities. *Organization Science, 22*, 1464–1480.

Farrell, H. (2012). The Consequences of the Internet for Politics. *Annual Review of Political Science, 15*, 35–52.

Fischer, E. & Reuber, a. R. (2011). Social interaction via new social media: (How) can interactions on Twitter affect effectual thinking and behavior? *Journal of Business Venturing, 26*(1), 1-18.

Fisher, D., Smith, M., & Welser, H. T. (2006). You Are Who You Talk To: Detecting Roles in Usenet Newsgroups. *In Proceedings of the 39th Annual Hawaii International Conference on*

System Sciences, 3, 59b–59b.

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40, 35.

Gillin, P. & Schwartzman, E. (2011). Social Marketing to the Business Customer: Listen to Your B2B Market, Generate Major Account Leads, and Build Client Relationships. Wiley.

Gleave, E., & Welser, H. T. (2009). A conceptual and operational definition of'social role'in online community. *42nd International Conference on System Sciences, 1–11*.

Gloor, P., Laubacher, R., Zhao, Y., & Dynes, S. (2004). Temporal Visualization and Analysis of Social Networks. NAACSOS Conference, 27–29.

Hanneman, A. & Riddle, M. (2005). *Introduction to social network methods.* University of California, Riverside.

Hallo! Community. (2015). Welkom bij Hallo!. Home – Hallo!. Retrieved July 10 2015, from: http://hallo.kvk.nl/hallo/

Hansen, D. L., Shneiderman, B. & Smith, M. a. (2011). Analyzing Social Media Networks with NodeXL. *Analyzing Social Media Networks with NodeXL*, 11–29.

Hawn, C. (2009). Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care. *Health Affairs (Project Hope), 28*(2), 361–8.

Helms, R., & Buijsrogge, K. (2006). Application of Knowledge Network Analysis to identify knowledge sharing bottlenecks at an engineering firm. *Proceedings of the 14th European conference on Information Systems, 1-13*.

Helms, R. & Majdan, G, (2015). DYNAMICS IN ONLINE COMMUNITIES: A LONGITUDINAL NETWORK ANALYSIS. Unpublished manuscript, Utrecht University.

Hoffman, D.L. & Fodor, M. (2010). Can You Measure the ROI of Your Social Media Marketing? *MIT Sloan Management Review*. 52, 1.

Howison, J., Inoue, K., & Crowston, K. (2006). Social dynamics of free and open source team communications. In *Proceedings of the IFIP 2nd International Conference on Open Source Software, IFIP International Federation for Information Processing*, 203, 319-330.

Howison, J., Wiggins, A., & Crowston, K. (2010). Validity Issues in the Use of Social Network Analysis for the Study of Online Communities. *Journal of the Association for Information Systems, 1*2(12).

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277–1288.

Huffaker, D. (2010). Dimensions of Leadership and Social Influence in Online Communities. *Human Communication Research, 36*, 593–617.

Hunter, B. (2002) 'Learning in the Virtual Community Depends upon Changes in Local Communities', in K.A. Renninger and W. Shumar (eds) *Building Virtual Communities: Learning and Change in Cyberspace*, 96–126. Cambridge: Cambridge University Press.

Igarashi, T. (2005). Gender differences in social network development via mobile phone text messages: A longitudinal study. *Journal of Social and Personal Relationships, 22*, 691–713.

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS ONE, 9.

Jalali, S. & Wohlin, C. (2012). Systematic literature studies: database searches vs. backward snowballing, *Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement*, 19-20.

Jones, Q. (1997). Virtual-Communities, Virtual Settlements & Cyber-Archaeology: A Theoretical Outline. *Journal of Computer-Mediated Communication*, 3: 0.

Jussila, J. J., Kärkkäinen, H., & Leino, M. (2011). Benefits of social media in business-to-business customer interface in innovation. *Proceedings of the 15th International Academic MindTrek Conference on Envisioning Future Media Environments - MindTrek '11*, 167.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons, 53*(1), 59–68.

Kazienko, P., Musial, K., & Kajdanowicz, T. (2008). Profile of the Social Network in Photo Sharing Systems. In AMCIS 2008 Proceedings.

King, R., Morgan, B. J. T., Gimenez, O., Brooks, S. P., Crc, H., & Raton, B. (2010). *Journal of Statistical Software, 36*(August), 1–2.

Knoke, D & Yang, S. (2008). *Social network analysis*. Sage, Newberry Park, CA

Komito, L. (1998). The net as a foraging society: Flexible communities. *The Information Society, 14*, 97–106.

Kossinets, G., & Watts, D. J. (2006). Empirical analysis of an evolving social network. Science (New York, N.Y.), 311(5757), 88–90.

Lazard, J. & Preece, J. (1998). Classification schema for online communities. *In Proceedings of the 1998 Association for Information Systems,* 84-86

Lee, J. & Lee, H. (2008). The computer-mediated communication network: exploring the linkage between the online community and social capital. *New Media & Society, 12,* 711–727.

Leek, J., Collado-Torres, L. & Reich, N. (2013). How to share data with a statistician. GitHub. Retrieved April 7, 2015, from https://github.com/jtleek/datasharing.

Luce, R., & Perry, A. (1949). A method of matrix analysis of group structure. *Psychometrika, 14,* 95– 116.

Majdán, G. (2012). *The success and sustainability of Online Communities: a Social Network Analysis approach*. Unpublished manuscript, Utrecht University.

Mascolo, C. (2011). Social and Technological Network Analysis. University of Cambridge. Retrieved from https://www.cl.cam.ac.uk/teaching/1314/L109/stna-lecture3.pdf

McKerlich, R., Ives, C., & McGreal, R. (2013). Measuring use and creation of open educational resources in higher education. *International Review of Research in Open and Distance Learning, 14*(4), 90–103.

McLure Wasko, M., & Faraj, S. (2005). Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice. *MIS Quarterly, 29*(1), 35 - 57.

Moody J., McFarland D., Bender-DeMoll S. (2005). Dynamic Network Visualization. *American Journal of Sociology, 110*(4), 1206–1242.

Mangold, W. G., & Faulds, D. J. (2009). Social media: The new hybrid element of the promotion mix. *Business Horizons, 52*(4), 357–365.

Moran, M., Seaman, J., & Tinti-Kane, H. (2011). Teaching, Learning, and Sharing: How Today's Higher Education Faculty Use Social Media. *Babson Survey Research Group*, (April), 1–16.

Moss, C. P. a, & Salkind, E. N. J. (2014). Encyclopedia of Research Design " Validity ", 1589–1592.

NetworkX. (2015). Reference. Reference - NetworkX. Retrieved Auguest 12 2015, from: http://networkx.readthedocs.org/en/stable/reference/index.html

Newman, M. (2010). *Networks*. Oxford University Press.

Otte, E. & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science, 28*(6), 441–453.

Owston, R. (1998) *Making the Link: Teacher Professional Development on the Internet.* Portsmouth, NH: Heinemann.

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, 911–932.

Petróczi, A., Nepusz, T., and Bazsó, F. (2006). Measuring tie-strength in virtual social networks. *Connections, 27*(2) , 39-52

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*, 829–854.

Preece, J. and Maloney-Krichmar, D. (2005), Online Communities: Design, Theory, and Practice. *Journal of Computer-Mediated Communication*.

Ridings, C., Gefen, D., & Arinze, B. (2002). Some antecedents and effects of trust in virtual communities. *Journal of Strategic Information Systems, 11* (3–4), 271–295.

Ridings, C. & Gefen, D. (2004). Virtual Community Attraction: Why People Hang Out Online. *Journal of Computer-Mediated Communication. 10*(1). 0-0.

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

Sack,W. (2000). Conversation map: an interface for very-large-scale conversations. *J. Manag. Inf. Syst. 17*, 73–92

Scott, J. (2000). Social Network Analysis: A Handbook. Contemporary Sociology.

Smith, A. (2011). Why Americans use Social Media. *Pew Research Center*, 1–7.

Snijders, T. A. B. (1996). Stochastic actor-oriented dynamic network analysis. *Journal of Mathematical Sociology*, 21, 149–172.

Soffer, S. & Vázquez, A. (2005). Clustering coefficient without degree correlations biases. *Physical Review E*, 71:057101.

Stanoevska-Slabeva, K., & Schmid, B. F. (2001). A typology of online communities and community supporting platforms. *Proceedings of the 34th Annual Hawaii International Conference on System Sciences, 00*(c), 10. IEEE Comput. Soc.

Stanoevska-Slabeva, K. (2002). Toward a community-oriented design of internet platforms. *Int J Electr Commerce, 6*(3), 71095

Straub, D., Boudreau, M., & Gefen, D. (2004). Validation guidelines for IS positivist research. *Communications of the Association for Information Systems, 13*, 380-427.

Travers, J., & Milgram, S. (1969). An Experimental Study of the Small World Problem. *Sociometry, 32*, 425–443.

Thackeray, R., Neiger, B. L., Smith, A. K., & Van Wagenen, S. B. (2012). Adoption and use of social media among public health departments. *BMC Public Health, 12*(1), 242.

Toivonen, R., Kumpula, J. M., Saramaki, J., Onnela, J.-P., Kertesz, J., & Kaski, K. (2007). The role of edge weights in social networks: modelling structure and dynamics. *Proceedings of SPIE, 6601*, 1–8

Toral, S. L., Mart ńez-Torres, M. R., & Barrero, F. (2010). Analysis of virtual communities supporting OSS projects using social network analysis. *Information and Software Technology, 52*, 296–303.

Trier, M. (2008). Towards dynamic visualization for understanding evolution of digital communication networks. *Information Systems Research, 19*, 335–350.

Varik, F. & Oostendorp, H. (2013). Enhancing Online Community Activity: Development and Validation of the CA Framework. *Journal of Computer Mediated Communication, 18*(4), p. 454-475.

Wasserman, S., & Faust, K. (1994). Social Network Analysis: Methods and Applications. *Social Networks, 8*, 825.

Wasko, M., & Faraj, S. (2005). Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *Mis Quarterly, 29*(1), 35-57.

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

Webster, J. & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review, *MIS Quarterly, 26*(2), 13-23.

Weerd, I. van de, & Brinkkemper, S. (2008). *Meta-Modeling for Situational Analysis and Design Methods* (pp. 38–58). Hersey: Idea Group publishing.

Wiggins, A., Howison, J., & Crowston, K. (2008). Social dynamics of FLOSS team communication across channels. *In Proceedings of the Fourth International Conference on Open Source Software (IFIP 2.13)*. Milan, Italy.

Yao, Y., Zhou, J., Han, L., Xu, F., & Lü, J. (2011). *Comparing linkage graph and activity graph of online social networks. Lecture Notes in Computer Science*, 6984 LNCS, 84–97.

Appendix A

NumberOfReplies

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	843	15.9	15.9	15.9
	2	758	14.3	14.3	30.1
	3	579	10.9	10.9	41.0
	4	524	9.9	9.9	50.9
	5	422	7.9	7.9	58.9
	6	367	6.9	6.9	65.8
	7	317	6.0	6.0	71.7
	8	236	4.4	4.4	76.2
	9	217	4.1	4.1	80.3
	10	159	3.0	3.0	83.3
	11	110	2.1	2.1	85.3
	12	101	1.9	1.9	87.2
	13	88	1.7	1.7	88.9
	14	82	1.5	1.5	90.4
	15	57	1.1	1.1	91.5
	16	55	1.0	1.0	92.5
	17	41	.8	.8	93.3
	18	49	.9	.9	94.2
	19	37	.7	.7	94.9
	20	38	.7	.7	95.7
	21	21	.4	.4	96.0
	22	26	.5	.5	96.5
	23	16	.3	.3	96.8
	24	23	.4	.4	97.3
	25	16	.3	.3	97.6
	26	11	.2	.2	97.8
	27	12	.2	.2	98.0
	28	5	.1	.1	98.1
	29	12	.2	.2	98.3
	30	1	.0	.0	98.3
	31	9	.2	.2	98.5
	32	5	.1	.1	98.6
	33	6	.1	.1	98.7
	34	6	.1	.1	98.8
	36	2	.0	.0	98.9
	38	3	.1	.1	98.9
	39	4	.1	.1	99.0
	40	5	.1	.1	99.1
	41	3	.1	.1	99.2
	42	2	.0	.0	99.2
	43	2	.0	.0	99.2
	44	1	.0	.0	99.2
	45	2	.0	.0	99.3
	46	3	.1	.1	99.3
	47	2	.0	.0	99.4
	48	3	.1	.1	99.4
	49	1	.0	.0	99.5
	50	2	.0	.0	99.5
	51	3	.1	.1	99.5
	52	1	.0	.0	99.6
	53	3	.1	.1	99.6
	54	1	.0	.0	99.6
	56	1	.0	.0	99.7
	57	1	.0	.0	99.7
	58	2	.0	.0	99.7
	60	1	.0	.0	99.7
	64	1	.0	.0	99.8
	70	1	.0	.0	99.8
	71	1	.0	.0	99.8
	75	1	.0	.0	99.8
	77	1	.0	.0	99.8
	81	1	.0	.0	99.8
	87	1	.0	.0	99.9
	89	1	.0	.0	99.9
	90	2	.0	.0	99.9
	91	1	.0	.0	99.9
	99	1	.0	.0	100.0
	133	1	.0	.0	100.0
	332	1	.0	.0	100.0
	Total	5311	100.0	100.0	

Appendix B

Sub category 101, Start-ups					
	Manually	**Method 1**	**Method 2**	**Method 3**	**Method 4**
Number of nodes	105	105	105	105	**105**
Number of edges	170	692	410	138	**162**
Average degree	1.619	6.59	3.905	1.314	**1.543**
Avg. weighted degree	1.943	11.029	5.505	1.4	**1.924**
Network diameter	10	7	11	20	**8**
Graph density	0.016	0.063	0.038	0.013	**0.015**
Sub category 301, Networking					
	Manually	**Method 1**	**Method 2**	**Method 3**	**Method 4**
Number of nodes	100	100	100	100	**100**
Number of edges	141	588	352	124	**137**
Average degree	1.41	5.88	3.52	1.24	**1.37**
Avg. weighted degree	1.62	9.62	4.81	1.32	**1.87**
Network diameter	10	5	8	14	**7**
Graph density	0.014	0.059	0.036	0.013	**0.014**
Sub category 303, Promotion					
	Manually	**Method 1**	**Method 2**	**Method 3**	**Method 4**
Number of nodes	99	99	99	99	**99**
Number of edges	144	550	329	121	**142**
Average degree	1.455	5.556	3.323	1.222	**1.434**
Avg. weighted degree	1.657	8.788	4.384	1.283	**1.808**
Network diameter	13	6	11	19	**12**
Graph density	0.015	0.057	0.034	0.012	**0.015**
Sub category 306, Internet, online marketing & sales					
	Manually	**Method 1**	**Method 2**	**Method 3**	**Method 4**
Number of nodes	91	91	91	91	**91**
Number of edges	141	480	301	129	**143**
Average degree	1.549	5.275	3.308	1.418	**1.571**
Avg. weighted degree	1.857	10.769	5.385	1.505	**2.374**
Network diameter	10	7	9	21	**10**
Graph density	0.017	0.059	0.037	0.016	**0.017**
Sub category 601, Hallo!					
	Manually	**Method 1**	**Method 2**	**Method 3**	**Method 4**
Number of nodes	70	70	70	70	**70**
Number of edges	119	384	239	107	**115**
Average degree	1.7	5.486	3.414	1.529	**1.643**
Avg. weighted degree	2.129	10.771	5.386	1.614	**2.286**
Network diameter	9	7	7	11	**9**

Graph density	0.025	0.08	0.049	0.022	**0.024**
	Sub category 605, Others				
	Manually	**Method 1**	**Method 2**	**Method 3**	**Method 4**
Number of nodes	80	80	80	80	**80**
Number of edges	118	428	257	104	**114**
Average degree	1.475	5.35	3.212	1.3	**1.425**
Avg. weighted degree	1.6	8.9	4.425	1.338	**1.85**
Network diameter	8	5	8	11	**8**
Graph density	0.019	0.068	0.041	0.016	**0.018**

Appendix C

Sub category 101, Start-ups				
Method	**Observed value**	**Significance**	**Average**	**Standard deviation**
Manual, method 1	0.7265	0.0002	-0.0002	0.0135
Manual, method 2	0.6580	0.0002	-0.0001	0.0134
Manual, method 3	0.4883	0.0002	-0.0004	0.0133
Manual, method 4	**0.8488**	**0.0002**	**0.0001**	**0.0140**
Sub category 301, Networking				
Method	**Observed value**	**Significance**	**Average**	**Standard deviation**
Manual, method 1	0.7664	0.0002	0.0001	0.0147
Manual, method 2	0.6979	0.0002	0.0002	0.0145
Manual, method 3	0.5448	0.0002	0.0000	0.0144
Manual, method 4	**0.8957**	**0.0002**	**0.0001**	0.0145
Sub category 303, Promotion				
Method	**Observed value**	**Significance**	**Average**	**Standard deviation**
Manual, method 1	0.8161	0.0002	0.0001	0.0149
Manual, method 2	0.7576	0.0002	0.0001	0.0143
Manual, method 3	0.5662	0.0002	-0.0001	0.0144
Manual, method 4	**0.9236**	**0.0002**	**-0.0001**	**0.0144**
Sub category 306, Internet, online marketing & sales				
Method	**Observed value**	**Significance**	**Average**	**Standard deviation**
Manual, method 1	0.8164	0.0002	0.0002	0.0161
Manual, method 2	0.7686	0.0002	0.0001	0.0157
Manual, method 3	0.6187	0.0002	0.0000	0.0158
Manual, method 4	**0.8867**	**0.0002**	**0.0001**	**0.0155**
Sub category 601, Hallo!				
Method	**Observed value**	**Significance**	**Average**	**Standard deviation**
Manual, method 1	0.8069	0.0002	-0.0006	0.0205
Manual, method 2	0.7415	0.0002	-0.0001	0.0207
Manual, method 3	0.6229	0.0002	0.0001	0.0201
Manual, method 4	**0.9299**	**0.0002**	**-0.0001**	**0.0206**
Sub category 605, Others				
Method	**Observed value**	**Significance**	**Average**	**Standard deviation**
Manual, method 1	0.7439	0.0002	-0.0001	0.0180
Manual, method 2	0.7026	0.0002	0.0002	0.0180
Manual, method 3	0.5374	0.0002	-0.0000	0.0179
Manual, method 4	**0.8789**	**0.0002**	**-0.0003**	**0.0178**

Appendix D

Category id	Number of replies	Occurrence	Relative percentage within sub category	Post duration in days
101	1	116	13.76%	8.5
	2	118	14.00%	7
	3	90	10.68%	10.6
	4	82	9.73%	13.6
	5	84	9.96%	40.7
	6	56	6.64%	24.3
	7	57	6.76%	26.8
	8	45	5.34%	17.2
	9	45	5.34%	32.3
	10	28	3.32%	11
	more_than_10	122	14.47%	58.3
104	1	29	13.49%	27.4
	2	26	12.09%	16.5
	3	25	11.63%	52
	4	29	13.49%	19.8
	5	15	6.98%	18.1
	6	11	5.12%	81.6
	7	14	6.51%	50.5
	8	10	4.65%	131
	9	6	2.79%	60.8
	10	8	3.72%	86.6
	more_than_10	42	19.53%	144.8
301	1	156	23.53%	23.6
	2	102	15.38%	23.7
	3	72	10.86%	28.9
	4	65	9.80%	34.5
	5	47	7.09%	42.3
	6	39	5.88%	30.5
	7	38	5.73%	60.8
	8	21	3.17%	51.1
	9	16	2.41%	9.4
	10	19	2.87%	52.1
	more_than_10	88	13.27%	96.4
302	1	33	11.83%	28.4
	2	33	11.83%	17.3
	3	31	11.11%	27.4
	4	29	10.39%	49.4

	5	26	9.32%	81.7
	6	20	7.17%	68.9
	7	13	4.66%	77.2
	8	13	4.66%	38.5
	9	20	7.17%	53.2
	10	11	3.94%	72.4
	more_than_10	50	17.92%	79.9
303	1	30	8.50%	25.9
	2	39	11.05%	14.7
	3	26	7.37%	18.3
	4	33	9.35%	36.6
	5	31	8.78%	23
	6	31	8.78%	19.2
	7	18	5.10%	16.2
	8	22	6.23%	54.7
	9	11	3.12%	11.2
	10	15	4.25%	75.8
	more_than_10	97	27.48%	97.4
304	1	29	15.43%	7.4
	2	26	13.83%	36.8
	3	16	8.51%	2.1
	4	16	8.51%	34.8
	5	11	5.85%	28.9
	6	12	6.38%	82.2
	7	13	6.91%	50.5
	8	15	7.98%	95.9
	9	8	4.26%	54.8
	10	5	2.66%	47
	more_than_10	37	19.68%	116.7
305	1	19	14.39%	48.2
	2	23	17.42%	5
	3	14	10.61%	4.6
	4	12	9.09%	23.1
	5	9	6.82%	6.8
	6	12	9.09%	30.8
	7	10	7.58%	40.8
	8	4	3.03%	17.2
	9	6	4.55%	82.8
	10	2	1.52%	266
	more_than_10	21	15.91%	97
306	1	84	10.85%	10.9
	2	84	10.85%	16.9
	3	87	11.24%	47.5
	4	67	8.66%	13.9

	5	59	7.62%	37.3
	6	59	7.62%	35.1
	7	52	6.72%	45
	8	45	5.81%	59.3
	9	42	5.43%	37.8
	10	28	3.62%	28.1
	more_than_10	167	21.58%	86.3
404	1	17	12.14%	14.7
	2	21	15.00%	1.1
	3	17	12.14%	7.5
	4	21	15.00%	39.5
	5	8	5.71%	46.4
	6	11	7.86%	13.8
	7	11	7.86%	17.3
	8	7	5.00%	81.4
	9	8	5.71%	6.4
	10	4	2.86%	1.8
	more_than_10	15	10.71%	31.1
501	1	26	10.79%	29.4
	2	27	11.20%	2.2
	3	28	11.62%	30.4
	4	27	11.20%	9.3
	5	28	11.62%	34.7
	6	22	9.13%	29.5
	7	17	7.05%	18.6
	8	9	3.73%	49.9
	9	9	3.73%	72.4
	10	10	4.15%	46.6
	more_than_10	38	15.77%	105.2
502	1	17	15.89%	2.9
	2	11	10.28%	1.2
	3	14	13.08%	3.4
	4	10	9.35%	20.4
	5	8	7.48%	67.1
	6	8	7.48%	7.4
	7	3	2.80%	9.3
	8	5	4.67%	2.4
	9	6	5.61%	12.8
	10	1	0.93%	5
	more_than_10	24	22.43%	90.7
503	1	26	13.98%	3.9
	2	32	17.20%	15.8
	3	22	11.83%	4.5
	4	20	10.75%	2.6

	5	20	10.75%	15.4
	6	11	5.91%	80.5
	7	16	8.60%	5.4
	8	4	2.15%	109.8
	9	4	2.15%	2.2
	10	2	1.08%	39
	more_than_10	29	15.59%	30.3
601	1	79	19.90%	11.6
	2	65	16.37%	4.3
	3	43	10.83%	12.9
	4	40	10.08%	8
	5	26	6.55%	7.4
	6	28	7.05%	17.1
	7	24	6.05%	9.4
	8	12	3.02%	54.7
	9	12	3.02%	28.5
	10	8	2.02%	39.1
	more_than_10	60	15.11%	53.8
604	1	77	21.57%	11.8
	2	76	21.29%	19.8
	3	48	13.45%	46.7
	4	35	9.80%	63.1
	5	29	8.12%	54.8
	6	20	5.60%	29.6
	7	11	3.08%	27.6
	8	14	3.92%	67.5
	9	6	1.68%	120.5
	10	4	1.12%	213.8
	more_than_10	37	10.36%	161
605	1	105	24.08%	14.9
	2	75	17.20%	19
	3	46	10.55%	30.8
	4	38	8.72%	51.8
	5	21	4.82%	28.9
	6	27	6.19%	22.5
	7	20	4.59%	74.2
	8	10	2.29%	4.8
	9	18	4.13%	50.2
	10	14	3.21%	33.1
	more_than_10	62	14.22%	55.8

Appendix E

cate_id	NOR	1-7 days	8-14 days	15-21 days	22-28 days	29-60 days	61-90 days	91-120 days	120+ days
101	1	105	2	0	2	2	2	0	3
	2	223	4	2	1	2	1	0	3
	3	247	3	4	4	5	1	2	4
	4	313	6	1	0	0	0	3	5
	5	352	6	8	7	8	11	4	24
	6	288	12	3	0	1	7	10	15
	7	365	17	4	0	3	1	1	8
	8	315	19	9	5	4	2	0	6
	9	345	23	2	1	10	6	9	9
	10	241	14	10	2	4	9	0	0
	10+	1733	147	67	39	93	42	66	406
	Total	4527	253	110	61	132	82	95	483
104	1	23	1	1	0	0	0	1	3
	2	40	1	2	1	3	3	0	2
	3	61	2	0	0	4	3	0	5
	4	101	5	3	0	2	2	1	2
	5	65	2	1	0	3	2	0	2
	6	51	1	1	3	5	0	0	5
	7	74	6	1	0	6	5	0	6
	8	57	1	3	0	6	3	0	10
	9	42	6	4	0	0	0	0	2
	10	56	8	2	2	6	1	3	2
	10+	450	38	21	15	44	71	11	82
	total	1020	71	39	21	79	90	16	121
301	1	113	7	10	4	7	3	2	10
	2	155	11	4	3	15	7	0	9
	3	172	3	4	4	17	3	1	12
	4	216	12	6	1	3	6	1	15
	5	194	6	5	7	5	3	4	11
	6	183	7	14	5	5	1	1	18
	7	201	24	14	9	6	1	0	11
	8	115	15	7	0	13	8	1	9
	9	130	4	1	1	7	1	0	0
	10	129	14	9	17	11	0	0	10
	10+	1030	144	95	54	129	47	49	215
	total	2638	247	169	105	218	80	59	320
302	1	28	0	1	1	1	0	0	2
	2	53	5	0	1	2	0	2	3

	3	80	4	2	0	1	1	1	4
	4	97	1	1	6	1	0	1	9
	5	92	1	3	1	18	2	0	13
	6	101	4	2	3	6	0	0	4
	7	83	2	1	0	1	0	0	4
	8	89	7	0	0	6	0	0	2
	9	167	4	0	0	1	0	1	7
	10	91	9	1	0	4	0	0	5
	10+	557	35	58	34	28	23	4	101
	total	1438	72	69	46	69	26	9	154
303	1	22	3	0	1	0	1	0	3
	2	66	2	2	0	4	2	0	2
	3	71	0	1	0	2	3	0	1
	4	104	7	3	0	5	2	0	11
	5	129	8	7	7	1	0	0	3
	6	167	9	1	0	4	1	0	4
	7	102	1	4	3	13	3	0	0
	8	145	4	4	1	11	0	0	11
	9	83	1	2	4	4	5	0	0
	10	119	7	2	0	4	3	0	15
	10+	1216	127	98	65	116	60	43	255
	total	2224	169	124	81	164	80	43	305
304	1	26	0	1	0	0	0	1	1
	2	44	1	0	0	1	1	1	4
	3	44	3	0	0	1	0	0	0
	4	53	1	0	3	1	0	0	6
	5	44	1	2	2	3	2	0	1
	6	58	7	0	0	1	0	0	6
	7	75	3	0	5	4	0	0	4
	8	97	4	5	3	6	0	0	5
	9	53	4	1	0	4	3	4	3
	10	44	1	3	0	0	0	1	1
	10+	369	63	19	7	70	11	4	100
	total	907	88	31	20	91	17	11	131
305	1	13	1	1	0	0	1	1	2
	2	44	0	0	0	0	0	2	0
	3	38	2	1	1	0	0	0	0
	4	39	0	1	0	0	0	3	5
	5	40	1	2	2	0	0	0	0
	6	58	5	0	0	2	0	1	6
	7	56	5	0	4	2	0	1	2
	8	24	5	0	2	1	0	0	0
	9	45	0	0	2	1	2	0	4
	10	5	0	0	0	0	0	4	11

	10+	248	23	15	2	4	8	7	23
	total	610	42	20	13	10	11	19	53
306	1	68	4	1	1	1	1	0	8
	2	144	9	1	1	3	1	2	7
	3	218	3	5	3	7	0	1	24
	4	235	8	5	1	3	0	1	15
	5	251	4	9	9	4	5	4	9
	6	291	22	10	3	10	3	2	13
	7	309	18	3	5	4	3	1	21
	8	309	12	1	1	10	2	1	24
	9	338	21	2	6	2	1	0	8
	10	242	2	18	0	10	1	0	7
	10+	2206	222	122	86	140	104	58	256
	total	4611	325	177	116	194	121	70	392
404	1	16	0	0	0	0	0	0	1
	2	39	3	0	0	0	0	0	0
	3	50	0	0	0	0	0	1	0
	4	76	0	2	0	0	0	1	5
	5	34	1	0	0	0	1	0	4
	6	62	0	0	0	0	0	0	4
	7	73	2	0	0	1	0	0	1
	8	41	6	5	0	0	0	0	4
	9	62	8	2	0	0	0	0	0
	10	40	0	0	0	0	0	0	0
	10+	328	30	20	7	37	4	0	7
	total	821	50	29	7	38	5	2	26
501	1	20	1	0	2	1	0	0	2
	2	49	1	3	0	0	1	0	0
	3	68	0	1	3	5	1	2	4
	4	96	5	2	1	2	0	2	0
	5	120	9	4	1	0	1	1	4
	6	101	10	2	7	7	3	0	2
	7	98	10	0	0	7	4	0	0
	8	60	5	0	1	0	0	0	6
	9	55	3	2	3	0	14	1	3
	10	72	3	1	2	6	4	1	11
	10+	416	63	38	15	24	15	7	87
	total	1155	110	53	35	52	43	14	119
502	1	15	1	0	0	1	0	0	0
	2	22	0	0	0	0	0	0	0
	3	38	1	2	1	0	0	0	0
	4	39	0	0	0	0	0	0	1
	5	36	1	0	0	2	0	0	1
	6	45	0	0	0	3	0	0	0

	7	15	1	2	3	0	0	0	0
	8	40	0	0	0	0	0	0	0
	9	44	4	3	1	2	0	0	0
	10	10	0	0	0	0	0	0	0
	10+	334	34	8	10	26	18	3	37
	total	638	42	15	15	34	18	3	39
503	1	24	0	0	0	1	1	0	0
	2	57	0	0	2	2	0	1	2
	3	60	3	0	0	3	0	0	0
	4	77	3	0	0	0	0	0	0
	5	83	1	0	1	14	1	0	0
	6	47	7	3	0	7	0	0	2
	7	105	2	1	1	3	0	0	0
	8	30	1	0	0	0	0	0	1
	9	36	0	0	0	0	0	0	0
	10	6	3	0	2	2	7	0	0
	10+	355	34	25	14	25	9	18	27
	total	880	54	29	20	57	18	19	32
601	1	71	0	1	1	4	0	0	2
	2	113	6	3	5	2	1	0	0
	3	109	8	1	2	1	1	0	7
	4	144	5	6	0	1	0	0	4
	5	117	2	5	0	5	1	0	0
	6	155	1	3	1	5	0	0	3
	7	140	9	7	4	4	0	0	4
	8	89	0	1	0	0	0	0	6
	9	88	1	1	0	4	4	8	2
	10	73	5	0	0	1	0	0	1
	10+	932	106	21	42	71	6	21	81
	total	2031	143	49	55	98	13	29	110
604	1	64	1	0	1	4	4	0	3
	2	117	7	7	0	11	2	3	5
	3	116	5	1	1	6	3	1	11
	4	97	6	1	9	5	6	7	9
	5	101	18	3	1	6	8	0	8
	6	102	6	0	5	2	2	1	2
	7	61	3	0	1	9	1	0	2
	8	75	1	11	1	14	1	0	9
	9	34	0	1	1	1	3	2	12
	10	22	0	0	0	7	0	0	11
	10+	527	91	14	3	54	18	21	144
	total	1316	138	38	23	119	48	35	216
605	1	89	3	3	1	2	3	0	4
	2	124	5	1	5	6	2	1	6

	3	115	2	0	2	3	4	5	7
	4	123	4	2	2	6	0	0	15
	5	86	4	1	0	12	1	0	1
	6	126	15	1	2	6	0	3	9
	7	97	8	4	2	9	8	2	10
	8	75	0	4	1	0	0	0	0
	9	124	7	8	0	6	4	1	12
	10	119	4	0	3	7	3	1	3
	10+	877	75	12	20	46	30	13	68
	total	1955	127	36	38	103	55	26	135

Appendix F Python, SNA implementation

1) Choice of programming language

Python was chosen as the main programming language to develop/implement statistical analysis of social network graphs for a number of reasons:

a) python is a cross-platform, easy to learn, easy to read language with support for both functional as well as object-oriented language constructs, allowing easy interfacing with third-party libraries which might be constructed in any of these styles

b) python is well-known in the scientific community as a go-to language for developing solutions to various problems, including Big Data and statistics programming problems, increasing the likelihood that a scientifically-robust library might already be at hand for problems encountered

c) python has well-documented interfaces to statistical programs/libraries like igraph, matlab, SPSS and even the R programming language.

d) Python is well-known for its ease and consistency in importing/exporting data in different formats, as well for its flexibility in selecting and transforming data between different formats, and its ability to handle/clean non-standard data sets.

2) Dependencies

The sna.py program makes heavy use of the functionality provided by the NetworkX software package, developed at Los Alamos National Laboratory by the Applied Mathematics and Plasma Physics group (Hagberg, Schult and Swart, 2008). NetworkX is a python language package for the exploration and analysis of networks and network algorithms, with one of its main goals being the ability to painlessly slurp in large non-standard data sets.

It is well-tested, with over 1800 unit tests covering more than 90% of the code, and released as free and open source software under the terms of the BSD license, with Aric Hagberg (hagberg@lanl.gov), Dan Schult (dschult@lanl.gov) and Pieter Swart (swart@lanl.gov) as copyright holders. The source code for the implementation of a specific algorithm is to be found at the following url: http://networkx.github.io/documentation/latest/reference/algorithms.html. Links on this page will lead to detailed algorithm/function call discussion; in general, the source code for the algorithm implementation can be found through a link [source code] next to the function call description on these pages.

The reader is referred to this url, and the relevant source code found through the links on this url, for a more in-depth look at the precise algorithm implementation. For now, only the algorithm implementations developed within the context of this thesis, i.e. those not readily available in the NetworkX package, will be discussed more in detail, further down.

3) Definitions

The following definitions for structures referenced in code are, where not explicitly mentioned, derived from definitions found in the NetworkX documentation. Where algorithms are implemented directly in python, reference to the source algorithm is provided, as well as the means of verification for a correct implementation of the referred to algorithm, typically through a third-party package like Gephi or the R programming language. Where necessary or deemed appropriate, a direct reference to the scientific literature describing the implemented algorithm is given.

1. Graph structure

A graph (network) is a collection of nodes together with a collection of edges that are pairs of nodes.

2. Weighted graph

A graph in which the edges differ in weight (aka importance).

3. Directed edge

An edge (pair relationship) in which the order of the edge pair matters i.e. the relation (u,v) differs from the relation (v,u). Many classical graph properties are defined differently for directed graphs.

4. Graph

Graphs hold undirected edges. Self loops between a node and itself are allowed but multiple (parallel) edges are not. Edges are represented as links between nodes with optional key/value attributes.

5. Directed Graph ('DiGraph')

DiGraphs hold directed edges. Self loops are allowed but multiple (parallel) edges are not. Edges are represented as links between nodes with optional key/value attributes.

6. MultiGraph

A MultiGraph holds undirected edges. Self loops are allowed. MultiGraphs can hold multi-edges, i.e. multiple edges between two nodes. Each edge can hold optional data or attributes. Edges are represented as links between nodes with optional key/value attributes.

7. Directed MultiGraph ('MultiDiGraph')

A MultiDiGraph holds directed edges. Self loops are allowed. MultiDiGraphs can hold multi-edges, i.e. multiple edges between two nodes. Each edge can hold optional data or attributes. Edges are represented as links between nodes with optional key/value attributes.

8. Density

The density for undirected graphs is

$$d=2m/(n(n-1))$$

and for directed graphs is

$$d=m/(n(n-1)),$$

where n is the number of nodes and m is the number of edges in the graph structure.

The density is 0 for a graph without edges and 1 for a complete graph. The density of multigraphs can be higher

than 1. Self loops are counted in the total number of edges so graphs with self loops can have density higher than 1.

9. Average clustering coefficient

Estimates the average clustering coefficient of a graph structure G. The local clustering of each node in G is the fraction of triangles that actually exist over all possible triangles in its neighborhood. The average clustering coefficient of a graph G is the mean of local clustering's.

The calculation is executed by repeating n times (defined in *trials*, an optional parameter) the following experiment: choose a node at random, choose two of its neighbors at random, and check if they are connected. The approximate coefficient is the fraction of triangles found over the number of trial. For the purposes of this research, trials was held at its default value of 1000.

10. Diameter

The diameter of a graph structure is the maximum eccentricity found for all nodes within the structure. The eccentricity of a node is the maximum distance from this node to all other nodes in the same graph structure.

11. Average shortest path length

The average shortest path length is the sum of all divisions of the shortest path from source node s to target node t, for all possible values of s and t, by a factor n(n - 1), with n being the number of nodes in the graph structure. Mathematically:

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

with V being the set of nodes in the graph, d(s,t) the shortest path between nodes s and t, and n as the total number of nodes in the graph. In case of disconnected graphs, as with a subject split into several different independent threads, the average shortest path length was calculated for each component subgraph, with the result for each component subgraph subsequently summed and averaged over all subgraphs. An independent thread, in this context, is a thread where nodes (actors, users) involved do not appear in other discussion threads within the same subject category.

12. Average degree

The calculation for average degree ignores edge weight/node strength completely, and simply looks at the number of unique connections to/from a node i.e. if a node A has 2 connections to other nodes B and C, and A → B appears 5 times in the list of edges for this graph while A → C only appears a single time in the graph, the average degree for A is 2.

13. Average weighted degree

The calculation for average weighted degree takes node strength into account. As an example, given the graph with edges:

A -> B

A -> B
A -> B
B -> A
C -> A

edge A → B weighs (node strength) 3, while edges B → A and C → A each weigh 1, bringing the total weight of A to 5, while B and C each weigh 1, so the average weighted degree would be (5+4+1)/3 = 3.3333333 for this graph.

14. Reciprocity

Reciprocity is a measure for the tendency to return a tie in a network/graph.It can be studied/determined at the level of a dyad: a representation of a pair of nodes and the possible relational ties between them. For the purpose of this thesis, dyadic reciprocity was calculated, per node and per category+time interval. The dyadic reciprocity is the proportion of dyads which are symmetric. In more practical terms, dyadic reciprocity is the number of reciprocated dyads divided by the number of adjacent dyads.

In a network with nodes A, B, C, D and E, and the following edges:
B → A
C → A
B → C
C → B
D → B
C → D
D → E
E → D

there are 2 reciprocated dyads (B → C, C → B and D → E, E → D) and 6 adjacent dyads, hence the reciprocity can be calculated as 2/6, or 33%. Implementation of the algorithm is based on the description given in the manual for the grecip function in the sna package, R statistics programming language, and the discussion of dyadic and triadic reciprocity in:

Unit 4: Dyads and Triads, Reciprocity and Transitivity of the ICPSR summer program in Quantitative Methods of Social Research at the University of Michigan, Ann Arbor, Summer 215 (Ann McCranie), source url: http://annmccranie.net/site/ICPSR.html.

15. Degree Centrality

The degree centrality for a node is the fraction of all graph nodes this node is connected to.
For multigraphs or graphs with self loops the maximum degree might be higher than n-1, with n being the total number of nodes in the graph, and values of degree centrality greater than 1 are possible.

16. Betweenness Centrality

Betweenness centrality of a node v is the sum of the fraction of all-pairs shortest paths that pass through this node v:

$$c(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

for all pairs (s, t) belonging to the graph structure's nodes collection:

where $\sigma(s,t)$ is the number of shortest (s,t)-paths, and $\sigma(s,t|v)$ is the number of those paths passing through some node v other than s,t. If s=t, $\sigma(s,t)=1$, and if $v \in s,t$, $\sigma(s,t|v)=0$ (Brandes, 2001; Brandes and Pich, 2007; Brandes, 2008).

17. Closeness Centrality

Closeness centrality of a node is the reciprocal of the sum of the shortest path distances from this node to all $n-1$ other nodes [n being the total number of nodes]. Since the sum of distances depends on the number of nodes in the graph, closeness is normalized by the sum of minimum possible distances $n-1$ (Freeman, 1979). Or, mathematically,

with u = node for which the closeness centrality is calculated, v any other node in the same Graph, d(v, u) the shortest path distance between node u and node v, and n the total number of nodes in the Graph. Take note that when the graph isn't completely connected, the closeness centrality is calculated for each connected part [component subgraph] separately.

18. Eigenvector Centrality

The calculation for eigenvector centrality is based on assigning relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes.

For directed graphs, like in this research, the calculated eigenvector centrality is "right" eigenvector centrality. The eigenvector calculation method is done by power iteration and has no guarantee of convergence – to be more precise, in 19 cases [edgelist selection based on category and time interval], no convergence was reached with the default settings.

Appendix G

Edge-Ratio analysis on 60 days time interval												
Sub-category 101	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10	P_11	P_12
Number of new users	74	48	63	93	202	250	257	173	155	214	231	94
Number of old users		16	13	23	45	60	76	78	77	94	101	72
New user's %		75.00%	82.89%	80.17%	81.78%	80.65%	77.18%	68.92%	66.81%	69.48%	69.58%	56.63%
Edges between new users %		26.29%	38.51%	34.08%	30.16%	33.82%	26.26%	19.98%	12.75%	15.34%	17.97%	13.21%
Edges between old users %		37.40%	11.63%	31.85%	16.56%	14.00%	21.88%	24.45%	41.83%	32.80%	28.02%	33.83%
Edges between old and new users %		36.31%	49.86%	34.08%	53.28%	52.18%	51.86%	55.57%	45.42%	51.86%	54.02%	52.96%
Impact of new users		35.05%	46.46%	42.50%	36.88%	41.94%	34.02%	28.99%	19.08%	22.07%	25.82%	23.33%
Post/user	1.47	2.05	2.28	2.00	2.89	2.52	2.36	2.48	2.66	2.85	2.54	2.31
Sub-category 104	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10	P_11	P_12
Number of new users	41	63	58	48	69	55	46	28	43	36	25	23
Number of old users		12	14	24	25	25	22	26	27	32	22	21
New user's %		84.00%	80.56%	66.67%	73.40%	68.75%	67.65%	51.85%	61.43%	52.94%	53.19%	52.27%
Edges between new users %		59.68%	14.94%	34.39%	29.26%	51.40%	42.44%	10.50%	22.52%	3.93%	15.63%	15.91%
Edges between old users %		3.56%	45.18%	27.38%	18.52%	6.93%	18.60%	48.47%	24.32%	77.49%	27.05%	39.39%
Edges between old and new users %		36.76%	39.89%	38.24%	52.21%	41.68%	38.95%	41.03%	53.15%	18.58%	57.32%	44.70%
Impact of new users		71.05%	18.54%	51.58%	39.86%	74.76%	62.74%	20.24%	36.66%	7.43%	29.39%	30.43%
Post/user	1.66	1.95	1.63	1.93	2.17	1.81	1.85	2.19	2.10	2.22	1.96	1.93
Sub-category 301	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10	P_11	P_12
Number of new users	107	83	69	58	178	167	140	58	97	77	144	25
Number of old users		28	22	41	62	66	80	59	56	67	83	27
New user's %		74.77%	75.82%	58.59%	74.17%	71.67%	63.64%	49.57%	63.40%	53.47%	63.44%	48.08%
Edges between new users %		33.04%	29.01%	47.48%	23.36%	28.78%	18.01%	10.46%	18.30%	14.09%	31.60%	21.17%
Edges between old users %		16.85%	21.53%	14.32%	22.23%	24.94%	35.19%	59.83%	33.23%	36.76%	24.45%	23.36%
Edges between old and new users %		50.11%	49.47%	38.20%	54.41%	46.28%	46.80%	29.71%	48.46%	49.15%	43.96%	55.47%
Impact of new users		44.18%	38.26%	81.04%	31.50%	40.15%	28.30%	21.10%	28.87%	26.35%	49.81%	44.03%
Post/user	1.89	2.63	1.82	1.87	2.42	2.39	2.99	2.14	2.78	2.27	1.94	1.42
Sub-category 302	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10	P_11	P_12
Number of new users	73	48	69	37	94	58	69	52	47	56	56	25
Number of old users		13	18	23	32	23	29	32	48	32	34	25
New user's %		78.69%	79.31%	61.67%	74.60%	71.60%	70.41%	61.90%	49.47%	63.64%	62.22%	50.00%
Edges between new users %		16.12%	48.65%	15.52%	53.03%	35.84%	32.42%	26.58%	18.54%	33.33%	27.95%	7.33%
Edges between old users %		37.91%	11.35%	34.83%	11.31%	10.62%	15.66%	27.45%	36.90%	14.83%	15.64%	61.33%
Edges between old and new users %		45.97%	40.00%	49.66%	35.66%	53.54%	51.91%	45.96%	44.56%	51.84%	56.41%	31.33%
Impact of new users		20.49%	61.34%	25.16%	71.09%	50.05%	46.05%	42.94%	37.47%	52.38%	44.93%	14.67%
Post/user	1.66	2.21	1.80	2.65	2.10	1.86	2.18	2.05	2.07	1.88	1.84	1.72

Sub-category 303	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10	P_11	P_12
Number of new users	97	83	86	86	93	90	109	107	63	76	64	36
Number of old users		20	28	38	35	37	60	64	44	56	62	43
New user's %		80.58%	75.44%	69.35%	72.66%	70.87%	64.50%	62.57%	58.88%	57.58%	50.79%	45.57%
Edges between new users %		18.69%	22.46%	13.42%	27.70%	42.02%	35.37%	20.27%	14.85%	6.32%	5.49%	10.94%
Edges between old users %		29.71%	36.59%	57.15%	26.37%	18.14%	21.69%	33.32%	41.67%	69.48%	70.29%	36.72%
Edges between old and new users %		51.60%	40.95%	29.43%	45.93%	39.84%	42.94%	46.41%	43.48%	24.20%	24.22%	52.34%
Impact of new users		23.19%	29.78%	19.35%	38.13%	59.29%	54.84%	32.40%	25.21%	10.97%	10.81%	24.00%
Post/user	1.68	2.09	2.14	2.24	1.99	2.00	2.11	2.20	2.41	3.08	2.08	1.89
Sub-category 304	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10	P_11	P_12
Number of new users	51	40	57	48	65	85	46	24	31	58	45	12
Number of old users		8	10	20	28	25	17	18	18	20	35	16
New user's %		83.33%	85.07%	70.59%	69.89%	77.27%	73.02%	57.14%	63.27%	74.36%	56.25%	42.86%
Edges between new users %		46.59%	74.70%	32.65%	55.78%	49.60%	36.31%	14.89%	18.63%	50.05%	12.88%	4.20%
Edges between old users %		11.36%	0.89%	12.08%	8.18%	4.29%	14.88%	43.26%	29.50%	6.94%	29.61%	43.89%
Edges between old and new users %		42.05%	24.40%	55.27%	36.05%	46.11%	48.81%	41.84%	51.86%	43.01%	57.51%	51.91%
Impact of new users		55.91%	87.81%	46.25%	79.80%	64.19%	49.73%	26.06%	29.45%	67.31%	22.89%	9.80%
Post/user	1.63	1.48	1.97	1.99	2.11	1.44	1.41	1.50	1.65	1.94	1.61	2.04
Sub-category 305	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10	P_11	P_12
Number of new users	34	38	34	27	18	32	32	11	19	24	59	4
Number of old users		8	6	6	10	5	23	11	16	19	20	4
New user's %		82.61%	85.00%	81.82%	64.29%	86.49%	58.18%	50.00%	54.29%	55.81%	74.68%	50.00%
Edges between new users %		65.85%	70.37%	53.89%	21.74%	51.80%	36.18%	21.62%	9.30%	23.70%	34.39%	7.14%
Edges between old users %		1.22%	0.00%	11.98%	14.01%	5.76%	12.35%	13.51%	81.86%	30.37%	14.34%	78.57%
Edges between old and new users %		32.93%	29.63%	34.13%	64.25%	42.45%	51.47%	64.86%	8.84%	45.93%	51.28%	14.29%
Impact of new users		79.72%	82.79%	65.87%	33.82%	59.89%	62.18%	43.24%	17.14%	42.47%	46.04%	14.29%
Post/user	1.38	1.70	1.43	1.97	2.25	1.81	1.89	1.82	1.77	1.91	1.94	1.38
Sub-category 306	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10	P_11	P_12
Number of new users	180	107	125	117	115	146	117	91	67	117	120	42
Number of old users		36	52	63	56	69	71	69	53	69	73	48
New user's %		74.83%	70.62%	65.00%	67.25%	67.91%	62.23%	56.88%	55.83%	62.90%	62.18%	46.67%
Edges between new users %		28.06%	30.99%	27.21%	28.46%	31.27%	36.38%	11.10%	6.81%	10.72%	13.19%	3.59%
Edges between old users %		18.18%	21.47%	30.93%	16.99%	22.96%	15.89%	50.27%	54.24%	40.26%	33.35%	72.91%
Edges between old and new users %		53.75%	47.54%	41.85%	54.55%	45.77%	47.73%	38.64%	38.95%	49.02%	53.47%	23.51%
Impact of new users		37.51%	43.88%	41.87%	42.32%	46.05%	58.45%	19.51%	12.19%	17.04%	21.21%	7.68%
Post/user	2.97	3.16	3.26	2.94	2.38	2.69	2.81	2.71	3.35	4.26	4.23	3.23
Sub-category 404	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10	P_11	P_12
Number of new users	8	1	25	29	21	22	59	43	31	44	35	17
Number of old users		1	2	8	14	10	13	17	30	29	16	9
New user's %		50.00%	92.59%	78.38%	60.00%	68.75%	81.94%	71.67%	50.82%	60.27%	68.63%	65.38%

Edges between new users %		0.00%	67.78%	54.84%	6.73%	56.76%	45.38%	41.49%	10.63%	19.27%	30.05%	36.21%
Edges between old users %		0.00%	0.00%	3.10%	51.59%	1.35%	7.63%	8.10%	37.32%	23.44%	13.66%	0.00%
Edges between old and new users %		100.00%	32.22%	42.05%	41.68%	41.89%	46.99%	50.41%	52.05%	57.29%	56.28%	63.79%
Impact of new users		0.00%	73.20%	69.97%	11.21%	82.56%	55.38%	57.89%	20.92%	31.97%	43.79%	55.38%
Post/user	1.50	1.00	2.96	2.49	2.57	1.66	1.82	2.40	3.21	2.01	1.75	2.04
Sub-category 501	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10	P_11	P_12
Number of new users	40	36	56	50	90	50	50	42	43	43	80	16
Number of old users		9	11	20	37	30	24	17	31	24	30	13
New user's %		80.00%	83.58%	71.43%	70.87%	62.50%	67.57%	71.19%	58.11%	64.18%	72.73%	55.17%
Edges between new users %		64.44%	65.76%	18.91%	22.26%	19.29%	9.70%	25.93%	31.45%	16.67%	49.30%	9.38%
Edges between old users %		2.22%	1.09%	30.70%	28.56%	24.44%	37.15%	19.91%	15.72%	29.44%	6.40%	25.00%
Edges between old and new users %		33.33%	33.15%	50.39%	49.18%	56.27%	53.15%	54.17%	52.83%	53.89%	44.31%	65.63%
Impact of new users		80.56%	78.67%	26.48%	31.41%	30.87%	14.36%	36.42%	54.12%	25.97%	67.78%	16.99%
Post/user	1.53	1.36	2.54	1.66	2.64	2.01	2.70	1.71	1.65	2.15	1.64	1.45
Sub-category 502	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10	P_11	P_12
Number of new users	21	49	31	37	14	15	45	19	22	68	50	10
Number of old users		1	9	18	9	8	10	8	12	31	25	11
New user's %		98.00%	77.50%	67.27%	60.87%	65.22%	81.82%	70.37%	64.71%	68.69%	66.67%	47.62%
Edges between new users %		92.36%	17.90%	41.70%	20.83%	31.68%	34.26%	27.78%	22.77%	36.29%	46.48%	12.36%
Edges between old users %		0.00%	43.28%	8.86%	0.00%	11.88%	12.87%	0.00%	24.88%	14.01%	8.54%	19.10%
Edges between old and new users %		7.64%	38.83%	49.45%	79.17%	56.44%	52.87%	72.22%	52.35%	49.71%	44.97%	68.54%
Impact of new users		94.25%	23.09%	61.98%	34.23%	48.58%	41.87%	39.47%	35.19%	52.83%	69.72%	25.96%
Post/user	1.10	1.42	1.68	1.71	1.43	1.39	1.73	1.52	1.97	2.02	1.48	1.86
Sub-category 503	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10	P_11	P_12
Number of new users	35	21	17	17	36	58	51	22	35	42	35	18
Number of old users		5	5	10	13	20	26	15	20	25	25	17
New user's %		80.77%	77.27%	62.96%	73.47%	74.36%	66.23%	59.46%	63.64%	62.69%	58.33%	51.43%
Edges between new users %		9.13%	66.50%	35.14%	29.36%	18.92%	23.94%	7.01%	41.46%	18.69%	20.24%	15.67%
Edges between old users %		61.54%	2.81%	10.81%	17.89%	26.35%	25.87%	28.57%	18.13%	32.71%	24.29%	23.50%
Edges between old and new users %		29.33%	30.69%	54.05%	52.75%	54.73%	50.19%	64.42%	40.42%	48.60%	55.47%	60.83%
Impact of new users		11.31%	86.05%	55.80%	39.96%	25.44%	36.14%	11.79%	65.15%	29.82%	34.70%	30.47%
Post/user	1.54	1.81	3.36	1.67	2.41	2.29	1.94	2.84	1.93	2.36	2.20	1.91
Sub-category 601	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10	P_11	P_12
Number of new users	84	31	28	28	74	60	71	50	31	62	73	55
Number of old users		20	18	33	31	35	41	38	35	38	45	27
New user's %		60.78%	60.87%	45.90%	70.48%	63.16%	63.39%	56.82%	46.97%	62.00%	61.86%	67.07%
Edges between new users %		10.76%	6.75%	10.51%	21.76%	32.97%	21.43%	18.34%	1.22%	16.12%	56.67%	1.01%
Edges between old users %		62.25%	58.35%	47.46%	35.00%	22.65%	40.56%	42.65%	88.65%	33.61%	13.91%	82.71%
Edges between old and new users %		26.99%	34.90%	42.03%	43.24%	44.38%	38.02%	39.01%	10.13%	50.26%	29.42%	16.28%
Impact of new users		17.70%	11.10%	22.89%	30.88%	52.21%	33.80%	32.28%	2.60%	26.00%	91.60%	1.51%

Post/user	2.68	4.55	4.22	3.23	2.32	2.53	2.85	2.65	3.15	2.89	2.05	2.16
Sub-category 604	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10	P_11	P_12
Number of new users	91	124	79	86	69	63	77	22	49	98	69	12
Number of old users		16	32	36	20	20	26	12	28	40	39	8
New user's %		88.57%	71.17%	70.49%	77.53%	75.90%	74.76%	64.71%	63.64%	71.01%	63.89%	60.00%
Edges between new users %		61.24%	40.58%	58.14%	43.43%	42.42%	25.24%	46.94%	25.89%	13.78%	18.40%	38.41%
Edges between old users %		4.96%	19.54%	7.79%	13.71%	9.85%	20.44%	10.20%	19.05%	36.40%	33.31%	15.94%
Edges between old and new users %		33.80%	39.87%	34.07%	42.86%	47.73%	54.32%	42.86%	55.06%	49.82%	48.28%	45.65%
Impact of new users		69.15%	57.02%	82.48%	56.02%	55.89%	33.76%	72.54%	40.69%	19.40%	28.81%	64.01%
Post/user	1.35	2.06	1.77	2.03	1.37	1.54	2.00	1.65	1.60	2.20	1.81	1.70
Sub-category 605	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10	P_11	P_12
Number of new users	71	89	41	33	106	90	63	79	48	70	99	35
Number of old users		16	15	19	32	48	55	43	39	53	58	48
New user's %		84.76%	73.21%	63.46%	76.81%	65.22%	53.39%	64.75%	55.17%	56.91%	63.06%	42.17%
Edges between new users %		29.35%	67.14%	27.13%	34.55%	22.76%	13.24%	17.91%	22.07%	12.33%	23.18%	6.02%
Edges between old users %		14.24%	2.55%	36.18%	27.72%	30.74%	46.20%	30.13%	21.79%	34.72%	27.68%	50.68%
Edges between old and new users %		56.41%	30.31%	36.69%	37.73%	46.49%	40.56%	51.96%	56.15%	52.95%	49.15%	43.30%
Impact of new users		34.63%	91.70%	42.75%	44.98%	34.90%	24.80%	27.66%	40.00%	21.66%	36.75%	14.27%
Post/user	1.48	2.90	1.98	1.50	2.22	1.95	2.17	2.32	1.83	2.64	2.40	2.06

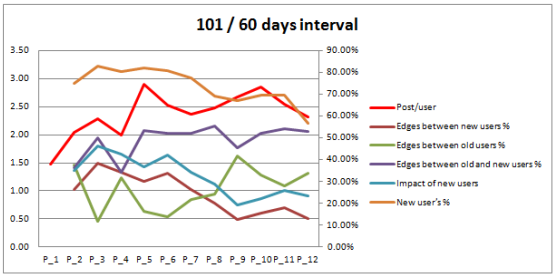
Edge-Ratio analysis on 90 days time interval								
Sub-category 101	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Number of new users	103	82	235	310	337	248	345	194
Number of old users		18	29	72	86	88	110	103
New user's %		82.00%	89.02%	81.15%	79.67%	73.81%	75.82%	65.32%
Edges between new users %		58.48%	49.79%	34.75%	26.35%	14.68%	17.42%	17.86%
Edges between old users %		3.98%	13.55%	15.50%	21.81%	35.63%	30.85%	29.83%
Edges between old and new users %		37.54%	36.66%	49.76%	51.84%	49.69%	51.73%	52.31%
Impact of new users		71.32%	55.93%	42.81%	33.07%	19.89%	22.97%	27.35%
post/user	1.75	2.33	2.71	2.65	2.39	3.03	2.95	2.57
Sub-category 104	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Number of new users	90	72	73	99	57	60	53	31
Number of old users		27	28	40	29	32	34	24
New user's %		72.73%	72.28%	71.22%	66.28%	65.22%	60.92%	56.36%
Edges between new users %		16.54%	40.40%	38.66%	45.95%	21.80%	8.07%	15.21%
Edges between old users %		44.09%	19.48%	16.91%	11.79%	21.68%	66.28%	46.13%
Edges between old and new users %		39.37%	40.11%	44.44%	42.26%	56.52%	25.64%	38.66%
Impact of new users		22.74%	55.90%	54.27%	69.32%	33.43%	13.25%	26.98%
post/user	1.79	1.72	1.95	2.09	2.01	2.37	2.39	2.18
Sub-category 301	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8

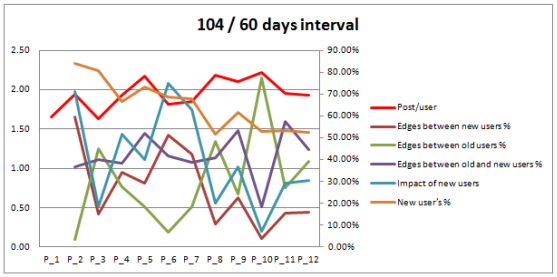
Number of new users	158	101	172	231	178	117	169	77
Number of old users		28	53	74	91	67	87	66
New user's %		78.29%	76.44%	75.74%	66.17%	63.59%	66.02%	53.85%
Edges between new users %		33.37%	39.62%	25.66%	18.22%	22.86%	20.91%	4.86%
Edges between old users %		20.67%	17.87%	23.07%	35.05%	28.45%	29.78%	63.53%
Edges between old and new users %		45.95%	42.51%	51.27%	46.73%	48.69%	49.31%	31.61%
Impact of new users		42.63%	51.83%	33.88%	27.54%	35.95%	31.67%	9.03%
post/user	2.27	2.34	2.56	2.45	2.99	2.88	2.34	1.70
Sub-category 302	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Number of new users	96	94	114	75	97	71	83	54
Number of old users		19	30	33	37	49	37	38
New user's %		83.19%	79.17%	69.44%	72.39%	59.17%	69.17%	58.70%
Edges between new users %		46.12%	66.23%	29.19%	32.87%	20.17%	33.72%	21.83%
Edges between old users %		6.98%	5.34%	19.19%	14.47%	39.42%	13.59%	31.59%
Edges between old and new users %		46.90%	28.43%	51.62%	52.66%	40.41%	52.69%	46.58%
Impact of new users		55.45%	83.65%	42.03%	45.41%	34.10%	48.75%	37.20%
post/user	1.92	2.03	2.62	1.82	2.22	2.38	2.04	1.87
Sub-category 303	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Number of new users	135	131	144	125	190	89	116	60
Number of old users		34	43	45	72	57	69	56
New user's %		79.39%	77.01%	73.53%	72.52%	60.96%	62.70%	51.72%
Edges between new users %		31.52%	21.68%	37.97%	37.55%	18.07%	7.90%	15.47%
Edges between old users %		26.32%	47.38%	15.88%	17.28%	33.47%	60.63%	30.62%
Edges between old and new users %		42.17%	30.95%	46.15%	45.16%	48.46%	31.47%	53.91%
Impact of new users		39.70%	28.15%	51.64%	51.78%	29.64%	12.61%	29.91%
post/user	1.98	2.15	2.36	2.03	2.41	2.47	3.10	2.09
Sub-category 304	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Number of new users	76	72	86	112	48	53	82	33
Number of old users		11	26	28	21	22	27	29
New user's %		86.75%	76.79%	80.00%	69.57%	70.67%	75.23%	53.23%
Edges between new users %		77.33%	29.97%	43.49%	25.61%	25.97%	45.07%	27.10%
Edges between old users %		0.53%	14.27%	11.85%	26.83%	21.04%	9.07%	27.74%
Edges between old and new users %		22.13%	55.76%	44.67%	47.56%	52.99%	45.86%	45.16%
Impact of new users		89.15%	39.04%	54.36%	36.81%	36.76%	59.91%	50.91%
post/user	1.59	1.99	2.37	1.60	1.54	1.69	2.09	1.76
Sub-category 305	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Number of new users	60	46	39	38	32	30	51	36
Number of old users		5	10	7	24	18	24	20
New user's %		90.20%	79.59%	84.44%	57.14%	62.50%	68.00%	64.29%
Edges between new users %		73.33%	49.72%	50.32%	35.65%	11.74%	22.98%	29.61%

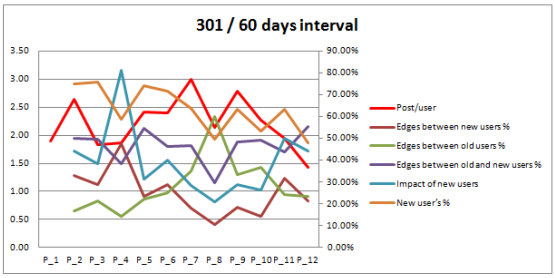
Edges between old users %		0.00%	7.58%	5.73%	13.62%	68.02%	26.95%	24.58%
Edges between old and new users %		26.67%	42.70%	43.95%	50.72%	20.24%	50.07%	45.81%
Impact of new users		81.30%	62.47%	59.59%	62.39%	18.79%	33.79%	46.06%
post/user	1.75	1.51	2.29	1.84	1.96	2.00	2.12	1.55
Sub-category 306	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Number of new users	239	173	186	192	151	124	188	91
Number of old users		55	69	85	83	67	84	85
New user's %		75.88%	72.94%	69.31%	64.53%	64.92%	69.12%	51.70%
Edges between new users %		34.31%	29.24%	32.64%	37.12%	11.13%	15.22%	5.04%
Edges between old users %		18.15%	27.33%	15.73%	15.42%	45.77%	33.28%	52.83%
Edges between old and new users %		47.54%	43.42%	51.63%	47.47%	43.10%	51.51%	42.13%
Impact of new users		45.21%	40.09%	47.08%	57.52%	17.15%	22.01%	9.76%
post/user	3.15	3.56	2.96	2.74	3.05	3.40	4.54	3.78
Sub-category 404	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Number of new users	9	25	43	29	88	45	71	25
Number of old users		2	9	13	17	27	32	12
New user's %		92.59%	82.69%	69.05%	83.81%	62.50%	68.93%	67.57%
Edges between new users %		67.78%	38.10%	43.59%	45.29%	19.03%	22.72%	34.85%
Edges between old users %		0.00%	12.80%	9.40%	7.53%	24.12%	20.57%	0.00%
Edges between old and new users %		32.22%	49.11%	47.01%	47.18%	56.85%	56.71%	65.15%
Impact of new users		73.20%	46.07%	63.13%	54.04%	30.45%	32.96%	51.58%
post/user	1.56	2.96	2.83	2.10	2.05	3.56	2.12	1.92
Sub-category 501	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Number of new users	53	79	100	90	68	67	76	63
Number of old users		9	29	33	27	28	29	25
New user's %		89.77%	77.52%	73.17%	71.58%	70.53%	72.38%	71.59%
Edges between new users %		67.70%	32.82%	31.41%	11.38%	34.38%	29.81%	50.70%
Edges between old users %		1.69%	18.50%	25.00%	35.34%	10.90%	21.13%	5.96%
Edges between old and new users %		30.61%	48.68%	43.59%	53.28%	54.72%	49.06%	43.34%
Impact of new users		75.41%	42.34%	42.92%	15.90%	48.75%	41.19%	70.81%
post/user	1.58	2.36	2.23	2.63	2.51	1.95	1.99	1.78
Sub-category 502	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Number of new users	57	44	44	22	60	26	108	20
Number of old users		9	18	9	11	12	32	17
New user's %		83.02%	70.97%	70.97%	84.51%	68.42%	77.14%	54.05%
Edges between new users %		26.20%	39.58%	33.33%	35.50%	22.12%	42.57%	14.68%
Edges between old users %		32.72%	8.33%	11.11%	12.40%	25.73%	11.67%	20.18%
Edges between old and new users %		41.08%	52.08%	55.56%	52.10%	52.14%	45.76%	65.14%
Impact of new users		31.56%	55.78%	46.97%	42.00%	32.33%	55.18%	27.16%
post/user	1.21	1.74	1.89	1.35	1.72	2.13	2.04	1.76

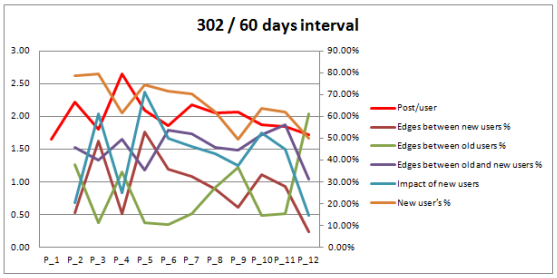
Sub-category 503	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Number of new users	44	29	44	67	63	45	62	33
Number of old users		5	11	21	27	19	29	27
New user's %		85.29%	80.00%	76.14%	70.00%	70.31%	68.13%	55.00%
Edges between new users %		67.24%	70.42%	21.05%	27.08%	42.11%	20.09%	14.91%
Edges between old users %		2.71%	1.68%	23.80%	17.51%	12.28%	31.04%	23.65%
Edges between old and new users %		30.05%	27.90%	55.14%	55.42%	45.61%	48.87%	61.44%
Impact of new users		78.83%	88.03%	27.65%	38.68%	59.88%	29.48%	27.11%
post/user	1.86	2.74	2.29	2.45	2.39	2.27	2.57	2.05
Sub-category 601	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Number of new users	97	46	60	102	96	56	101	89
Number of old users		16	38	40	43	51	48	40
New user's %		74.19%	61.22%	71.83%	69.06%	52.34%	67.79%	68.99%
Edges between new users %		20.11%	13.11%	40.25%	23.73%	2.53%	17.44%	26.42%
Edges between old users %		30.31%	42.40%	14.80%	35.91%	75.69%	32.60%	48.23%
Edges between old and new users %		49.58%	44.49%	44.95%	40.36%	21.79%	49.96%	25.36%
Impact of new users		27.11%	21.42%	56.03%	34.36%	4.83%	25.73%	38.29%
post/user	3.73	4.66	3.38	2.46	2.98	3.23	2.86	2.19
Sub-category 604	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Number of new users	161	133	136	82	89	59	153	26
Number of old users		25	40	20	26	29	49	19
New user's %		84.18%	77.27%	80.39%	77.39%	67.05%	75.74%	57.78%
Edges between new users %		53.91%	58.83%	45.99%	27.15%	28.41%	19.37%	31.18%
Edges between old users %		10.80%	7.22%	8.36%	19.74%	18.66%	30.74%	13.53%
Edges between old and new users %		35.28%	33.96%	45.64%	53.11%	52.92%	49.89%	55.29%
Impact of new users		64.05%	76.13%	57.21%	35.08%	42.38%	25.57%	53.96%
post/user	1.80	2.02	1.95	1.52	2.03	1.73	2.27	1.67
Sub-category 605	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Number of new users	131	70	86	143	106	84	139	65
Number of old users		22	31	52	59	50	63	62
New user's %		76.09%	73.50%	73.33%	64.24%	62.69%	68.81%	51.18%
Edges between new users %		53.25%	32.46%	34.49%	22.89%	22.05%	18.03%	16.29%
Edges between old users %		8.88%	31.02%	19.37%	31.65%	24.61%	30.51%	32.38%
Edges between old and new users %		37.87%	36.52%	46.14%	45.45%	53.34%	51.46%	51.32%
Impact of new users		69.99%	44.16%	47.03%	35.63%	35.17%	26.20%	31.83%
post/user	2.50	2.10	2.09	2.10	2.40	2.25	2.90	2.26

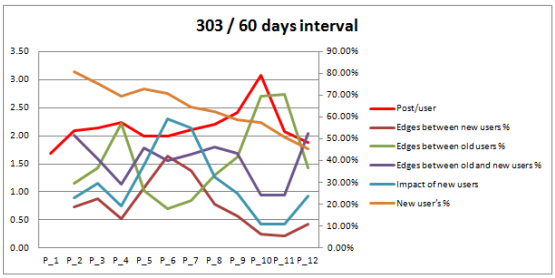
Appendix H

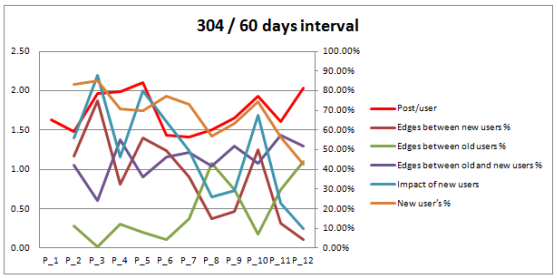

101 / 60 days interval
Post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

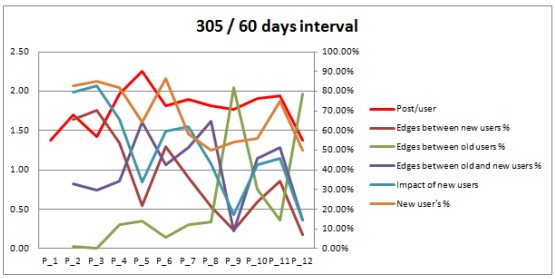

104 / 60 days interval
Post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

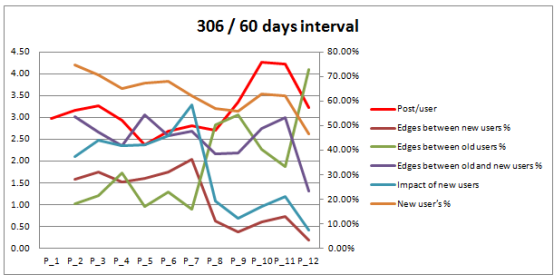

301 / 60 days interval
Post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

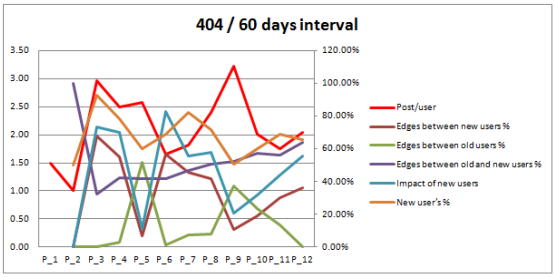

302 / 60 days interval
Post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

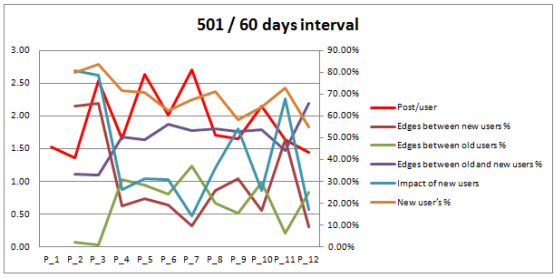

303 / 60 days interval
Post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

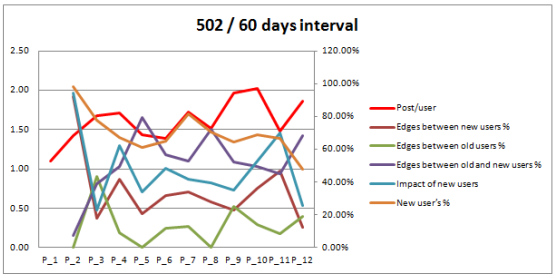

304 / 60 days interval
Post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

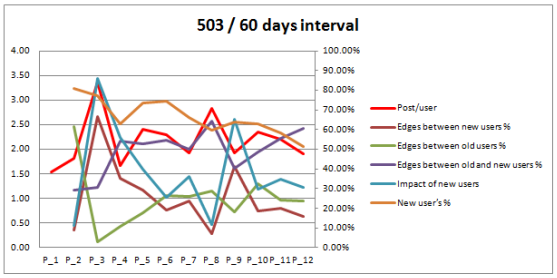

305 / 60 days interval
Post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

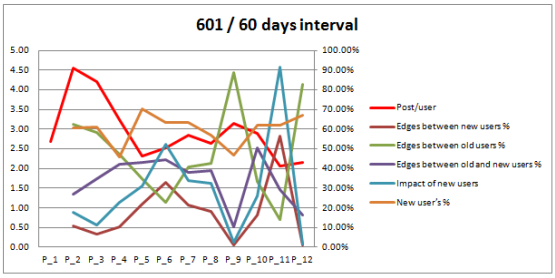

306 / 60 days interval
Post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

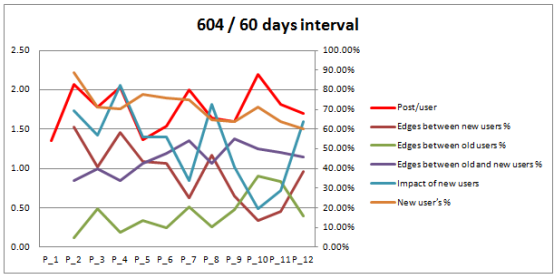

404 / 60 days interval
Post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

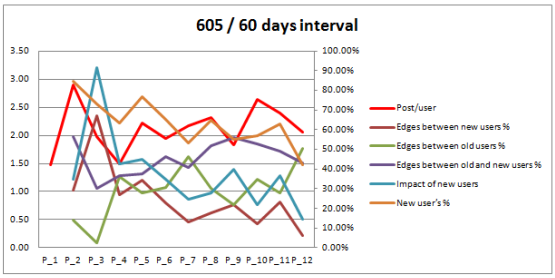

501 / 60 days interval
Post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

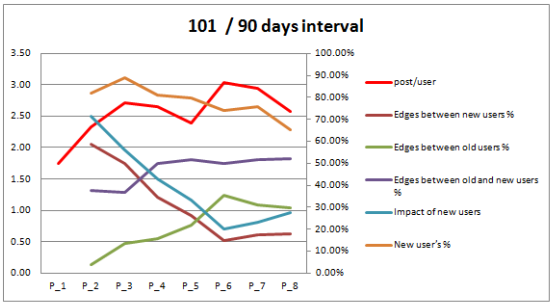

502 / 60 days interval
2.50
2.00
1.50
1.00
0.50
0.00
120.00%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
P_1 P_2 P_3 P_4 P_5 P_6 P_7 P_8 P_9 P_10 P_11 P_12
Post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

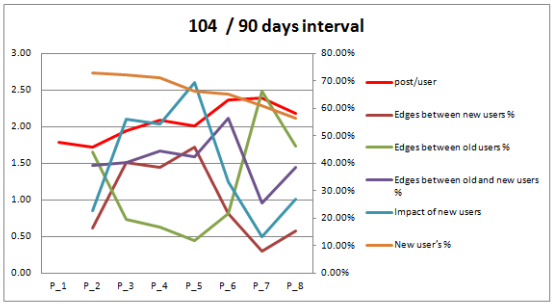

503 / 60 days interval
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
P_1 P_2 P_3 P_4 P_5 P_6 P_7 P_8 P_9 P_10 P_11 P_12
Post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

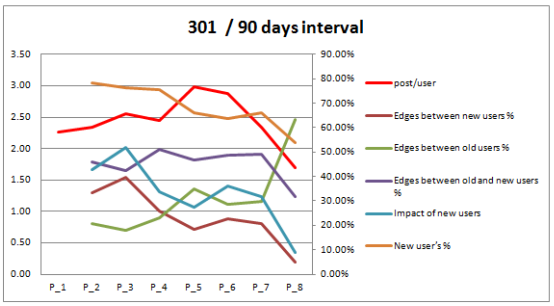

601 / 60 days interval
5.00
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
P_1 P_2 P_3 P_4 P_5 P_6 P_7 P_8 P_9 P_10 P_11 P_12
Post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

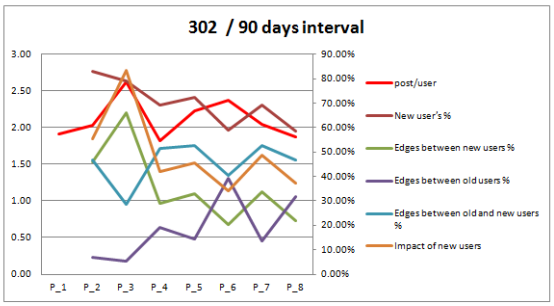

604 / 60 days interval
2.50
2.00
1.50
1.00
0.50
0.00
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
P_1 P_2 P_3 P_4 P_5 P_6 P_7 P_8 P_9 P_10 P_11 P_12
Post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

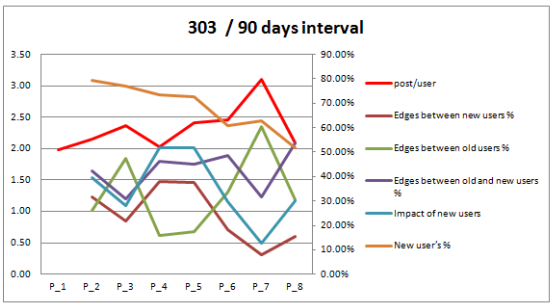

605 / 60 days interval
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
P_1 P_2 P_3 P_4 P_5 P_6 P_7 P_8 P_9 P_10 P_11 P_12
Post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

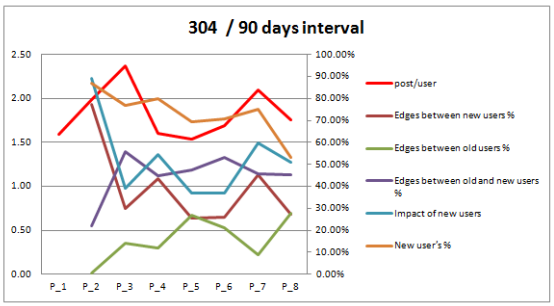

101 / 90 days interval
post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

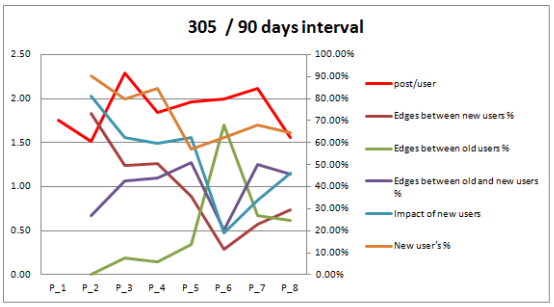

104 / 90 days interval
post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

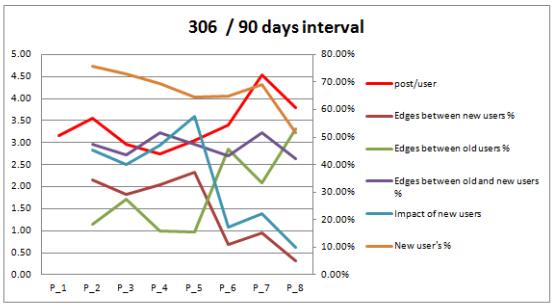

301 / 90 days interval
post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

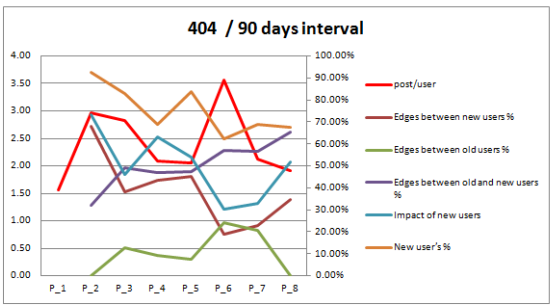

302 / 90 days interval
post/user
New user's %
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users

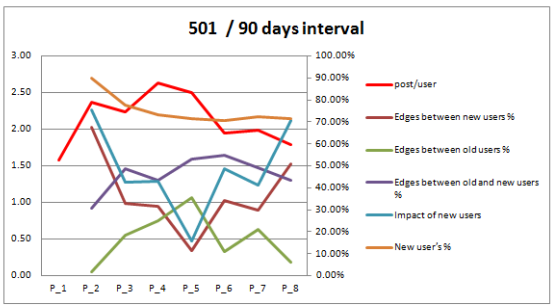

303 / 90 days interval
post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

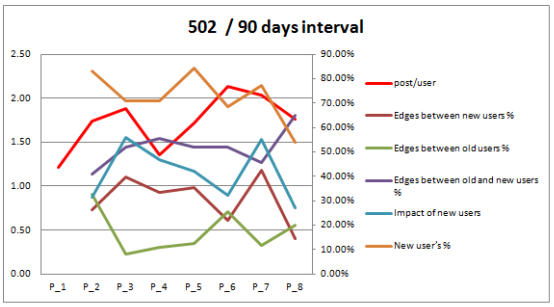

304 / 90 days interval
post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

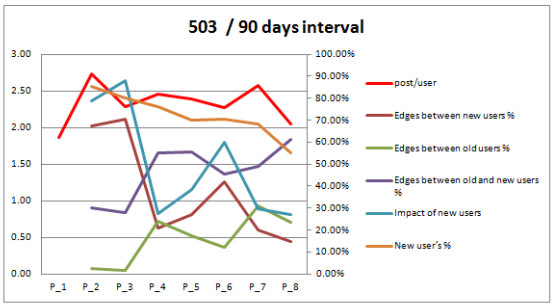

305 / 90 days interval
post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

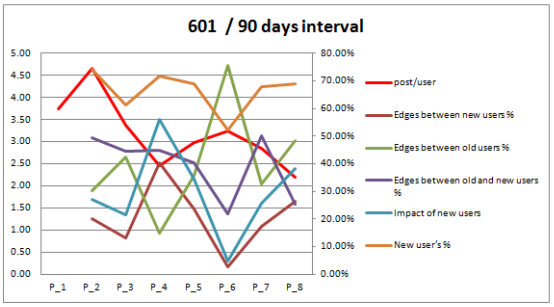

306 / 90 days interval
post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

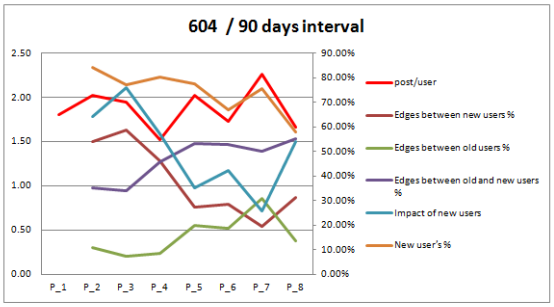

404 / 90 days interval
post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %

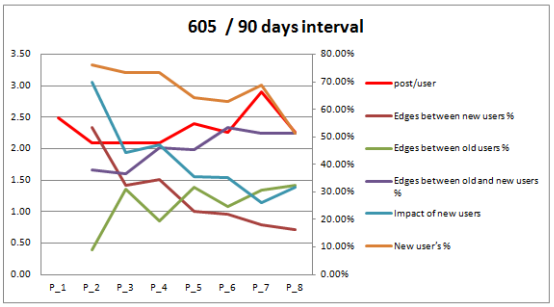

501 / 90 days interval
post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %


502 / 90 days interval
2.50
2.00
1.50
1.00
0.50
0.00
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
P_1 P_2 P_3 P_4 P_5 P_6 P_7 P_8
post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %


503 / 90 days interval
3.00
2.50
2.00
1.50
1.00
0.50
0.00
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
P_1 P_2 P_3 P_4 P_5 P_6 P_7 P_8
post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %


601 / 90 days interval
5.00
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
P_1 P_2 P_3 P_4 P_5 P_6 P_7 P_8
post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %


604 / 90 days interval
2.50
2.00
1.50
1.00
0.50
0.00
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
P_1 P_2 P_3 P_4 P_5 P_6 P_7 P_8
post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %


605 / 90 days interval
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
P_1 P_2 P_3 P_4 P_5 P_6 P_7 P_8
post/user
Edges between new users %
Edges between old users %
Edges between old and new users %
Impact of new users
New user's %