The Diffusion of Information in Emerging Scientific Fields:

A Different Methodological Approach to Measure Collaboration in Research

Joachim Bråthen

Master Thesis at the Center for Technology, Innovation and Culture

UNIVERSITY OF OSLO

March 2016

2016

The Diffusion of Information Emerging Scientific Fields: A Different Methodological Approach to Measure Collaboration in Research

Joachim Bråthen

http://www.duo.uio.no

Print: Reprosentralen, University of Oslo

Abstract

The fields of innovation studies, entrepreneurship and science and technology studies are all relatively new additions to the academic landscape. Research on the development of these new fields has been undertaken, yet there has not been a thorough investigation of the co-authorship networks that shape them on an individual researcher level. As co-authorship networks best represent the diffusion of tacit information between researchers it is of importance to understand the structural characteristics that facilitate the flow of knowledge. The main method employed here is social network analysis of bibliometric data spanning ten years for the 20 most influential journals for these three fields. Additionally, a text-mining approach is used to uncover their differing research themes and their importance. Together it provides the most holistic and thorough investigation of this theme done to date. The results show that the fields are less separated than earlier research has argued. Many of the researchers involved are not limited to publishing within one of the fields, but rather publish in several of them. There has been a convergence of the fields in that sense. There are, however, differences in the collaborative structures of the fields. The fields are highly fractured in terms of collaboration between researchers; however they are all characterized by tightly knit groups of researchers that collaborate extensively. Further, the research themes are not as similar as thought previously. Overall innovation and entrepreneurship are closer to one another than to STS. Some individuals in the network have proven to be extremely important for the diffusion of information, while most are not particularly central for diffusion because of the fractured characteristics.

Preface

This thesis is the product of several months of work. Although the research presented here has been arduous and time consuming in the making, it has been a tremendously rewarding experience that I would not want to be without. As innovation studies, entrepreneurship and STS continue to progress and to an increasing degree influence important decisions in society it is important to take a step back and evaluate their impact and to analyze how emerging fields of study evolve to better understand the necessary steps that needs to be taken to secure academic development and sharing of information. Through this research I aim to do both.

Social network analysis is increasingly a method used to analyze the structures of academic fields. This set of methods constitutes a valuable contribution to the study of innovation. As of yet this method is not commonly taught at university bachelor and master courses, at least not in social science departments. To acquire the necessary experience and skills to be able to do the research I knew that I wanted, I took take a PhD level course in addition to my regular classes, the product of which this research is built upon. Through this research I hope that this method will become an increasingly popular approach in general and in innovation, entrepreneurship and STS in particular. As these fields are all concerned with the flow of knowledge I hope to see more of this method taught and employed at the centers that focus on these issues, as there are still many aspects left unexplored.

I wish to sincerely thank my supervisor Magnus Gulbrandsen for his excellent help and guidance through my work. I also whish to thank my good friend and brother Aleksander Bråthen and my good friend Mads Motrøen for their invaluable help and support through my academic education and this thesis in particular. It is greatly appreciated. Lastly I wish to thank the Center for Technology, Innovation and Culture where I have written this thesis, as well as the field of innovation studies, entrepreneurship and STS for providing me with new ways of thinking and evaluate innovation, technology and their impact on society. An ever-increasingly important way of viewing the world we inhabit.

Table of Contents

List Of Figures

1 Introduction

Innovation, entrepreneurship and technology research is a fairly recent addition to academia. Contributions to further knowledge about innovation include scholars from several older academic disciplines. Prominent examples are: economics, business, history, sociology, political science and management. Innovation research has been cross-disciplinary from its beginning. This collaborative attribute captures the importance of studying innovation from multiple perspectives in order to understand it completely (Fagerberg, 2005). The multitude of disciplines that are present in the field is widely considered to be crucial and important for the study of innovation. At the same time it might lead to a clustered community of researchers, as collaboration between scholars from different disciplines might be low (Katz & Martin, 1997). Because the field is young and comprised of several research disciplines, there is a distinct possibility that these studies exhibit certain characteristics of fracturing. Meaning that the cooperation between researchers and/or the research themes discussed are not particularly well defined.

In this thesis I will investigate the differences and similarities between sub-fields that focus on innovation. The fields in question are innovation studies, entrepreneurship and science and technology studies. Although there are several ways to divide the academic fields that are occupied with different relevant research themes, there is a convention for using these three. Since all of these are new additions to academia, there is reason to suspect that they have not fully developed into clear and separate fields. In this thesis I will answer whether or not the fields are growing further apart into distinct and separate fields, or if they are more or less the same, but separated by methodological and historical differences.

To achieve this I will answer four concise research questions. The first being: *How is information diffused through the networks of researchers within their fields, and are there any systematic differences in the flow of information between them?* This question is important to answer as the flow of information through a community of researchers are of the utmost importance in any scientific field, and quite possibly to an even larger extent for emerging scientific fields. As any well functioning network needs to have a well functioning system of diffusion on knowledge, this investigation can potentially lead to new insights that could shape research policy in general and the three fields under study in particular.

The second research question is: *Have the fields developed into three clearly separated ones when analyzed on an individual researcher level?* Earlier research have concluded that the fields are separate entities, but the methods employed here are different and on a more granular level than previous work. From this we will learn both how the empirical development and the state of three highly relevant and steadily more important fields of research, and the methodological aspect of difference in results when analyzing similar data on a micro-level as opposed to macro-level. Insights that in turn is important for research on this theme moving forward, as the methodological best practices has yet to be agreed upon.

Who are most important for the diffusion of information through these networks? By using centrality measures we can find who are the most important in terms of the spread of information, as well as who are the most important gatekeepers of information. By studying this we can find if the information is evenly distributed in the networks and if there are some researchers that are significantly more important for the diffusion than others. It is also important to look at the differences between the fields to determine if they are indeed separate.

Which social network method is most fitting to measure cooperation in science and what type of challenges do these have when studying emerging fields? Thus a part of my contribution will be to find out what we can learn in terms of developing accurate methods. As social network analysis of bibliometric data on emerging fields is still relatively new, it is important to get a good understanding of what methods work best to answer different questions, as well as which metrics we should focus on.

This study is thus more inclined towards the empirical development of the three fields because of their relevance and importance, not to contribute towards a more general theory of the dynamics of emerging scientific fields, as there is already a significant amount of work done in the sociology of science on this. Additionally I will focus considerably on the methodological aspects of this form of research through new approaches to social network analysis of emerging fields. In that sense a large portion of this study is related to what we can learn through using these cutting-edge methodological approaches.

The way information and knowledge is produced and diffused is to an extent dependent on the social structure of in which they take place. This is as mentioned a common theme of study in the sociology of science. This study is occupied with the empirical development of three fields, not the sociology of science in more general terms, but it is useful to shortly discuss it as some of the previous contributions detail the importance of studying individual researchers. An important contribution in this aspect is Diana Crane´s notion of invisible colleges (1972). An invisible college refers to a group of researchers that work within the same paradigm and shares many of the same research themes and objectives.

Crane herself defines it as “a communication network of a subgroup of researchers within a research area” (Crane, 1972). Further Tiegland (2003) explains her idea of invisible colleges as:

> […] scientists within a research field organize themselves into subgroups of informal networks of personal relationships, or invisible colleges that are characterized by strong ties based on informal collaboration. These invisible colleges are then linked to individuals within other research fields through weak ties by their members, thus facilitating the diffusion of information both to and from each field. A common language based on a similar orientation towards research facilitates communication between individuals from different fields. With regard to performance, Crane suggests that the position of a scientist in the invisible college impacts his or her awareness of existing research as well as how rapidly he or she obtains information (Tiegland, 2003).

To research the diffusion of information and the invisible colleges that might arise in the networks of these new scientific fields, we must thus primarily look at the individual researcher and how they collaborate and share information between each other. This is to a large extent the reasoning behind opting for a co-authorship study on an individual level. The research of Crane and others delve deep into general development of new scientific fields, this is not a large focus in this thesis. Rather I will investigate the empirical development of the three fields in question in particular while also focusing on development of useful methodological approaches to this kind of research.

It can be argued that innovation studies is mostly occupied with factors that promote growth through innovation, entrepreneurship is more focused on the individuals and businesses and how they grow, where STS is more occupied with the specifics of knowledge creation and its effect on society. I will expand on earlier research, with a broader focus than what has been done so far and in so doing expanding on the knowledge base on how this field of study is put together.

This thesis will investigate these issues in a new and more holistic way than previous research. By utilizing different types of unstructured data, I will quantify the information by using several methods, with heavy reliance on social network analysis of co-authorship. Co-authorship represents a relational statistical indicator, meaning that it measures the actual cooperation between researchers as opposed to passive citation. This allows us to better investigate the actual flow of knowledge through networks. Earlier research has mostly focused on articles and books as their level of analysis, by using authors I will achieve a much higher level of granularity.

Combined, the methods employed describe the current state and to what extent cooperation take place within the fields, as well as the most important research themes. Specifically I will use the 20 most central journals in innovation, entrepreneurship and STS as the basis for a bibliometric social network analysis. I will use all the articles published in these journals in a ten year period to withdraw data on who has co-authored articles with whom, and how the fields are structured. Additionally I will use the keywords from these articles as a basis for establishing what the fields are researching. The primary types of analysis I will do is:

- Use article keywords to determine the most important research themes.
- Social network analysis of the bibliometric data for the fields separately and then the entire field seen as one.
- Identify the most central authors in the fields and who are important for the diffusion of information.
- Community detection analysis to identify if there are clear boundaries between them.

I will analyze the fields on a researcher level using co-authorship as opposed to citations on the level of books and handbooks, which has been a common way of analyzing new academic fields in earlier research. By doing this I will explore the research theme of newly established disciplines in the most comprehensive way done so far.

Several earlier studies and research projects have focused on the structure of new scientific fields. The studies focus on the field of innovation studies historically while also showing which seminal works and authors have helped shape the field as well as general accounts of what themes are researched and which researchers has been important for the field in their earlier phases. Yet there is little knowledge about the specifics of what the field is preoccupied with explaining, and who are the most important researchers in the field. Further, it is not yet clear whether or not they have overcome the challenges emerging scientific fields often are faced with. One previous study concluded that:

> The overall main conclusion from our analysis is that the social science literature on knowledge, technological change and innovation has developed in a progressively more compartmentalized manner. In terms of their citation profiles, the three fields of science and technology studies (STS), innovation studies (INN) and entrepreneurship (ENT) now appears as largely distinct, not as part of a strongly connected field (Bhupatiraju et al., 2012).

Although earlier studies conclude that they are indeed separate, the results of this thesis show that although sub-fields are arguably separate when analyzing them on the level of books and important contributions, they are far less removed when analyzing them on an individual researcher level. Especially there is a large amount of overlap between innovation studies and entrepreneurship, while STS is more separate from the other two. This is in contrast to some earlier research that has shown that entrepreneurship is more separated to the other two. Both keyword analysis and social network analysis point in this direction. Many of the researchers involved seem to publish under the domain of innovation studies, entrepreneurship and STS alike, perhaps not as surprising as it might seem, when considering that all of these fields have a similar focus. This indicates a convergence of the fields despite historical differences. However, all the fields show clear signs of being fractured, with small groups of researchers that actively collaborate. All three fields, as well as all of the fields seen as one, are strikingly sparse when looking at co-authorship networks, yet they are all highly clustered. This means that there are many small clusters of researchers that choose to collaborate extensively; but that the cooperation between these tightly knit groups are low. The fields are indeed separate when analyzing them historically but when investigating individual roles, the fields are not as different as previously thought. This is possibly due to the fact that researchers often choose to collaborate with other researchers that share their interest, affiliation or educational background. There is, however, evidence that the fields differ in their structure and level of cooperation. Whereas innovation studies and entrepreneurship are quite similar in this sense, STS show a lower level of collaboration and a propensity for published articles that are written by fewer authors, as well as also being the field with the highest amount papers written by a single author.

Because of the fracturing of the networks there are a few nodes that are tremendously important for the diffusion of tacit information through them. Similarly most authors do not contribute substantially to the diffusion of tacit information because of the cliques that have developed, but are instead containing much of the information within the groups. Given that the fields would have diverged into clearly separate ones, as measured on an individual level, we would expect the cluster analysis to result in clear signs that there are indeed three fields. This is not the case. The cluster analysis shows that the network has not evolved into clearly separate entities. Again supporting the finding that individual researchers are not necessarily deeply embedded in only a single field.

This study is explorative in its nature, and investigates to what extent an academic field, consisting of several sub-disciplines, share information and tacit knowledge and the level of cooperation in new emerging fields. Additionally, it serves as a detailed descriptive paper on the state of research on innovation, technology and its impact on society today, as there is a growing need to understand a field that hugely impacts innovation and R&D policy in most countries.

This study consists of six parts. In the second part of this paper I will discuss theory, research and the motivation behind this study. In the third part I will quickly explain social network analysis and the metrics that are used. The fourth part is description of the data and some descriptive statistics. The fifth part is a thorough investigation of the results, before I conclude in part six. The analysis of data is done with the R programming language and software environment (http://www.r-project.org/), with heavy reliance on the iGraph package for R that allows for effective social network analysis (http://www.igraph.org/r/).

2 Theory, Framework and Research Questions

In recent decades several new fields of study have emerged that investigates the economic and societal impact of knowledge, technology and innovation. This newfound academic attention has been important in setting policy debates, both locally and globally, on how a society can manage and promote economic growth, technological development and innovation policy. The study of innovation can help us understand "what governments have done and could do to promote the production, diffusion, and use of technical knowledge in order to realize national objectives" (Lundvall & Borrás, 2005). It is thus important to understand the various aspects of the fields that focus on technology and innovation because of the need to understand what impact the technology and innovation policies. Innovation in itself however is in no way a new phenomenon. Fagerberg describes the importance of innovation:

> Innovation is not a new phenomenon. Arguably, it is as old as mankind itself. There seems to be something inherently 'human' about the tendency to think about new and better ways of doing things and to try them out in practice. Without it, the world in which we live would look very different. Try for a moment to think of a world without airplanes, automobiles, telecommunications, and refrigerators, just to mention a few of the more important innovations from the not-too-distance past. Or – from an even longer perspective – where would we be without such fundamental innovations as agriculture, the wheel, the alphabet, or printing? (Fagerberg, 2005).

Despite the general agreement that innovation and technological development is of the utmost importance for societal change, it has received relatively little academic attention up until the 1960's when research centers, such as Science Policy Research Unit (SPRU), began focusing on innovation. Since then the field has proliferated and is now a fast growing multi-disciplinary field of study, mainly within the social sciences. There are now many fields or sub-fields of study that focus on some of these issues. These areas of research are similar in what they study, but they are also in many ways different and distinct from one another.

In this chapter I will detail the theory, framework and research motivation of this thesis. Additionally, I will describe research already done trying to answers some of the research questions. Ultimately, this thesis will provide a clear and more nuanced picture of the state of research on innovation and technology.

2.1 Characterizing the Fields of Innovation and Technology Development

As mentioned, we can argue that there are different, and to some extent separate, fields of research focusing of innovation and technology. It is important to understand the characteristics of these fields, what themes are studies, how the knowledge base is built, how this information spreads through the network of researchers and how the fields cooperate. Through this we can understand how this impact policy decisions, and therefore a key driver of economic growth and societal change. However, relatively little work has been done to explore these aspects of the different fields, as Fagerberg and Verspagen describe: "Despite the popularity of the phenomenon, very little has been written on the community of scholars that study innovation and contribute to the knowledge base necessary for designing innovation policy" (Fagerberg & Verspagen, 2009). The article cited here is one of the first studies done on exploring the characteristics in the field. However, the article is fairly limited in terms of what is studied. Fagerberg and Verspagen mainly focus on the economical and Schumpterian perspective of innovation studies. After the publication of their article there has been a research project conducted called EXPLORE, the most extensive project focusing on these issues to date, which I will describe and discuss below. I will thus expand on what has been written on the field, with a broader focus than what has been done so far and in so doing expanding on the knowledge base on how this field of study is put together.

2.1.1 What is Innovation, Entrepreneurship and STS?

As the study of innovation and technology has advanced, there has been a proliferation of sub-disciplines that all focus on aspects of knowledge, innovation and technology. There has been proposed many different conceptual separations of this overarching field of study, yet a separation between three quite distinct sub-fields has become the norm. Additionally, this distinction is reflected in many of the research centers that specialize in these research themes. Previous research often also use this norm when investigating the difference between the sub fields. The research of technology and innovation has developed mainly along three paths, namely innovation studies (INN), entrepreneurship (ENT), as well as science and technology studies (STS). These fields are all concerned with the economic and societal aspects of technology and innovation, albeit in different ways. There are discussions in the academic literature about in what ways and to what extent these three fields actually comprise three distinct fields of study, or if they rather are three analytical divides in what is actually one rather large field of study. Whether the three fields – this paper will refer to them as fields – are best viewed as three distinct separate fields with partial overlap of research topics, or whether they should be seen as one overarching field will not be a main discussion in this thesis. We will, however, discuss the usefulness of this analytical separation briefly. We can note that there is a current convention to refer to them as separate fields, and further the theoretical and empirical work this thesis is largely based upon discuss them as different fields. The discussion of what constitutes an academic discipline or field is as mentioned beyond the scope of this thesis and we will allow to state that there is "relationships between the three fields. As they are all concerned with broadly related topics, namely the social (including economic) context and usage of knowledge, there are linkages between them" (Bhupatirajuet al., 2012). The assumption that it is correct, or at least useful, to refer to them as separate fields are further supported by the discussion and statement that:

> One might even hypothesize that instead of three separate literatures, there is really only one (large) social science literature about knowledge and innovation, although prior causal impression held by many participants in the EXPLORE project (based on their experience as practitioners in one or more of the three fields) was that this would probably not be the case (Bhupatiraju et al., 2012).

The three fields of innovation, entrepreneurship and STS will thus be referred to as three separate fields in this thesis. This also provides the basis for the analysis, without an a priori establishment of them as three separate fields, the comparison of them would not be possible. As this thesis will compare and expand on existing comparisons between them, it is important to bear in mind that the convention of separating them into three fields will to some extent be reinforced in the analysis in terms of a certain circularity in that we are comparing three entities that could arguably actually be thought of as a single entity. The analytical process of treating them as three fields will to some extent force the analysis towards the conclusion that they are, indeed, separate. This is, however, not of considerable importance in this thesis as we accept the assumption that they are separate. Along with the convention of discussing these three specific sub fields there are also general characteristics about the fields that are used to discuss their similarity and differences.

The term innovation is frequently used in today's society, yet this has not always been the case. Schumpeter is the scholar associated with establishing the modern science of innovation studies in the early 19-hundreds. He viewed innovation as the most significant factor of economic growth, and thus something that needs to be studied and understood to better understand how society develops. However, the field did not begin in earnest until after WW2 where scholars such as Arrow, Nelson and Winter began their investigations on the effects of R&D and innovation in an economic sense in the US.

At about the same time, the economist Freeman began his work on the same effects in Britain. He later went on to develop the system that is used to codify and gather statistics on innovation today. These early innovation scholars drew much of their inspiration from the fields of economics and sociology, it was only in the 1960's that the arguably most important innovation studies institution Science Policy Research Unit (SPRU) in Britain was established, distinguishing innovation studies as a separate academic field.

Later many research centers have been established, to a large extent modeled after SPRU, where practitioners from several disciplines were not only seen as a strength, but rather a necessity for understanding innovation holistically. This meant including not only social scientist, but engineers and natural scientists as well (Fagerberg, Martin, & Andersen, 2013). The practitioners of innovation studies define it as "the scholarly study of how innovation takes place and what the important explanatory factors and economic and social consequences are" (Fagerberg, Fosaas, & Sapprasert, 2012). Innovation studies are therefore more inclined towards macro factors that influence society and specifically the role of innovation in this process.

Entrepreneurship studies rose out of management and business studies to a larger extent than both innovation studies and STS. The entrepreneurship scholars are often preoccupied with what characteristics entrepreneur and start-up businesses posess, and how a firm can be innovative, as opposed to innovation studies where whether a particular business fails or not is not really given much importance choosing rather to focus on aggregate effects. However, entrepreneurship shares the view that entrepreneurial activity and innovation is the main driver of economic growth in society. Early entrepreneurial theorists include well-known figures such as Knight who emphasized that entrepreneurs take on uncertainty and not risk that can be calculated using probability theory. The uncertainty faced by entrepreneurs is, according to Knight, immeasurable and therefore have potential high risk, but also high reward. Schumpeter is also a founding figure in entrepreneurship and he saw the entrepreneur as the agent that sees and acts on innovative solutions, thus keeping an economy out of a steady equilibrium. Later, the focus of entrepreneurship studies shifted gradually to look at who and what makes a company successful, often seen as partly a result of these entrepreneurial people (Kariv, 2011). Entrepreneurship can be difficult to define since the term is used interchangeably between the entrepreneurial mindset of people, and the task of running a self-started business. Some practitioners of entrepreneurship studies work with the following definition of entrepreneurship studies:

> The field of entrepreneurship [is] the scholarly examination of how, by whom, and with what effects opportunities to create future goods and services are discovered, evaluated and exploited.' Thus they argue that entrepreneurship involves *sources* of as well as the *processes* of discovery, evaluation and exploitation of opportunities, but also the set of *individuals* who, discover, evaluate and exploit these opportunities (Landström, Harirchi, & Åström, 2012).

We need not delve too deep in the discussion of precisely what constitutes entrepreneurship; we can rather focus on the definition of entrepreneurship studies. The definition of entrepreneurship studies places it in a research tradition that investigate micro-aspects of how innovation takes place and what is needed for successful entrepreneurship.

Science and technology studies also appeared at around the same time as the others and gained an academic foothold in the 1960's. An early and immensely influential contribution to the field was Kuhn's 'The Structure of Scientific Revolutions', which still retains its importance today. Other important and influential STS scholars followed, such as Bijker, Pinch, Merton, Collins and Latour, all of who are tremendously important and influential in the field today. STS arose partly in response to a growing concern for the negative externalities associated with scientific endeavor, and scientists and academics began to realize that there was a need for a 'science of science' to address these concerns. STS sprung out of mainly history and the sociology of science, yet includes practitioners from many other disciplines. In STS, science and technology is a social process. Entailing that scientists can never fully be disconnected from the social processes and environment surrounding them, i.e. norms. It is these constructs that to a large extent are studied in STS (Sismondo, 2010). Science and technology studies are to a large extent preoccupied with knowledge creation, and several methodological approaches are utilized to investigate this. The field is mostly associated with the sociology of science and knowledge, but also has practitioners that focus on science indicators, science policy and the history of science (Martin, Nightingale, & Yegros-Yegros, 2012). STS practitioners are very well aware of the field's historical roots and differences in thinking between some of the field's most important actors. Additionally, the field seems to be more based around ideas, rather than assumptions as innovation and entrepreneurship is.

From the short descriptions of the three fields, we can clearly see that they are indeed related, yet focus on clearly different aspects of what some would claim to be the same subject. Where innovation is largely a study of macro factors promoting growth through innovation, entrepreneurship is a micro study of what is needed on an individual level, and STS is the study of how scientific knowledge is produced. The fields are therefore highly interconnected in their overall perspective, yet fundamentally different in their approach to the problem.

2.2 Earlier Research

Bibliometric analysis of the development of innovation, entrepreneurship and STS over time is not a wholly new endeavor (Bhupatiraju et al., 2012). The goal of that research was to use bibliometric analysis to investigate the three fields in question as based on the core contributions specified in the research. They did this by looking at the citation network between researchers, but with a time dimension included in their analysis. This allowed them to also look at the development over time. They then analyzed between-field citation patterns and found that there are relatively few between-field citations between the fields, with innovation studies being most involved with the other fields.

They then did a cluster analysis using social network analysis (SNA) methods. A clustering analysis is an algorithm that organizes nodes in a network into groups based on maximizing between-group dissimilarity while also maximizing within-group similarity. This is a more open way to detect if the fields are indeed separate, as the citation analysis was based on the a priori assumption that there are three distinct fields. With a clustering algorithm the analysis may show other groups than the specific innovation, entrepreneurship and STS divide that is the conventional separation. There are several different clustering algorithms available where computing time quickly becomes a relevant factor for choosing the specific algorithm. I will discuss this further in the methodology chapter, and for now state that the research based their clustering algorithm on the principle of modularity. Through this analysis they concluded that the three fields of innovation studies, entrepreneurship and STS are indeed clear subfields by stating that "[t]he impression that this gives is that the complete citation network indeed consists of subfields, which largely corresponds to the three subfields under study" (Bhupatiraju et al., 2012). They also found that innovation studies is positioned between entrepreneurship and STS. Meaning that innovation studies have more in common with both entrepreneurship and STS than entrepreneurship and STS have with each other. Additionally, innovation studies are closer to entrepreneurship than STS, indicating that STS is further removed from the two other fields, and thus represents a more separate field.

The study concluded that "the social science literature on knowledge, technological change and innovation has developed in a progressively more compartmentalized manner" (Bhupatiraju et al., 2012), and that it "now appears as largely distinct, not as part of a strongly connected field." They do, however, also draw the conclusion that innovation studies and STS have basically the same common denominator, in terms of the earlier works cited, while entrepreneurship appears to have its beginnings at an altogether later point in time. Combining what we discussed earlier about innovation studies being 'in-between' entrepreneurship and STS, we can therefore assume that entrepreneurship have sprung out of innovation studies and have little in common with STS, at least directly. This also seems to be the case when consulting the paper of Bhupatiraju et.al. (2012).

Where this research used the core contributions identified through three case studies, I will use the most central journals in each field as the basis for my data collection. Additionally, they based their study on citations over a large time horizon, an approach that has many weaknesses in terms of discovering how tightly knit the fields are together in a specific timeframe. While citation analysis is excellent at discovering theoretical common roots over time, it is not particularly well suited to discover how researchers on an individual level interact in a specific period of time. To be able to investigate this, I have chosen a different approach, while contributing and expanding on the research already undertaken. Thus creating a fuller and deeper understanding of how the fields are interacting with each other and expanding the current research already in place. In effect earlier studies use scholarly work, i.e. articles or books, as nodes in the network, by using the authors of works we can achieve a much higher level of granularity while also being able to look at how the practitioners themselves interact with each other. Thus, by moving the level of analysis down to the author we will be able to investigate specific interactions, not only the diffusion of innovation over long periods of time. The point of this approach is not to discredit the framework and theory employed by the researchers in earlier projects, but rather to expand on the investigation they have led by contributing to the research with a new approach to the research topic.

To conclude that the fields are either separate entities or part of a larger social science field that focus on innovation and technological development it is important to look beyond the structure that makes up the field. Earlier research has done little in terms of exploring the fundamental differences in what separates them, i.e. what they research. They have, however, provided a good overview of the different backgrounds when they investigated the historical development of the fields. However, this does not capture the specific research themes. Different fields might have differing roots and be dispositioned to publish in different journals, while still researching the same topic. This is in my opinion a flaw with the research. Where the social network of citation reveals one aspect of sameness or co-development, a more fundamental analysis on what they research might provide a different and additional aspect:

> Social scientists studying the societal and economic impact of knowledge gradually cluster into distinct subgroups that are driven by specific norms, beliefs and values that evolve in each subgroup. The results of our quantitative analysis are consistent with such a view, but only more specific qualitative research can provide further support for such a hypothesis (Bhupatiraju et al. 2012).

This qualitative research can be undertaken in a quantitative way by using unstructured data in the form of article keywords and article abstracts. This allows us to quantitatively investigate qualitative data and the differences in research themes between the three fields. By analyzing the differences in research topics we can both get insight into what the different fields are researching, but we can also provide statistical indicators on the similarity of between-fields research topics. Thus, not only providing a purely qualitative study of what the research focus is in the fields, but also accurately pinpoint the differences and similarities. Provided that the dataset withdrawn from Web of Science provides two ways to do this, namely by analyzing the keywords, or by analyzing the abstracts, it is fitting to shortly discuss which way would best serve the research goals of this thesis. Keywords provide a basic overview of the themes of a specific article, as an article typically contains no more than six keywords. Whereas abstracts are longer, with more complex sentences without any specific added benefit in terms of analysis. I will therefore opt for keywords as the subject of text-mining analysis, as they are indeed keyword summaries of the longer abstract.

By combining the different levels of granularity, focus and time, and combining this with new analysis of unstructured data, we will be able to draw the most detailed and substantial picture of the fields of innovation studies, entrepreneurship and science and technology studies done to date. This will be done both by approaching the theme from a more fundamental level, e.g. data mining of research themes, as well as providing a more granular social network analysis on the researcher level, as opposed to the article or book level. The next chapter will discuss the theory, reasoning and framework behind social network analysis and bibliometric analysis, as this is the main theoretical and methodological framework for the thesis. Since I will increase the granularity of the analysis to the author level I will not use citation analysis. Rather I will opt for co-authorship analysis. The reasoning behind this will be explained in detail below.

2.3 Bibliometric Analysis

Bibliometric analysis has become an important and useful way to analyze academic disciplines. A thorough discussion of the methodology itself will be undertaken in the methodology chapter, but as bibliometric analysis and SNA are both a methodology and a theoretic framework to look at certain phenomena, I will briefly discuss it as well as the theoretical reasoning for opting for co-authorship as opposed to citation analysis. Bibliometrics, or scientometrics as it is sometimes referred to, is the analytical framework used to quantitatively analyze scientific publications (Gauthier, 1998). In this thesis I will utilize co-authorship as it represents a relational statistical indicator, meaning that it seeks to measure the actual collaboration between researchers and not the more passive act of referencing another researcher. This approach will allow us to look specifically at the flow of knowledge between the three fields in question, as Gauthier explains: "Co-author analysis is the most frequent relational indicator. It helps identify links and interactions between the actors of national and international systems of science and technology. Such interactions constitute the flow of knowledge" (Gauthier, 1998). This is in turn related to the importance of policy makers having a complete and accurate view of the field to be able to effectively make policy decisions.

2.4 Co-Authorship Research Using Social Network Analysis

When studying the importance of researchers in a specific scientific area, the most intuitive way of measuring the importance of scholars within their field is arguably by counting the number of citations the authors have. Research on citation networks is abundant and a common way to rank researchers is by simply summing the number of citations any researcher or specific paper may have. Citation network research began with a paper from Garfield et al. (Garfield, Sher, & Torpie, 1964). Since then, citation research has become a popular strand of research with many additions (Small, 1973; Small & Griffith, 1974; Collins, 1974; Hoffman & Holbrook, 1993; Ramos-Rodríguez & Ruíz-Navarro, 2004).

Social networks are studied for many different reasons. The importance of studying the networks themselves is tied to the observation that information is spread through social interaction or networks and that "because their structure has important implications for the spread of information and disease. It is clear, for example, that variation in just the average number of acquaintances that individuals have (also called the average degree of the network) might substantially influence the propagation of a rumor, a fashion, a joke, or this year's flu" (Newman M., 2001).

Co-authorship networks are similar in concept to the perhaps more well-known citation networks, and many of these networks have been studied (Newman, 2001; Smeaton et al. 2002; Cunningham & Dillon, 1997; Liu et al.). The motivation behind choosing a co-authorship network as opposed to a citation network is linked to the level of information diffusion that can be transferred from one researcher to another (Liu et al.). Citations are not necessarily linked temporally nor indicate a personal relationship between scholars. The amount of knowledge transferred from author to reader is likely less than when co-authoring a paper, where cooperation, discussion and information sharing are important aspects. This gives the approach an advantage over the more common citation studies as:

> This type of network is not only depicting an academic society but also representing the structure of knowledge […]. Somewhat similar to much studied citation network, co-authorship implies a much stronger bond among authors than citation. Unlike citation networks where nodes are papers and the links between them are citations, in a co-authorship network nodes are representing authors and link between nodes implies a scientific collaboration (Uddin et al. 2012).

The information being transferred while co-authoring a scientific paper therefore represent an active transferal, or diffusion of information, through the network of researchers, as opposed to the more passive and often temporally disconnected sharing of information through reading of publications. It is because of this that "Co-authorship is the preferred indicator used to describe collaboration and co-operation in all areas of research. Such collaborative efforts, or flow, lead to publications within the formal network of scientific journals" (Gauthier, 1998).

Co-authorship therefore seems like the best way to capture the insight I aim to accomplish with this thesis, in addition to providing a new and additional analysis to the same empirical background material already used to study the themes presented in this thesis. The idea of 'flow' presented by Gauthier in the quote above represents an idea already well established within academia, namely diffusion of innovations. A term I will describe later in this chapter how it relates to the transferal of knowledge through networks. There is, however, criticism of utilizing co-authorship as a measure of collaborative activity. This discussion will be taken in the next section.

2.5 Critique of Using Co-Authorship as a Measurement of Collaboration

The advantage pertaining to sharing of information through large networks have long been acknowledged in innovation policy research, often known as heterogeneity or heterogeneous groups within the field, or rather collaboration between heterogeneous groups (Powell & Grodal, 2005). Collaboration in research is widely and in general considered to be something good that help to increase a collective knowledge base and help build the interdisciplinary research disciplines through sharing of information and ideas, i.e. diffusion of information. However, there is some critique related to how well co-authorship actually capture what it seeks to measure, i.e. collaboration between researchers.

There are some aspects of this theoretical framework that have received substantial criticism over the years. The debate whether co-authorship is in fact a valid measure for collaboration is especially important, seeing as researchers collaborate in many ways, not only through publications. A second debate is about why researchers chose to collaborate at all. We must remember that by using co-authorship data as a proxy for studying collaboration in scientific endeavors, we are implicitly assuming that the notion and action of cooperation is in some way meaningful and significant, not merely a result of, say, forced collaboration or economic incentives. Which leads us into the observed increase in co-authored papers, which might be explained by incentives and institutional structures in academia. We will discuss these three particular areas further below as they are important for the validity of this study, as well as providing us with a backlight on the results of this study. After discussing the three instances that might be problematic when conducting a co-authorship study, I will discuss the justifications for using this method despite some of its shortcomings and unresolved issues.

2.5.1 Co-Authorship as a Proxy for Collaboration

In innovation and technological development research co-authorships have been used as a proxy for collaboration for decades, after researchers noticed that the level of co-authorship was increasing. However, the method of choice has been to count the number of co-authored articles and the number of collaborators of specific researchers (Katz & Martin, 1997). These metrics are also a part of social network analysis. For instance, the number of researchers a specific researcher has collaborated with is in this setting known as degree centrality in social network analysis. And although degree centrality is an important metric in determining the importance of any given node, it is subject to some weakness, luckily other metrics allowed through the use of social network analysis help create a fuller, and more detailed picture of the composition of a scientific field.

The underlying assumption that co-authorship is a proxy for collaboration, and in turn sharing of information is subject to some criticism, ranging from the problem of researchers being credited as authors of a paper for purely social reasons, or that co-authorship is only one of several ways in which scientist can collaborate and share information (Katz & Martin, 1997). Co-authorship is therefore not a perfect measure for collaboration among scientist. It is nonetheless a measure that captures information on collaboration, and additionally in a highly quantifiable way, and thus enabling researchers to use quite large datasets that would be difficult to utilize by using other methods that enables researchers to compare co-authorship levels between academic fields, journals, time periods etc.

A perhaps striking issue with using co-authorship as a way to measure collaboration is that papers are not the only product that is created through a research project. There are other outputs as well, including patents, which is often an output in the natural sciences and computer science. Or the output might be absolutely nothing at all. Researchers might collaborate on something that produce no scientific output, but might produce relationships or new ideas among researchers (Melin & Persson, 1996). The fields under investigation in this study is not really prone to producing patents, so this will most likely not be an issue here. However, collaboration that have led to some other document like a book, or absolutely nothing will not be picked up by the data in this survey. This is an obvious flaw in the data and more generally the method and framework of using co-authorship data to study research collaboration.

Thus, we run the risk of not being able to capture all the collaborative efforts present in a field. In fact, we are definitively not capturing all the collaborative efforts that exist. On the other hand, we also run the risk of measuring collaboration that does not exist. This is related to researchers getting their names on papers due to purely social reasons like authoritative figures demanding to be recognized as one of the authors of a paper that individual did not contribute to (Melin & Persson, 1996).

There is therefore two opposing and equally skewing ways that co-authorship is not a perfect measure for collaboration. On the one hand, there is a risk of capturing and measuring collaborative activities that did not truly happen, as well as not being able to measure collaborative activities that did happen, yet did not result in a research paper output. Thus, co-authorship studies as this should not be seen as an ultimate measure of collaborative activities, but more as a rough investigation, because of the factors discussed. On the other hand, it is difficult to estimate the impact these factors potentially have. A co-authorship study normally deals with vast amounts of articles at the same time. Thus, the indication of collaboration should still be quite good. However, as Melin and Persson notes: "We will simply have to accept a certain level of uncertainty" (Melin & Persson, 1996). This uncertainty is unfortunately not feasible to reduce.

We have discussed some of the issues relating to if the method of using co-authorship data is a good measurement of collaboration. We have not, however, tackled the issue of why and under which circumstances researchers choose to collaborate, something we will do in the next section.

2.5.2 Why do Researchers Collaborate?

While we have discussed advantages and disadvantages of measuring collaboration through co-authorship, we have yet to discuss why researchers collaborate in the first place. It is important to quickly discuss this as this research aim to uncover how researchers collaborate across fields, this will help us later in the analysis to shed light on the findings of how three fields are tied together.

Several reasons for collaboration between scientific personal has been proposed, and as many reasons for them choosing to collaborate in the form of co-authorship. It is therefore important to first note that scientific collaboration takes several forms, as discussed above. Co-authorship of papers in journals is one of them. As Smith notes: “Nothing short of a complete description of the kinds of relationships and activities of all persons concerned in the final product would give an approximation of the amount of group effort going into the papers presented” (Smith, 1958). Thus, we have two different possible ways to discuss collaboration: one where we look at why researchers collaborate in a broad sense, meaning all the activities and interactions they have that constitute a form of collaboration. This could include relations such as training new PhD students, delivering speeches, attending seminars and chatter in a break room. This form of relation certainly is withheld within the term collaboration.

On the other hand, we can discuss a more stringent version of collaboration, namely co-authorship. Although not the only form of communication and relation among researchers, it is more fruitful for this. A reason for this, which will be discussed later in the methodology chapter that pertains to social network data, is that many of the forms of collaboration that fall under the broad description, seem to not be readily codeable and therefore highly difficult to get valid data on. Imagine for a moment that we wanted to collect data on all research collaboration. We would then have to measure, in a way, who talks to whom how often and about what, which is simply infeasible. This part of the thesis will therefore discuss the reasons behind collaboration in the form of co-authorship.

Apart from this we can observe that researchers tend to organize collaboration themselves, meaning that they are usually not in any way forced or directly pressured into collaborative activities with other researchers (Melin & Persson, 1996). This would seem to suggest that collaboration is an activity that researchers value in some way. Be it personal, or structural. A slightly different yet related possible incentive, more related to personal gain is that researchers chose to collaborate on papers due to the fact that they seek to increase their scientific output: “The co-authorship of papers stems from the researchers desire to increase their scientific productivity, both in quantity and in quality” (Acedo, Barroso, Christóbal, & Galán, 2006). A slightly different version of this argument is presented by Price when he notes that co-authored papers “arises more from economic than from intellectual dependence and […] the effect is often that of squeezing full papers out of people who only have fractional papers in them at that particular time" (Price, 1987). Which would indicate a need for researchers to publish extensively and see collaboration as a good way of increasing one's scientific output. On the other hand, it has been argued that co-authorship is in fact a strongly social and intellectual process, motivated purely by these factors (Edge, 1979). Yet another factor contributing to researchers choosing to collaborate is the simplicity of actual cooperation:

> In addition, spatial proximity seems to encourage collaboration since it tends to generate more informal communication. The closer two potential collaborators are, the more likely they are to engage in informal communication. This is consistent with the result of a recent study which shows that co-authorship decreases exponentially with the distance separating pairs of institutional partners (Katz & Martin, 1997, p. 5).

As the geographical distance between scholars increase, the amount of collaboration decreases. There are then many different ways explaining why collaboration happens, and the behavior is likely motivated by several factors, not just one. To answer the question of why researchers collaborate, it is then reasonable to put it in some perspective. Studies have shown that the amount of collaboration in the form of co-authorships is increasing, meaning that researchers have chosen to collaborate more often than before. To investigate what has spurned this increase, we will also indirectly investigate why cooperation is taking place. In the next section I will discuss the increase in co-authorship cooperation.

2.5.3 Growth in Co-Authored Papers

Many scholars have observed a dramatic growth in the number of co-authored papers, noting that the first 50 years of the last century, co-authored papers were in fact quite rare, most scholars opting for single authored papers (Acedo et al. 2006). This theme has been studied on par with the reasons why scholars collaborate in the first place on the assumption that some underlying effect has driven the growth of collaboration among scientists. It is important to get an understanding of the reasons for the massive growth in co-authorship as the underlying assumption in the network analysis part of this thesis is founded on the idea that collaboration represents something more than a mere economic, or other, incentive. There are many proposed reasons for this, ranging from the institutional, economic, technological and scientific. In this section I will discuss some of these theories as to why this sudden increase has happened, and what might be the underlying processes that drive it.

An early observation regarding collaboration was that researchers that are more theoretical in their approach to scientific work, as opposed to those more inclined toward the empirical, tend to publish fewer co-authored papers (Smith, 1958). This observation has been studied by other researchers, and is now generally accepted as a trait of theoretical scientific endeavor (Katz & Martin, 1997). If taken as an assumption, this would seem to indicate that the growth in collaboration as measured through co-authorship is linked with an increase in empirical work as opposed to theoretical work. It also seems to imply that the rise in co-authorship is dependent on intellectual dependence. This, however, seems an unlikely explanation, as there is little reason to believe that empirical research has been growing on account of theoretical work. There must therefore be an underlying driver of this rapid development.

A far more reasonable explanation lies in the institutional framework surrounding research collaboration. It has been suggested that collaboration is a trait associated with what is known as 'big science' (Pao, 1992), which in turn is related to the use of complex and complicated devices and instruments that is needed for many of the big science projects (de Solla Price, 1963). This is in turn connected to the complexity of big science research, and the need for specialized competence to be able to perform highly complex research as no one individual might possess all the necessary skills and functions required for the work.

Consider the particle accelerator at CERN as a prime example of big science, where the skills needed to perform an advanced study necessitates a small army of specialized scientists, engineers and management all collaborating on a specific and highly complex research objective. This would be a principal example of how the numbers of authors increase with the complexity of the study.

However, this is in contrast to studies showing that more 'basic' fields produce more co-authored papers (Frame & Carpenter, 1979). There is therefore evidence supporting both sides of the claim.

If we again allow ourselves to consider CERN as an example, it becomes clear that there are tremendous economic benefits of in collaboration among scientists from different fields, institutions and countries, as the economics behind building several particle accelerators is quite simply infeasible. The economics of scientific work therefore seems to be, at least in some cases, a highly relevant parameter of collaboration.

Pao (Pao, 1992) notes that: “Although collaboration has existed since the beginning of science, it was noteworthy that its significant growth rate in recent decades was matched by the exponential increase in research dollars”. This might mean two things: Either the rise in research funding has spurred collaboration when researchers afford the costs related to collaboration. This might seem reasonable as collaboration is generally considered a good policy instrument to increase scientific quality and quantity. On the other side, it might entail that increased collaboration has led to more research funding as researchers apply for more funding to support their collaborative activities.

Both interpretations seem to hinge on the assumption that collaboration is something that policy makers are willing to invest in. We do not have direct evidence for this, but seems like a plausible assumption seeing as collaboration is widely regarded as a uniquely positive endeavor among research policy makers. In fact, the last couple of decades have seen collaborative activities becoming a prerequisite for research funding applications (Katz & Martin, 1997). This is an indication that science policy makers encourage collaboration, and that the institutional and economic incentives behind collaboration are of great importance when discussing why researchers collaborate.

Katz and Martin (Katz & Martin, 1997) discusses the reasons for scientific collaboration thoroughly in a paper titled ‘What is research collaboration?’ This paper may be referred to for a more detailed discussion, which also contains the points discussed above. The list of potential factors involved in the observed increase in research funding is long, and there is as of yet no consensus on which of these factors contribute the most. They therefore present these findings, or reasons, before concluding that, “collaboration is an intrinsically social process and, as with any other human interaction, there may be at least as many contributing factors as there are individuals involved” (Katz & Martin, 1997).

The reasons behind the dramatic increase in collaboration through co-authorship can therefore be difficult to grasp, and will likely vary between research projects, as well as between individual researchers. Whatever the cause for collaboration on a specific paper, it is clear that they represent a social bond between researchers in one or more aspects. Luckily for us, this social bond allows us to study its structure using social network analysis. This will be discussed in detail later.

2.5.4 Justification for Using Co-Authorship Data

As discussed in this chapter there is some criticism of using co-authorship as a valid measure of collaborative activities. Despite these critiques co-authorship has been used as a proxy for collaboration for a long time: “For decades the multiple-author publication, frequently referred to as a co-authored publication, has been used as a basic counting unit to measure collaborative activity” (Katz & Martin, 1997). This might be related to the fact that data on co-authorship is readily available for researchers, both for statistical and econometric modeling, as well as social network analysis (Uddin et al. 2012).

As discussed above, the shortcomings of the method and framework are small compared to the amount of data that can be used in the analysis. Also, there is good reason to believe that co-authorship is a good measure of cooperation and “[t]here is general consensus that the observed growth in multiple-authorship is evidence of an increase in collaboration”(Katz & Martin, 1997):

> Although the assessment of collaboration using co-authorship is by no means perfect, it nevertheless has certain advantages. First, it is invariant and verifiable; given access to the same data-set, other investigators should be able to reproduce the results. Secondly, it is a relatively inexpensive and practical method for quantifying collaboration. Furthermore, the size of sample that it is possible to analyze using this technique can be very large and the results should therefore be statistically more significant than those from case-studies. Finally, some would argue that bibliometric studies are unintrusive and indeed non-reactive – that is, the measurement does not affect the collaboration process (Katz & Martin, 1997).

Both the topics discussed in this chapter and the prevailing convention in academic research to treat co-authorship as a valid measure of cooperation gives us a good case to use the framework and method discussed in this thesis. Although there are shortcomings, the benefits of the approach far outweigh the negative aspects. Indeed, I would argue that among the available approaches to study cooperative efforts in academia, co-authorship studies are by far the best suited seeing as all other approaches have severe shortcomings without the added benefit of looking at social links, reproducibility, the sheer amount of data one can analyze at the same time, its non-reactive nature etc.

However, to investigate the phenomenon of co-authored papers in detail one can also approach it more holistically, through an examination on an individual level the motivation behind collaboration through interviews and smaller case studies of the fields of innovation, entrepreneurship and STS. This paper will thus contribute to the understanding of diffusion of information through co-authorship networks, and further investigate the level of similarity between the fields of innovation studies, entrepreneurship and science and technology studies. The fundamentals behind the idea of diffusion of innovation will be discussed in the next section.

2.6 Diffusion of Information

The process of information sharing through a community of researchers can be thought of in terms of diffusion of innovations. Diffusion is defined by Everett M. Rogers in his seminal work 'Diffusion of Innovations' as:

> […] the process in which an innovation is communicated through certain channels over time among the members of a social system. It is a special type of communications, in that the messages are concerned with new ideas. Communication is a process in which participants create and share information with one another in order to reach a mutual understanding. This definition implies that communication is a process of convergence (or divergence) as two or more individuals exchange information in order to move towards each other (or apart) in the meanings that they give to certain events (Rogers, 2005, pp. 5-6).

The sharing of ideas between researchers is therefore diffusion of new ideas. The diffusion of information through a network of researchers thus represent knowledge that is new to the recipients. 'Objectively' new knowledge is of little importance in this context, as it is the perception of new information that knit the network together. For example, a researcher in economics might perceive information diffused to him or her through a historian as 'new', although it might be a long established principle is history and vice versa.

In innovation research the term tacit knowledge is widely used to signify a type of knowledge that is difficult to codify and diffuse, and can be exemplified by the simple statement that we often know a lot more about something than we are able to express orally or in writing (Polanyi, 1956). Often exemplified through the notion of teaching someone to ride the bicycle. It would be extremely difficult to teach bicycle riding through written material, and relatively easy to teach through practical demonstration. This type of knowledge is difficult to codify, but possible to diffuse though other methods, such as working together and gaining shared experience.

Explicit knowledge, on the other hand, is knowledge that is easy to codify and therefore easy to publish (Powell & Grodal, 2005). I therefore argue that citation networks are better if you want to study the diffusion of explicit knowledge in a network of scholars, while co-authorship is better able to capture the diffusion of tacit knowledge. Katz and Martin (Katz & Martin, 1997) also notes this while discussing the benefits of collaboration on research:

> A second closely related type of benefit is the transfer of knowledge or skills. As noted earlier, it can be time consuming for an individual to update their knowledge or to retain. Furthermore, not all the details concerning new advances are necessarily documented. Much of the knowledge may be tacit and remains so until researchers have had the time to deliberate and set out their findings in a publication. Frequently, considerable time elapses before the knowledge appears in written form. Collaboration is one way of transferring new knowledge, especially tacit knowledge.

The parts of this thesis directly related social network analysis of the three fields are thus aimed towards the tacit knowledge diffusion in the network present through collaboration on scientific work. This will give us a good understanding of how knowledge in fields related to innovation and technological development is diffused through the network, and in turn give us valuable insights into how this impacts policy decisions about science and technology development related politics, and therefore a key driver of economic growth and societal change.

In the next section of this chapter I will quickly summarize the discussions so far in this chapter and on that basis conceptualize and detail the research motivation behind this thesis, as well as explaining the main research questions I seek to answer.

2.7 Research Question and Motivation

The motivation behind this thesis is to expand the literature on innovation studies, entrepreneurship and STS by investigating which research themes are similar, and dissimilar, between the three fields, and how the academic fields shares information though its co-authorship network, and importantly investigating who is important for the diffusion of information in the network and the positioning of the most important researchers. Seeing as the fields are all recent addition to academia we might expect that the fields are highly clustered when analyzed as a whole. Several groups of researcher might not collaborate all that much with others, but choosing to cooperate with members that shares a similar educational background or thematic interest. This might also be the case when analyzing the fields separately, seeing as they are all relatively new additions to academia. This investigation will be driven by global and local measures of the network, many of them related to the centrality of nodes. I will investigate these themes through the four concise research questions as explained in the introduction, which are:

- *How is information diffused through the networks of researchers within their fields, and are there any systematic differences in the flow of information between them?*
- *Have the fields developed into three clearly separated ones when analyzed on an individual researcher level?*
- *Who are most important for the diffusion of information through this the networks?*
- *Which social network method is most fitting to measure cooperation in science and what type of challenges do these have when studying emerging fields?*

Additionally, the paper is concerned with the current structure of the scientific field. Through investigation of clustering of research collaboration, I aim to uncover which researchers collaborate with each other, and thus where diffusion of information from different academic disciplines occurs. A highly possible outcome of this investigation is that there are a few sets of researchers than span between the fields, while most researchers prefer to publish with authors already highly situated in a particular field. The reasoning behind this investigation is that if there exist a group of tightly knit researchers, all specialized in similar fields and currently conducting research on innovation and technological development, one of these researchers that connects to a different community might be an influential carrier of information to this other group, and visa versa.

The research motivation of this thesis is therefore grounded in the fact that “In order to make rational decisions, public policy makers need to have a firm understanding of scientific and technological activities. Bibliometric indicators provide the only overall picture of the scientific output of a country” (Gauthier, 1998). This is perhaps now more important than ever seeing as the literature on innovation and technological development is steadily increasing and the contributions of these academic fields on the policy makers (Fagerberg, 2005).

3 Methodology

There will be several methods employed throughout this thesis. In addition to the bibliometric network analysis there will also be analysis of unstructured data, sometimes also referred to as text mining. Text mining applications allows us to utilize the large amount of unstructured data present in the meta-data on scholarly articles to analyze the content of the articles, not only looking at the structural aspects of them. This additional analysis of the data represents a novel approach that compliments the social network analysis that is the main part of this thesis. Analysis of unstructured data has not been undertaken in the earlier research that we build upon and will provide us with new and meaningful insight into what specifically separates the three fields of innovation studies, entrepreneurship and STS.

The methods employed in the text mining part of this thesis differs substantially from the social network approach, primarily in that it is a qualitative, rather that SNA's quantitative, measure. However, unstructured data in large quantities is possible to analyze quantitatively. Further, by combining the two approaches the results will be more holistic than simply using one of the two approaches. The fact that the data for both the text mining and social network analysis is withdrawn from the same raw data set allows us to gain a new level of insight rooted in the two methodologies complimentary nature.

This paper will anyhow mainly rely on social network analysis (SNA), with text mining providing further insights on the subject as well as placing the social network analysis in the context of content. SNA methodology has been used in many diverse fields and disciplines (Wasserman & Faus, 1994; Barbási, 2002; Watts, 2001). SNA is based on the notion that social relationships can be presented trough graph theory, which is then subject to formal analysis. In the next chapter I will first outline the methods used for text mining followed by a brief outline of SNA and its applications, as well as the measurements I will use for this paper.

3.1 Text Mining and Unstructured Data Analysis

Text mining can be defined as "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources". (Hearst, 2003). Text mining or text analytics is a research and business intelligence method and the basis for the analysis of keywords in this thesis.

Unstructured data makes up most of relevant information that can be used for research, analysis and business intelligence. It's widely held-that as much as 80 % of business and research relevant data is unstructured, through what is called the 80 percent rule (Tan). The amount of information behind the 80 % estimate is thought to have a high value seeing all the potential insight and information possible to retrieve from it, if thorough analysis is undertaken.

Most of the unstructured data is in the form of free text. Text mining is implemented by firms to draw on the abundant text material in their data warehouses. "(…) [A]nalysts and consultants have begun leveraging text mining capabilities to sift through vast amounts of textual data with the aims of creating usable forms of business intelligence, noting trends, identifying correlations, and researching references to specific transactions, corporate entities, or persons" (Feldman & Sanger, 2007, p. 273). Making text mining a relevant and important research methodology to be able to derive new and useful information in research.

The data systematization and analysis follows a content analysis methodology, which can be defined as: "Content analysis is a research technique for making replicable and valid inferences from data to their context" (Krippendorff, 1980). Content analysis can be utilized for both structured and unstructured data and can be applied in an inductive and a deductive way, with differences in how the two should be conducted. Whether an inductive or deductive approach is most useful depends on the motivation of the study being undertaken.

An inductive approach is a movement from the specific to the general, and a deductive approach is a movement from the general to the specific. A deductive approach to content analysis is therefore useful when wanting to test a theory, whilst an inductive approach is best suited for situations where the research is more exploratory (Elo & Kyngäs, 2007). In my analysis I will utilize an inductive approach where the categories are derived from the text itself, as opposed to already being established. Particular events will thus be observed and then combined into broader categories, meant to be accessible and effective for analysis. Meaning that I will not set a specific number of keywords or themes that I will try to fit the data into at a later stage. Rather I will allow the data itself to shed light on which keywords are the most prominent, and therefore the most researched themes.

A further goal with the analysis of unstructured data is to create insightful graphics that can illustrate the findings of the fairly large amount of text being analyzed in a clear way. Unstructured data is often difficult to present, seeing as there are few quantitative measurements with substantial meaning attached to them, in addition to being difficult to graph or model mathematically. In text mining, interactivity of analysis results is a central part of the process. Since such text-mining software includes numerous visualization options (Feldman & Sanger, 2007, p. 189) there are several advantages in using visualization techniques as opposed to quantitative or query listings. Below is a set of advantages described by Feldman and Sanger (2007, p. 191):

- Concision: the capability of showing large amounts of different types of data all at once
- Relativity and Proximity: the ability to easily show clusters, relative sizes of groupings, similarity and dissimilarity of groupings, and outliers among the data in query results
- Focus with Context: the ability to interact with some highlighted feature while also being able to see the highlighted featured situated in some of its relational context
- Zoomability: the ability to move from micro to macro quickly and easily in one big step or in increments
- 'Right Brain' Stimulation: the ability to invite user interaction with textual data that is driven not only by premeditated and deliberate search intentions but also as a result of intuitive, reactive, or spatially oriented cognitive process for identifying interesting patterns.

The number of advantages and the importance of visualization to derive meaningful insights from a large set of unstructured data will make visualization a key component of the analysis in my thesis. However, there are some difficulties with text mining which we will discuss in the next part.

3.1.1 Methodological Difficulties

There are some difficulties with using free text as data for analysis. The difficulties arise because of the discrepancy between how a human being interprets text and the way a computer interprets the same text. In this part of the chapter we will discuss the implications of two possible difficulties that may arise as a result of this discrepancy: how natural language is created to be read and understood by humans and not computers, as well as the persistent, yet highly solvable case of "stop words".

Natural language

Analyzing text can be somewhat difficult due to the fact that the free text is written in natural language. Natural language, the way people write and speak, is not easy for a computer to understand. At least not the context of the written text: "text is written for people to read. We do not have programs that can 'read' text and will not have such for the foreseeable future. Many researchers think it will require a full simulation of how the mind works before we can write programs that read the way people do" (Hearst, 2003). This thesis will analyze the keywords associated with the articles. Where computers have a distinct and large advantage over humans to process large amounts of text in mere minutes and seconds, they do not share people's ability to make sense of natural language:

> Natural language has developed to help humans communicate with one another and record information. Computers are a long way from comprehending natural language. Humans are able to distinguish and apply linguistic patterns to text, overcoming obstacles (such as slang, spelling, variations, and contextual meaning). Computers do not handle them easily. Meanwhile, although our language capabilities allow us to comprehend unstructured data, we lack the computer's ability to process text in large volumes at high speed (Weiguo Fan & Wallace, 2006).

Luckily, keywords are simple words in a list not meant to be read as sentences where part of the meaning of the individual words are derived from their context, so this issue is not a problem in relation to the keywords. Specifically I will analyze which words appear most often in the keywords. This methodological approach to analyzing unstructured data is often referred to as categorization. This method allows us to identify the main themes of the text by treating the words in the collection of documents as a "bag of words", meaning that the algorithm does not attempt to analyze the contextual meaning of the words in any way. This method is generally considered efficient for finding useful insights from large text sources: "Surprisingly it is sufficient for many applications to simply count the number of occurrences of each word in a document, the so called bag–of–words representation" (Leopold, May, & Paaß, 2005). The bag-of-words metaphor stems from the fact that the order of the words are insignificant to the result of the analysis because the algorithm treats only words, not sentences and could just as well have been drawn from the dataset in a random order (Weiguo Fan & Wallace, 2006). This provides a relatively good approximation of the most important themes due to the shape of the frequency distribution of words that usually exhibits traits of a power law function with an extremely long tail. This can be further conceptualized as: "Preprocessing is concerned with the elimination of textual information which is irrelevant or even misleading to solving the subsequent data mining task. As a rule of thumb half of the words occur only once even in a large text corpus of some million running words" (Leopold, May, & Paaß, 2005). This is typical in text-mining and also allows us the added benefit of removing half of the words from the analysis without any difference in the results of analysis if the data processing time proves to be an issue.

By utilizing this approach, we will be able to find the most common themes and keywords in the journals associated with the three fields by ordering the words by their frequency distribution, thus avoiding the difficulties associated with computers attempting to read text, but loosing the contextual meaning that words create in sentences. However, this will inevitably lead to words such as "the" and "it" to rank highly in the distribution. As these words contain no meaning, at least not in terms of research themes, we will be forced to exclude the so-called stop words.

Stop Words

As mentioned above, stop words refer to words and phrases that is scattered in natural language, but that contains no significant meaning. Common stop words might include – for the English language – words such as 'the', 'it', 'such' etc. E.g. common words that are not particularly useful for discovering the content of text. This pose a problem as the words are often found to large extent and thus would be very much present in the frequency distribution. The principle is explained by Leopold May and Paaß, while also explaining the solution:

> The simplest method for the removal of uninformative words is to use a predefined list of stop words, and to delete all words in the text that match an element of the list. Stop word lists typically consist of function words (articles, pronouns, and conjunctions). The problem of stop word lists is that they may be inappropriate to the corpus or the task in question. In a corpus of texts on computers the word 'computer' will probably be equally distributed amongst the documents and thus fairly uninformative. In such a case the word 'computer' should be treated as a stop word (Leopold, May, & Paaß, 2005).

By treating these words as noise and simply removing them from the analysis will simply and effectively solve this issue. Luckily, text-mining packages for the R environment contains such pre defined lists of common stop words for the English language, making this part quite easy to overcome. However, there are words that in a context specific environment should be considered to be treated as stop words, if not we run the risk of generating highly generic overviews of the themes being studied by the different fields. In our case words such as ‘technology’, ‘innovation’, and ‘R&D’, might prove to be so common that we underestimate the true differences of the fields by including them. Meaning that the three separate frequency distributions will be so similar that they their differences convey no substantial meaning.

There is no apparent and easy way to solve this potential issue of when and if certain words should be treated as stop words. However, as it is uncertain that this will pose an issue, and as this is an exploratory study, we will investigate solutions for this problem if it occurs in the analysis. But, as discussed, this process of exploring the data is best left to the analysis part and we will end the discussion here and rather bring it back if it proves to be an issue later in the analysis.

3.1.2 Using Text-Mining as a Method

Through this discussion of the benefits and potential difficulties with text-mining for the sort of application I am pursuing in this thesis, it seems clear that the benefits are potentially large, and the disadvantages are mostly easily dealt with. The use of text-mining methods for the specific research question in this thesis will help shed new and valuable insights into what separates the three fields of innovation studies, entrepreneurship and STS.

3.2 Social Network Analysis

Social network analysis has developed from a fairly non-technical way of studying the structures of social interaction, to an increasingly technical field dependent on mathematics and specialized computer software and algorithms. The influence and popularity of social network analysis have also increased through the years and especially after the rise of social network websites, such as Facebook and LinkedIn, and is today used by scholars from many different disciplines such as sociology, physics, management and medicine (Scott, 2013). Additionally, its methods have been able to provide meaningful answers to comprehensive questions in social sciences like sociology, economics and psychology (Borgatti et al. 2009).

Social network analysis began in sociology, and is based on the premise that social interaction between nodes can be analyzed trough a graph. This allows for the usage of graph theory to explain social phenomena (Wasserman & Faus, 1994). Graph theory is a set of formal mathematical formulations that can be used to analyze a set of nodes with lines, or edges, connecting them, not to be confused with what is normally termed a graph (Scott, 2013).

Especially linear algebra and matrices are heavily used in network analysis (Borgatti et al. 2009). A node represents actors, which can range from individuals, large companies or events. Edges between nodes represents that there is a relationship between the nodes. In this article the nodes are researchers, and the terms node, actor, researcher and scholar will be used interchangeably. Accordingly, in this paper, the edges between nodes represents that the researchers in question have co-authored a paper together.

As an example, consider a graph (N, g), that consists of a set of nodes N=(1, 2, … , n) and a *n x n* matrix *g* with information on the connections between the nodes, for instance a friendship network between five individuals:

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

The network above is represented through an adjacency matrix, meaning that it gives information on which nodes in the network that are connected, or adjacent to each other. The graph (N, g) is a binary undirected graph. A binary graph is where the value attached to the connection can only be defined as 1 or 0, in this instance friendship. Undirected means that any two nodes i and j have the same relationship to each other, either they have a relationship or they do not, i.e: $g_{ij} = g_{ji}$. Thus, half of the information in the graph is redundant, yet convention dictates it to be written in the matrix.

Further, the diagonal from the upper left corner down to the lower right are all zeros. In the case of there being non-zeros in this diagonal, we call it a loop. A node may in some instances be connected to themselves, as in citation studies where scholars cite earlier articles they themselves have written. It is, however, of little substantial meaning to propose that the people in a friendship network, as the one above, are friends with themselves, and the network is therefore devoid of loops.

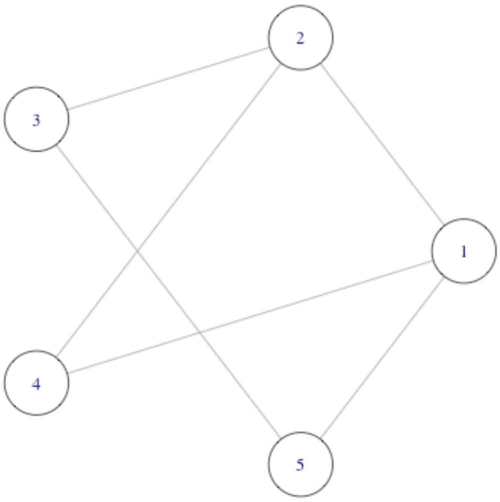

Figure 1 – Example Network

The matrix is therefore the structure that allows mathematics to be used in its analysis. But, when the number of nodes is large, the matrix becomes a troublesome carrier of information to readers because of its sheer size. It is therefore common to illustrate the network through visualization, as in figure 1 below. The visual "map" of nodes and their connections contain the exact same information as the matrix g, yet easier to comprehend. It is important to note that the spatial distance between nodes is almost purely arbitrary and carries no substantial meaning (the spatial distance between nodes are dependent on the algorithms used to visualize the network), yet the algorithm will tend to position nodes that are connected closer than to nodes it is not connected to.

The lengths of the edges between nodes are therefore also subject to the arbitrary spatial positioning of nodes, and thus do not carry any substantial meaning other than simply denoting an edge between the nodes. So, as opposed to a regular map, the visual network gives us information on the connection of relationship between the nodes, not on their spatial distance (Scott, 1988).

We can identify specific nodes as well as all connections between them in the visualization, thus providing an intuitive and informative way to get an overlook on the network. Visualizations of graphs are an important part of social network analysis, and one I will rely heavily upon in this paper. The graph also allows us two levels of analysis in terms of statistical metrics: A macro or global level that captures information on the network as a whole, and a micro or local level that captures information about specific nodes in the network and their centrality in it. The metrics used in this paper will be discussed shortly below.

3.2.1 Metrics

The study aims to explore the clustering of specific researchers and I have chosen an undirected, binary approach to the network analysis. The nodes in the network therefore consists of researchers where edges between any given nodes *i* and *j* are present if one or more article have been co-authored by authors *i* and *j*. Further, the edges can only have the value 1 or 0, making it a binary graph.

The metrics used for binary, undirected graphs are largely the same as that of other variants, with small differences in the calculation of particular metrics. In this paper, I will utilize four different local centrality measures, in addition to their normalized values: degree centrality, betweenness centrality, closeness centrality and eigenvector centrality. Additionally, two global measures will be used: graph density and transitivity. All metrics are explained below.

Degree Centrality

Arguably the simplest way to measure the importance of any given node in a network is to simply count how many vertices that connect to the node. In this instance, it would be to simply count the number of researchers with whom any given researcher has co-authored an article. Authors with a high degree centrality are thought to have a more central position in the network and therefore being more disposed to influence others. This is known as degree centrality and is simply calculated as:

$$C_D(n_i) = \sum_{j=1} x_{ji}$$

As the degree centrality metric is dependent on the size of the network g, the measure is poorly suited to be compared between networks or over time. A normalized version of the metric has therefore been developed and is a measure of the proportion of nodes that are connected to a specific node:

$$C'_D(n_i) = \frac{C_D(n_i)}{g-1}$$

where g is the number of nodes in a network, and (g-1) is the number of nodes the node n_i can connect to in an undirected network, so that $0 \leq C'_D(n_i) \leq 1$ and is suited to be compared across networks of varying sizes or the same network over time.

Betweenness Centrality

Betweenness centrality is a metric that seeks to capture aspects of the diffusion of information among the nodes in a network, and determining which nodes are important for the transmission of knowledge and information. The underlying assumption is that the diffusion of information is more likely to go through actors that lies on the shortest path between any node i and j. Information will spread though the network taking the shortest path possible, if a node k is on the shortest path, or geodesic, for many pairs of nodes j and k, the node i will be important for the spread of information through connection other researcher together. The metric is therefore calculated by measuring the amount of incidents where a node i lies in the geodesic between all nodes j and k:

$$C_B(n_i) = \sum_{j<k} g_{jk}(n_i)/g_{jk}$$

Where where $g_{jk}(n_{ij})$ is the number of geodesics linking nodes j,k that contains the actor i. In the case of there being more than one geodesic between a set of nodes the estimated probability for any one of them to be used is thus $g_{jk}(n_{ij})/g_{jk}$, where each geodesic is as likely to be used as the other. The betweenness centrality score is therefore the sum of these probabilities.

As with degree centrality this metric is dependent on the size of the network and the following normalization allows networks to be compared:

$$C'_B(n_i) = \frac{C_B(n_i)}{(g-1)(g-2)/2}$$

where g is the number of nodes in the network so that $0 \leq C'_B(n_i) \leq 1$.

Closeness Centrality

The closeness centrality metric captures how close a node is to all other nodes in the network, where a node is central when it is close to all other nodes in the network. This measure can therefore be thought of as a measure on how long it takes for information to spread from the nodes to other nodes in the network (Freeman, 1979). The metric is calculated by summing the length of the geodesics to all other nodes in the graph, and then taking the inverse:

$$C_C(n_i) = \left[\sum_{j=1}^{n} d(n_i, n_j)\right]^{-1}$$

where $d(n_i, n_j)$ is the length of the geodesic between nodes i and j. Since this is dependent on the size of the network g as well, it is more common to use the average length of geodesics to all other nodes:

$$C'_C(n_i) = C_C(n_i)(g-1)$$

and thus making it another metric where $0 \leq C'_C(n_i) \leq 1$.

Seeing as the closeness centrality metric is undefined in an unconnected graph, because the length to disconnected nodes is infinite and thus making the average length to all nodes infinite, this metric is inapplicable to networks consisting of several unconnected sub graphs. Closeness centrality measures must be undertaken in a connected graph, meaning that there is a path between all nodes i and j, and has a sizeable disadvantage with this particular metric.

Eigenvector Centrality

Eigenvector centrality is a measure similar in nature to Google's well-known search algorithm. The metric takes into account the centrality of a given node i, but also the degree centrality of the nodes connected to it. The mathematics and the ideas behind this measure is more intricate and elegant than other measures, and therefore somewhat more tedious. For more information on the mathematics of Eigenvector centrality refer to Phillip Bonachic's paper in centrality (Bonachic, 1987) and Matthew Jackson's book on social and economic networks (Jackson, 2008). The basics is however:

$$\lambda C_i^e(g) = \sum_j g_{ji} C_j^e\,(g)$$

where λ is a proportionality factor, and the nodes measured are dependent on the centrality of the nodes that connects to it.

Sergei Brin and Larry Page, the founders of Google, explain the intuition behind this way of measuring centrality (Brin & Page):

> [W]e have taken on the audacious task of condensing every page on the World Wide Web into a single number, its PageRank. PageRank is a global ranking of all web pages, regardless of their content, based solely on their location in the Web's graph structure.
>
> Using PageRank, we are able to order search results so that more important and central Web pages are given preference. In experiments, this turns out to provide higher quality search results to users. The intuition behind PageRank is that it uses information which is exernal to the Web pages themselves - their backlinks, which provide a kind of peer review. Furthermore, backlinks from "important" pages are more significant than backlinks from average pages. This is encompassed in the recursive denition of PageRank.

The PageRank algorithm, named after its inventor Larry Page and only a pun on the term web page, is somewhat different from the Eigenvector centrality measure, but conceptually it remains the same. In a co-authorship setting the intuition behind the metrics is that a central researcher who co-authors a paper with another researcher gives credibility to the quality of the other researcher in choosing to cooperate with him or her, under the assumption that central researchers are of a certain quality themselves. The Eigenvector centrality of a node is thus a reciprocal process where the centrality of a node is dependent on the centrality of the nodes that connects to it, and they in turn are dependent on the centrality of the nodes that are connected to them, and so on. Researchers with a high Eigenvector centrality will therefore be connected to many others with high scores. However, this implies that nodes with high Eigenvector centrality will often be found in very dense substructures of the network.

Graph Density

Graph density is a global metric on how well connected the graph is. There is a finite number of possible connections in a binary, undirected graph with no loops which is: g(g-1)/2. By calculating a ratio of the present connection between nodes in a network we get an impression on the density of the graph and, in the case of co-authorship networks, the level of cooperation within the field. The ratio is calculated as:

$$\Delta = \frac{L}{g(g-1)/2} = \frac{2L}{g(g-1)}$$

where Δ is the notation for the density and L is the number of lines present in the network. The ratio ranges from 0 in the case of a fully unconnected graph, and 1 in the case of a graph where all possible links are present.

Transitivity

Transitivity, which is sometimes called the clustering coefficient, measures the probability that the adjacent vertices of a vertex are connected. When calculating the transitivity, we therefore look at all instances where two edges spring from the same node, and then look at how often those two nodes are connected (e.g. ij and ik involve node i, how often, then, is jk connected). The formula is formulated by Barrat et.al. (Barrat et al. 2004) as:

$$C_i^w = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{(w_{ij} + w_{ih})}{2} a_{ij} a_{ih} a_{jh}$$

The metric is useful as a way to investigate the cliquishness, or clustering, of the network by asking how often friends of friends are connected. For instance, the network (N, g) in the example above, one would ask to what extent the friends of a particular node are friends (the transitivity score for the graph is 0.33). An interesting observation network scholars have made is that most networks tend to have a dramatically higher clustering value than what would be expected if edges were distributed completely at randomly with the same size and connectivity. Studies have reported clustering values upwards of 10.000 times higher than what would be expected from purely random networks (Jackson, 2008). Illustrating the point that connections in most networks are not random.

3.2.2 Community Detection

The last investigation in this study will be analyze and identify communities of researchers within the field of innovation studies. As computer power has increased and specialized programs for the analysis of social networks have arisen, there has been a trend towards more community detection analysis. This part of the field allows researchers to classify the nodes according to their structural position and therefore draw conclusions on cliques (Fortunato, 2010). The R programming package and the igraph package in R, which I used for community detection, offers a number of community detection algorithms. There is unfortunately no standard convention on what algorithm to use, and the results of community detection analysis can therefore differ quite significantly, depending on the specific algorithm used. The algorithm used in this paper is called the Fast greedy algorithm and divides the graph into communities such that there are many edges within a community, but few between them. For a detailed account of the algorithm see Clauset et.al. (Clauset, Newman, & Moore, 2004).

4 Data and Descriptive Statistics

Data on social networks can be somewhat troublesome to gather, seeing as a random sampling is of little use when analyzing relation data. The point of social networks is that relations are not random, and it's these connections one seeks to analyze (Scott, 2013). Data collection methods normal for other quantitative methods, like econometric analysis, are therefore not applicable to SNA because of the random sampling that is the basis for other such methods. On the other side neither is collecting data on the entire network of researchers in the world due to its tremendous size. All decisions on how to limit the data sets in social network analysis are in principle arbitrary. In a co-authorship network there will always be more connections that can be included. This is referred to as the small world problem, that in turn is known as 'six degrees of separation', the theory stating that all the people in the world is connected to all the other people in the world through six links or less.

As neither this thesis nor any other study of this sort aims to analyze the entire written corpus in the sciences throughout time, there is a need for a cut-off point. This cut-off point can be more or less well founded. The next section of this chapter will describe the process of determining this cut-off point, e.g. identifying the relevant literature, before we discuss the organization of data, as well as some descriptive statistics of the raw data.

4.1 Identifying the Most Central Journals

As mentioned above there has been studies conducted on how the three fields relate to each other. Fagerberg and Verspagen were some of the first to explore this theme with the article 'Innovation studies – The emerging structure of a new scientific field' (Fagerberg & Verspagen, 2009). The scope of their investigation was, however, somewhat limited and especially STS was not represented in the study in a good way, with the main focus being on innovation studies. A few years later the research project called EXPLORE took aim at taking this analysis further. They expanded on how the fields have evolved since its beginning in the 1960's, what the important research topics were, and who are central to the field. This research also led to a special edition in the journal 'Research Policy', a well-known scholarly journal within innovation studies, entrepreneurship, and STS alike. The edition consisted of several related articles where the fields were studied separately, but also articles where all three fields were studied and compared. These articles build on each other and provide the most detailed and thorough analysis of the scholarly disciplines concerned with knowledge, innovation and technological development from a social science viewpoint so far. The group of researchers who were a part of this project recognized that the Fagerberg and Verspagen approach was fairly limited in terms of addressing the development of all three fields, and wanted to expand on this area of study. As one of the research articles explains, this was:

> […] one of the reasons for the EXPLORE project, of which this contribution is part, to launch a detailed study of the three fields of Science and Technology Studies (STS), Innovation Studies (INN) and Entrepreneurship (ENT). These detailed studies were carried out as part of a larger bibliometric study, based on a new approach in which handbooks published for the three fields played a central role (Bhupatiraju et al. 2012).

One of the major contributions of this project was the three individual studies on the state of the three fields. Each study resulted in an article that details the structure of the field in question, and explore the knowledge bases they represent. They do this by analyzing and categorizing the most central contributions to the field, the most important research environments, and they use citation analysis to achieve a better understanding of these knowledge bases. The articles in question are titled: 'Innovation: Exploring the knowledge base (Fagerberg, Fosaas, & Sapprasert, 2012), Entrepreneurship: Exploring the knowledge base (Landström, Harirchi, & Åström, 2012), and Science and technology studies: Exploring the knowledge base (Martin, Nightingale, & Yegros-Yegros, 2012).

This earlier research has been aimed at uncovering how new research fields are established and how they evolve. They have especially been concerned with how multi-disciplinary fields can evolve into more or less similar fields, and how they might evolve into separate fields as time passes. This is of interest to see how the three fields are related:

> New research fields in the social and natural sciences often originate in the interstices of established disciplines when researchers from neighboring disciplines, with differing disciplinary perspectives, realized they share a common interest. Over time, by working together, they may start to develop their own shared conceptual, methodological and analytical frameworks. This then allows them to move from publishing in journals of their 'parent' disciplines and to establish their own journals, professional associations, specialized university departments or units […] and PhD programmes to train their own researchers (Martin, Nightingale, & Yegros-Yegros, 2012).

As all of the three fields are multi-disciplinary in their approach and if they remain separate or not are best answered by looking at who collaborates with whom, collaboration in terms of co-authorship would provide a good measure on the links between the fields.

The three articles are very similar in their scope and research goals, in that they all seek to establish the most central contribution to their literatures, both the most central researchers and the most central authors. There are, however, differences in the specifics of what they wish to achieve and what their results are. One of the main research objectives is, for all of them, to identify the core contributions in their fields and this initial exploration of core journals will be the foundation for my data gathering efforts. They did this with a new method to ascertain what are the most central journals in each field and who are the most central researchers. The path to establishing the core contributions to the literatures starts with the researchers identifying a number of handbooks, as this is seen as authoritative reference works that seeks to establish an overview of different research themes within a field (Fagerberg, Fosaas, & Sapprasert, 2012).

Handbooks also have the characteristic that they seek to educate the reader in several of the research areas within the field, and they will often contain citations to what the authors of said handbooks see as the most central and important scholarly works within them. The identification of handbooks as the basis for the further analysis is therefore under the assumption that these handbooks represents core contribution in that high standing experts have written the chapters in them, and that what they cite is possible to regard as the most important works in their fields (Landström, Harirchi, & Åström, 2012).

Further, the researchers collected all the citations from the handbooks and entered them into a database; they then proceeded to analyze what the core contributions were by simply looking at what scholarly works were cited. There were of course restrictions and methodological choices taken to ensure that the findings were robust. The details of this will be left out of this thesis for reasons related to length, and I refer you to the articles themselves, especially 'Innovation: Exploring the knowledge base' for detailed explanations on how the core contributions were established (Landström, Harirchi, & Åström, 2012; Fagerberg, Fosaas, & Sapprasert, 2012; Martin, Nightingale, & Yegros-Yegros, 2012).

When the core contributions were identified and sorted they ran a form of reverse citation study through the Web of Knowledge, which allows you to input a scholarly work and then see all the other scholarly works listed in the database that reference that work through a citation. In this way they gathered an overview of their fields by looking at who cite the most important works. From this they were able to identify the most prominent journals within the fields by looking at the number of times a specific journal references the core contributions.

There is one potential problem with this method, namely that it indirectly assumes that a specific field, e.g. entrepreneurship, is built directly on previous knowledge, in that the core contributions in the studies tend to be both slightly old, and that they mostly consist of books and not journal papers. The researchers are, however, aware of that by stating that:

> bibliometric analysis is not without limitations. For example, we have to bear in mind that it is based on the assumption that research is essentially cumulative – new research is built on and cites earlier high quality foundations – i.e. a 'normal science approach' (Kuhn, 1970). However, we know that this is not the only way to communicate and organize research, particularly in new and evolving fields that are organized and communicated through 'negotiations' between researchers and policy actors (Landström, Harirchi, & Åström, 2012).

There could feasibly be important areas of study within a field that draw their inspiration and history from other sources than what was identified as important in the handbook-stage of the studies. This is, however, unlikely seeing that the authors of the individual chapters in most of the handbooks are generally regarded as experts within the field. We can therefore assume that the assumption the studies build on, i.e. that they have a fairly excellent overview of their own research fields, holds true.

Additionally, these results are obviously dependent on the assumption that the scholars who conducted the research were able to choose handbooks in a meaningful way. There might also be a slight skew in the collected data if the scholars who conducted the research had an inclination toward including their own contributions to handbooks, which is the case in some instances. This, in turn, is also unlikely in that the authors of the three papers on knowledge bases are experts in their fields and it is therefore not especially strange that they also have chapters in handbooks, seeing that they were probably chosen to conduct the studies on the basis of their expertise. Regardless, we cannot be certain that the data isn't somewhat skewed towards the interests and experiences of the researchers behind the three studies in question. Unfortunately, there is no good way to test this and the unlikeliness of this phenomenon gives me reasonable certainty that the contributions and most important journals identify represents a robust research finding.

Using the methods described above, previous research has been able to identify the 20 most important journals in their fields. These will provide the basis for my data collection, and I will use these lists to draw a dataset from the Web of Knowledge. Many of the journals are represented in two and even all three of the fields, this is what we would expect considering that they are all researching similar themes, at least to some extent.

We will not discuss these findings in detail here, but rather discuss this in relation to the other findings in this thesis at a later stage, more specifically in the data chapter later in the thesis. These results will therefore provide me with the data framework for this thesis. In the next section I will discuss some of the findings in earlier network studies and discuss how I will expand and further develop their studies.

To investigate the subject matter of this thesis, I will need to first gather data that represents the fields of innovation studies, entrepreneurship and STS. The first step of the data retrieval process is to identify what is defined as the key literature in terms of the most important scholarly journals in the three fields. I will therefore build upon the data from earlier research, specifically the EXPLORE project. The project identified the most central journals for each of the fields. This represents the cut-of point for which journals that will be included in the set. The journals are presented in the table below.

There is another cut-off point that must be established, pertaining to time. Some of the journals are old, while some of them are quite a lot younger. This time issue therefore impacts what we can expect in terms of graph density and transitivity, i.e. how many of the researchers in the dataset have not been able to cooperate, and as we seek to measure cooperation it would invalidate the investigation if many pairs of scholars included in the dataset were not alive and/or active researchers at the same time.

I have therefore chosen a ten-year limitation for the dataset, including all the articles published between 2004-2014. The ten-year limitation is due to an assumption that most of the researches that have published one or more article in any of the journals during that time period have been active researchers at the same time. Meaning that most of the included researchers have at least have had a theoretical possibility of working together. If a wider time-span were to be used for analysis, several researchers would not have been active in the same time period, and we would thus expect a more decentralized network.

Table 1 – 20 most central journal by academic field

Nr.	Innovation	Entrepreneurship	STS
1	Research Policy	Strategic Management Journal	Social Studies of Science
2	Strategic Management Journal	Research Policy	Scientometrics
3	International Journal of Technology Managment	Academy of Management Journal	Science, Technology and Human Values
4	Academy of Management Review	Small Business Economics	Research Policy
5	Journal of Management Studies	Academy of Management Review	Studies in History and Philosophy of Science

6	Organization Science	Journal of Business Venturing	Social Science and Medicine
7	Academy of Management Journal	Journal of Management Studies	Isis
8	Technovation	Administrative Science Quarterly	Technology and Culture
9	Administrative Science Quarterly	International Journal of Technology Management	Minerva
10	Organization Studies	Organization Studies	Journal of the American Society for Information Science and Technology
11	Regional Studies	Journal of International Business Studies	Journal of Research in Science Teaching
12	Technological Forecasting and Social Change	Organization Science	Organization Studies
13	Management Science	Regional Studies	Strategic Management Journal
14	R&D Management	Journal of Business Research	American Sociological Review
15	Industrial and Corporate Change	Journal of Economic Behavior and Organization	Technological Forecasting and Social Change
16	Technology Analysis & Strategic Management	American Economic Review	Environment and Planning A
17	Human Relations	Management Science	Science Education
18	Small Business Economics	Journal of Business Ethics	Social Science Information sur les Sciences Sociales
19	Journal of International Business Studies	Journal of Management	Philosophy of the Social Sciences
20	Cambridge Journal of Economics	Journal of Economic Issues	Technology Analysis and Strategic Management

We can note straight away that there is significant overlap between the three lists, and seeing that the process to identify these three lists are the same methodologically, yet done separately, we can already from this list get a sense of the overlap between the three fields of study.

There are 43 distinct journals in the list. A quick look at the list reveal that innovation studies share more common journals with entrepreneurship studies than it does with STS, and these two also share more common journals than what is shared between entrepreneurship and STS. Innovation studies share 15 journals with one or both of the other fields, having five journals it does not share with any of the other fields. While entrepreneurship shares 13 of its journals with innovation studies (86.6%), it only shares 3 journals with STS (15%), while having 7 it does not share with any of the other fields. STS shares 5 journals (25%) with innovation studies, having 15 journals it does not share with any of the two other fields. Meaning that there are three journals shared by all three of the fields: ‘Research Policy’, ‘Organization studies’, and ‘Strategic Management Journal’.

The findings from the EXPLORE project that concluded that innovation studies and entrepreneurship are more overlapping than the other combinations therefore seems to hold when simply looking at the journals in the three lists and comparing them.

4.2 Attaining and Organizing the Data

The data pull is done through Thomson Reuters Web of Science, which is an online database containing massive amounts of meta-data on research journals, articles and scholars. The database contains all the relevant journals in the relevant timeframe, so the data I have drawn includes everything it is supposed to in relation to the top 20 journals for the three fields.

The meta-data from the Web of Science contains several dimensions of information per article. The relevant data for this thesis are the names of the authors of each article and the keywords. The structuring of the data meant removing the irrelevant meta-data and organizing three subsets. The subsets consist of the articles relevant for each of the particular fields, e.g. one subset with the top 20 journals for innovation, one for entrepreneurship, and one for STS.

However, there is some data missing in for some of the articles. Specifically, the data does not include keywords for all of the articles. This will potentially have an effect of the results of the analysis. This is probably due to the journals in question not conforming to the normal convention of including up to six keywords for all publications.

There are 33.385 articles with keywords in the data set of a total of 37.922 (88%). The amount of articles not containing keywords is therefore rather low, and will probably not skew the results in any dramatic way. Additionally, there are no good alternatives for finding keywords for the articles that do not have them when pulled from the database. It is a flaw in the dataset that is difficult to remedy, but it is important to note that there might be a small effect of this missing data.

The R environment is sensitive to small changes in spelling. Meaning that if there are variations in how a name is spelled, even as small as a difference in capital letter, R will treat that name as two separate entities. If two authors have the same name, R will treat them as the same person. We cannot with certainty exclude the possibility that these errors are present in the dataset.

4.2.1 Cleaning the Text-Data

There are several steps that should be undertaken to clean the data so that R can process it efficiently. There are algorithms in R that does a lot of this work, and I will shortly describe the necessary cleaning of the keywords and abstracts here.

After importing the text data into R as a .csv file there are certain characteristics of written text that must be removed or cleaned before we can proceed with an effective analysis. The first part is removing all punctuation so that we are left with just words. Next we make sure that all letters are in lower-case. As R would treat ‘Science’ as a different word than ‘science’, it is important to choose either lower-case or upper-case for all words, as this would have a large and dramatic effect on the results. We then remove all extra white space, meaning that all instances of extra spacing, tabs etc. are removed from the data allowing for instance ‘ research and development’ to be treated as ‘researchanddevelopment’ rather than the word ‘research’ and the word ‘development’. Further we run a stop-words algorithm that removes all the common stop-words in the English language, as described previously.

Lastly we run a stemming algorithm. Stemming entails reducing words to their basic form so that variations of the same words are processed as the same word. As R is sensitive to even the smallest variations in spelling, it will for instance treat the word ‘technology’ as a different word than its plural version ‘technologies’. The different versions of the words contain the same meaning and it is therefore necessary to treat them as the same word. The stemming algorithm follows a set of rules that was first described by Porter (Porter, 1980), and he describes how to basically remove suffixes from words to find their common root form. As the example below shows the suffixes are removed from the different versions of the same word to find the root:

technolog ⇐ technology
technologies
technological
technologist

After stemming the data, we are left with fewer unique words, and significantly higher validity of the analysis.

4.3 Descriptive Statistics

To get some general insights on how the data I will analyze looks like on a general level I will below present some descriptive statistics of the fields seen together, as well as the descriptive statistics of the fields separately. There are differences in the data between the fields in terms of the means, number of articles that have more than one author etc. that will provide an initial overview of the differences in collaboration among them. However, simple descriptives can be misleading and difficult to draw insightful conclusions from, so this will not be a large part of the analysis.

Firstly I will go through the descriptive statistics for all the fields seen as one, then proceed more granularly and look at how the sub-sets for the three different fields behave. I will comment on some general insight and differences in the sub-sets along the way.

4.3.1 Descriptive Statistics of Innovation, Entrepreneurship and STS

The tables below report descriptive statistics of the number of authors per article for the fields of innovation studies, entrepreneurship and STS combined, with a total of 37.922 unique articles, assuming that no articles have been published in two different journals. Median and arithmetic mean is fairly close and frequency distribution of authors per article as shown in table 3 shows that as much as 63.5 per cent of the articles have only one or two authors. The distribution further shows some outliers, with two articles having as much as 49 authors per article. Close cooperation among almost 50 researchers on a single paper seems intuitively unlikely. Seeing as this paper is concerned with the diffusion of information through co-authorship networks, these articles might skew the results in an unrealistic way under the assumption that the collaboration on these articles was low.

Table 2 – Descriptive statistics of Innovation, Entrepreneurship and STS

	N articles	Mean	Median	Sd	Min	Max
Authors	37922	2.33	2	1.40	1	49

Further the distribution of authors per article seems to follow a rough power law distribution with the majority of articles having quite few authors, and with a fairly long tail of papers having considerably more authors.

The table below also shows that there has been a collaborative effort to produce over 70 per cent of all the articles published. In general there is therefore quite a lot of collaboration taking place. We can also note that almost 95 per cent of the articles have 4 or fewer authors. Most collaboration therefore seems to be done in small groups, yet there are some articles with more authors.

Table 3 – Frequencies of Innovation, Entrepreneurship and STS

Authors per article	Frequency	Percent	Cumulative
1	11,100	29.27	29.27
2	12,981	34.23	63.50
3	8,576	22.61	86.12
4	3,123	8.24	94.35
5	1,137	3.00	97.35
6	574	1.51	98.86
7	194	0.51	99.38
8	89	0.23	99.61
9	67	0.18	99.79
10	24	0.06	99.85
11	14	0.04	99.89
12	16	0.04	99.93
13	7	0.02	99.95
14	4	0.01	99.96
15	3	0.01	99.97
16	1	0.00	99.97
17	2	0.01	99.97
18	1	0.00	99.98
20	1	0.00	99.98
22	2	0.01	99.98
23	1	0.00	99.99
26	1	0.00	99.99
30	1	0.00	99.99
41	1	0.00	99.99
49	2	0.01	100
Total	37,922	100	100

4.3.2 Descriptive Statistics of Innovation

In the top 20 journals for innovation studies there are 13.653 articles published in the ten year time frame. It is worth to note that the arithmetic mean for innovation studies is the lowest among the three fields, seeming to hint that there is slightly less collaboration in innovation studies than the other two.

Table 4 – Descriptive statistics of Innovation

	N articles	Mean	Median	Sd	Min	Max
Authors	13653	2.30	2	1.57	1	49

However, Innovation studies also have the lowest amount of articles with only one author, which seems to hint that it is the field with most collaboration in it.

Innovation studies have the fewest articles of the three fields; both entrepreneurship and STS have considerably more articles published in the time period.

Table 5 – Frequencies of Innovation

Authors per article	Frequency	Percent	Cumulative
1	3,184	23.32	23.32
2	5,434	39.80	63.12
3	3,585	26.26	89.38
4	1,078	7.90	97.28
5	250	1.83	99.11
6	68	0.50	99.60
7	13	0.10	99.70
8	14	0.10	99.80
9	10	0.07	99.88
10	3	0.02	99.90
11	3	0.02	99.92
12	2	0.01	99.93
13	2	0.01	99.95
15	1	0.01	99.96
20	1	0.01	99.96

22	1	0.01	99.97
26	1	0.01	99.98
30	1	0.01	99.99
41	1	0.01	99.99
49	1	0.01	100.00
Total	13,653	100	100

4.3.3 Descriptive Statistics of Entrepreneurship

Entrepreneurship has the largest amount of articles among the three fields. Significantly more than Innovation studies, which seems a bit odd considering that entrepreneurship shares 13 of its top 20 articles with innovation studies. The remaining seven articles that are a part of the entrepreneurship top 20, but not the innovation studies top 20 must therefore be large publications. When looking at the list of top 20 journals for the three fields, we can see that the remaining seven are journals relating to business, economics and management. This already seems to imply that entrepreneurship is more related to pure business and economics than the other fields.

Table 6 – Descriptive statistics of Entrepreneurship

	N articles	Mean	Median	Sd	Min	Max
Authors	19699	2.25	2	1.23	1	49

Table 7 – Frequencies of Entrepreneurship

Authors per article	Frequency	Percent	Cumulative
1	4,940	25.08	25.08
2	7,764	39.41	64.49
3	5,080	25.79	90.28
4	1,467	7.45	97.73
5	313	1.59	99.31
6	75	0.38	99.70
7	20	0.10	99.80
8	9	0.05	99.84

9	9	0.05	99.89
10	2	0.01	99.90
11	4	0.02	99.92
12	5	0.03	99.94
13	1	0.01	99.95
15	1	0.01	99.95
17	1	0.01	99.96
18	1	0.01	99.96
20	1	0.01	99.97
22	1	0.01	99.97
26	1	0.01	99.98
30	1	0.01	99.98
41	1	0.01	99.99
49	2	0.01	100
Total	19,699	100	100

4.3.4 Descriptive Statistics of STS

The data for STS consists of a total of 17.505 articles.

Table 8 – Descriptive statistics of STS

	N articles	Mean	Median	Sd	Min	Max
Authors	17505	2.42	2	1.60	1	26

STS has the highest arithmetic mean of the three fields, yet clearly the highest percentage number of articles with only one author. Further, STS is the only one of the three fields that shows more articles being written by one author than with two. For both innovation and entrepreneurship articles having two authors are the most common. STS seems do differentiate itself from innovation and entrepreneurship in how the researchers collaborate, but as there are more fine grained measurements of collaboration that we will look into later in this thesis, we will not conclude with anything yet.

Table 9 – Frequencies of STS

Authors per article	Frequency	Percent	Cumulative
1	5,856	33.45	33.45
2	5,044	28.81	62.27
3	3,356	19.17	81.44
4	1,574	8.99	90.43
5	810	4.63	95.06
6	490	2.80	97.86
7	175	1.00	98.86
8	79	0.45	99.31
9	60	0.34	99.65
10	22	0.13	99.78
11	10	0.06	99.83
12	11	0.06	99.90
13	6	0.03	99.93
14	4	0.02	99.95
15	2	0.01	99.97
16	1	0.01	99.97
17	1	0.01	99.98
20	1	0.01	99.98
22	1	0.01	99.99
23	1	0.01	99.99
26	1	0.01	100
Total	17.505	100	100

4.3.5 Comparison

As shown in the table below there are some differences in the collaborative activities for STS compared to both innovation and entrepreneurship. The frequency distributions look rather similar for innovation and entrepreneurship, while STS seems to display a more fragmented field, in terms of collaboration through co-authorships.

Table 10 – Descriptive statistics of Innovation, Entrepreneurship and STS

	N articles	Mean	p50	p95	p99	Singles
Innovation	13653	2.30	2	4	5	23.32%
Entrepreneurship	19699	2.25	2	4	5	25.08%
STS	17505	2.42	2	5	8	33.45%

While STS is in the middle in terms of the number of articles in its network, is it much more distinct from innovation and entrepreneurship than they are to each other in almost every other aspect presented here. The mean number of authors per article is high, while they have the highest percentage number of single author articles as well as the distribution of authors is different. STS thus show signs of both more collaboration and less collaboration, at least that is the impression from simple descriptive statistics. Through the network analysis later on we will be able to delve deeper into this.

The fields of innovation and entrepreneurship are very similar in every aspect expect of the number of articles included in its most important journals. This might be because of things such that the economical and business journals that entrepreneurship tend to publish in tend to be larger volumes of work than that of innovation studies.

The descriptive statistics presented above has provided an overview of how the data looks, and will be useful as a backdrop for the coming analysis. There are already signs in the descriptive statistics that there are differences among the fields. In the next chapter I will do the analysis and present the results of the status of collaboration in the three fields.

5 Analysis and Results

In this chapter I will present the analysis and results based on the data acquired. There are two analytical approaches used. The first is an exploratory and more qualitative text mining of keywords that will shed light and provide insight on what the similarities and differences are between the three fields in terms of their research focus. This will help us establish what research is undertaken, and will provide us with a possible reason why collaboration between-fields and within-fields might differ. If one field is more removed from the other two, there is an obvious reason to suspect that this field will also collaborate less with the other two.

After investigating the differences in research themes I will proceed to the main analysis of this thesis; a social network study of the three fields. There are several reasons to suspect that there are differences in collaboration based on the descriptive statistics presented above, which I will investigate thoroughly.

5.1 Text-mining Keywords

The first step of the analysis is run the text mining process for the keywords to the articles for all the fields. First we will find the common ground for all the fields by looking at all the keywords for all 37 unique journals seen as one. Below is a list of the top 30 keywords associated with innovation studies, entrepreneurship and STS along with their frequencies. The words are stemmed so that the words shown will often not be full words, rather the root word.

Table 11 – Most common stemmed keywords in all fields

Nr.	Keyword	Frequency
1	perform	5685
2	firm	4435
3	model	3994
4	manag	3809
5	innov	3316
6	behavior	2984
7	knowledg	2949
8	scienc	2404
9	industri	2343
10	market	2310
11	product	2281
12	organ	2264

13	technolog	2183
14	perspect	2169
15	network	2043
16	competit	1931
17	inform	1866
18	impact	1739
19	strategi	1733
20	organiz	1618
21	busi	1617
22	health	1497
23	system	1447
24	unitedst	1442
25	work	1415
26	growth	1387
27	research	1318
28	dynam	1299
29	advantag	1250
30	social	1240

A perhaps more useful representation of the results of the keyword analysis is the word cloud below, that shows the top 50 keywords where their relative size reflects the frequency of its occurrences.

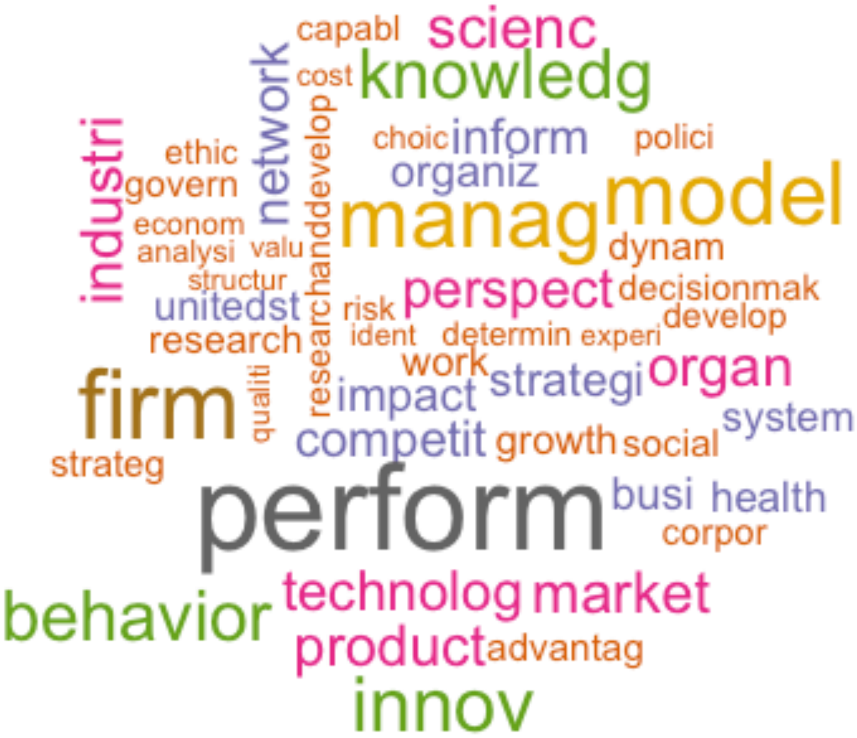

Figure 2 – Keywords All Fields

Performance leads the field with 5685 occurrences among the 37.922 articles in the data set. We can also see that the main themes seem to be related to business, economics and management. However, to get a picture of how the three fields differ we will need to run the same analysis for all three fields separately and compare it.

5.1.1 Comparing the fields

The information above gives us a general view of the themes being studied on the three fields seen as one. But to investigate the fields separately we must do a similar analysis of the three fields by themselves and then compare the results. This is a more qualitative analysis than quantitative, although I have included the frequencies in the table below to show differences in occurrences.

Table 12 – Most common keywords by field

Nr.	INN		ENT		STS	
1	perform	3372	perform	4464	scienc	2111
2	firm	3232	firm	3505	knowledg	1477
3	innov	2636	manag	2779	health	1374
4	manag	2285	model	2772	innov	1373
5	model	1801	behavior	2182	perform	1269
6	knowledg	1760	innov	1903	technolog	1185
7	industri	1737	market	1853	model	1113
8	technolog	1524	organ	1674	firm	1012
9	product	1502	industri	1604	network	993
10	organ	1344	competit	1564	manag	982
11	market	1236	knowledg	1529	impact	937
12	competit	1233	perspect	1465	system	797
13	network	1189	product	1447	industri	790
14	perspect	1181	busi	1384	research	789
15	strategi	1052	organiz	1332	product	771
16	organiz	1047	strategi	1317	unitedst	770
17	researchanddevelop	935	inform	1268	perspect	750
18	strateg	915	network	1082	behavior	718
19	dynam	880	corpor	1075	analysi	663
20	growth	878	advantag	992	organ	650
21	capabl	872	strateg	982	inform	578
22	develop	846	ethic	970	citat	577
23	advantag	841	technolog	959	polici	576
24	system	800	growth	934	mortal	543
25	behavior	779	decisionmak	903	inequ	529
26	inform	768	work	898	care	522
27	work	756	dynam	865	women	506
28	busi	667	govern	850	scientif	504
29	allianc	601	capabl	780	strategi	499
30	govern	600	impact	766	indic	494

For a better visualization of the data presented in the table above, see the word clouds with the top 50 recurring words for each of the three fields below.

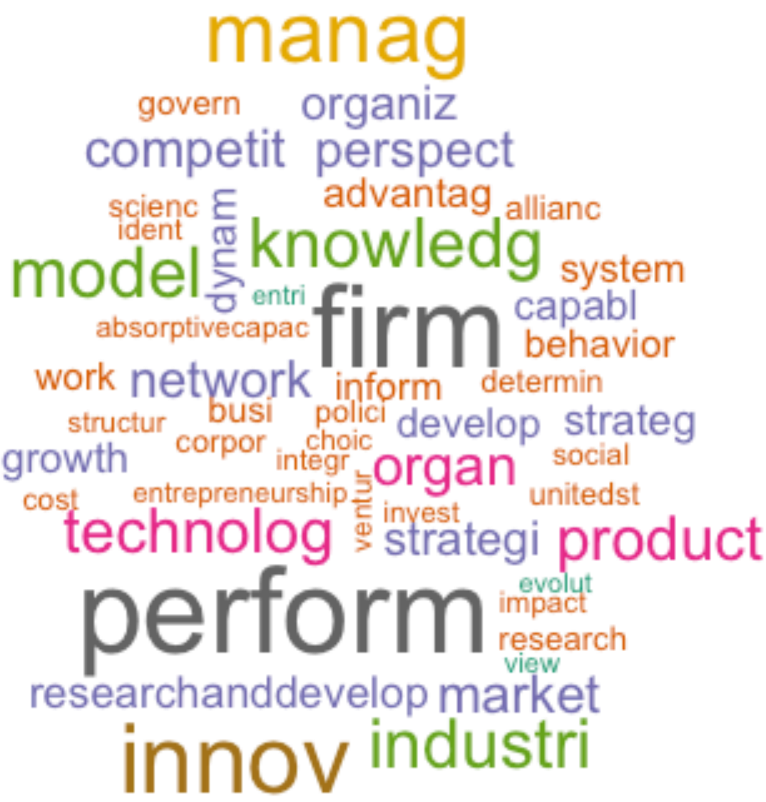

Figure 3 - Innovation Keywords

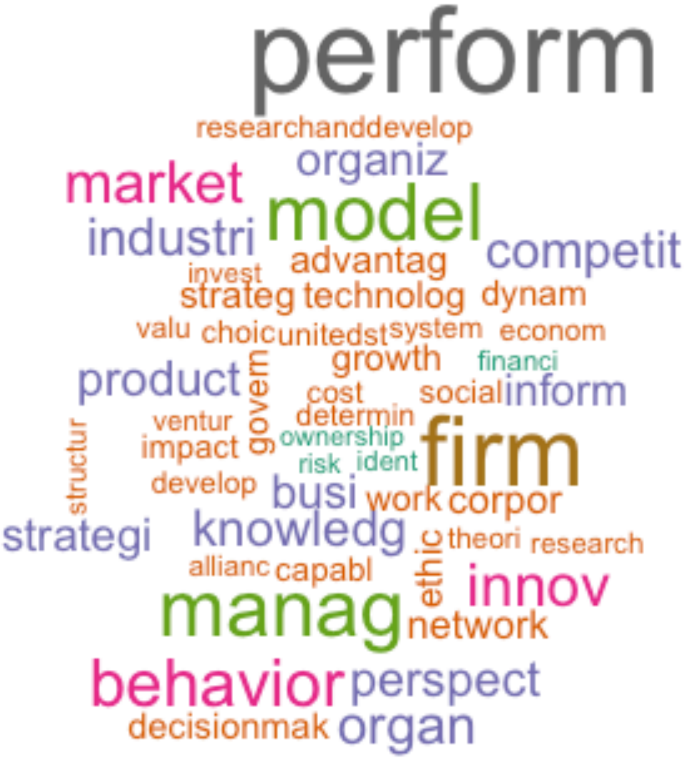

Figure 4 - Entrepreneurship Keywords

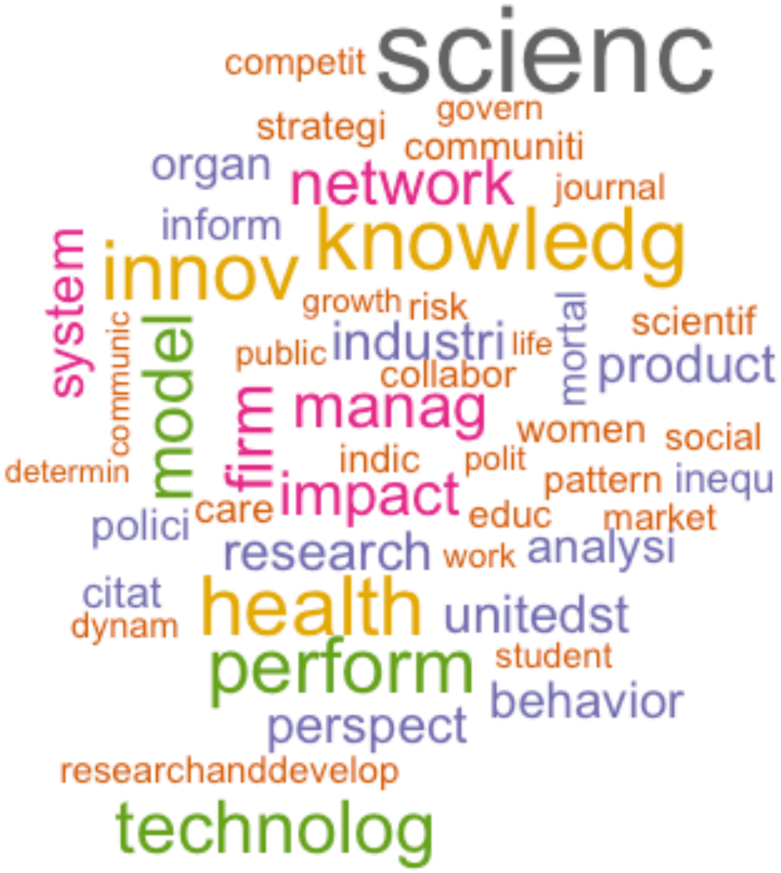

Figure 5 - STS Keywords

Seeing as the overlap in the journals are as large as they are most of the keywords that carry any importance are present in all three of the lists. I will therefore focus on the top 30 keywords to qualitatively see if there are differences in what are the most important research themes, here defined as the top 30 most common themes found in keywords for articles.

Of the 30 keywords 15 of them (50 %) are shared between the three lists. The overlap is in that case rather large. There is a large difference in the amount of keywords that are present in only one of the lists. While innovation and entrepreneurship have three keywords each among their top 30 that are only in their respective lists, STS have 13 keywords (43 %) that is only present in its. One of them, perhaps surprisingly, is the keyword 'scienc', which is also featured at the top of the list for STS. This is a strong indication that STS have a purer focus on how science affects and shapes society than the more business, growth and performance oriented fields of innovation and entrepreneurship.

Innovation and Entrepreneurship shares 11 keywords (37 %) that are not present in the top 30 for STS, making it the most overlapping fields. Innovation and STS shares one keyword that is not present in entrepreneurship, the word ‘system’. It seems that they are more concerned with innovation systems than entrepreneurship. Entrepreneurship and STS shares one keyword that is not present in innovation, which is ‘impact’.

There seems to be evidence that the themes researched among the fields are more similar between innovation and entrepreneurship, than any of them with STS. While innovation and entrepreneurship are more focused on matters related to business, strategy, growth etc. STS seems to have more focus on science and technology. This reflects the impression we got from looking at the journals that are considered the top 20 for each of the three fields. However, we should bear in mind that half of the keywords were present in all three fields and the field are therefore still very similar.

We have found that there are differences between the three fields in terms of research topics, and also we have confirmed the impression given to us by the overview of the most central journals that innovation studies and entrepreneurship overlap more than either of them do with STS. The research topics for innovation and entrepreneurship seem to be more focused on economic performance, business and management, while STS is more focused on science and technology in general, which is not a surprising result for practitioners in the fields. The results from the text mining give a backdrop for the upcoming social network analysis, which I will do in the next chapter. There we will be able to investigate how the fields cooperate between them and the level of cooperation within them. After seeing that the research topics are indeed different, we might suspect that the level of cooperation between the fields also differs.

This exercise has been useful in an other way, in that the qualitative assessment of the research themes between the fields are different between them, and a good fit with what you would expect the different fields to study. This lends credibility to the data set used, as the large amount of overlap could perhaps be a blurring factor, something it does not seem to be.

5.2 Social Network Analysis

In this section I will graph the networks of the three fields of innovation, entrepreneurship and STS. There are several measures and indicators that have been described in the methodology chapter that will be analyzed in this chapter. Further I will present illustrations of the networks, which is one of the key components of most social network analysis. As the networks in this study are fairly large I will use additional analytical methods to develop a deeper understanding of the unique characteristics of the three fields. As the methods and framework behind the data collected for this thesis are the same for the three fields, they lend themselves for comparison, which will be an important aspect of the analysis.

After analyzing the fields separately I will look at the entire corpus of journals from all three fields as one large field and see how this compares to the separate analysis undertaken. There are some interesting aspects we can look into when analyzing the entire corpus of journals, especially who hold important positions as 'gatekeepers' of information, i.e. are important for the effective diffusion of information through the network, this is of course highly related to the betweenness centrality of nodes.

5.2.1 Graphing the Innovation Network

Seeing as we are interested in the co-authorship network among the scholars we remove all of the isolates in the dataset. An isolate is an article with only one author. There are 1.384 isolates in the dataset, thus reducing the subset to be analyzed to N=12.269. Once the isolates are removed there remains 2610 sub graphs, with a total of 16.768 unique authors. In the figure below you can observe the complete network of authors, excluding the isolates.

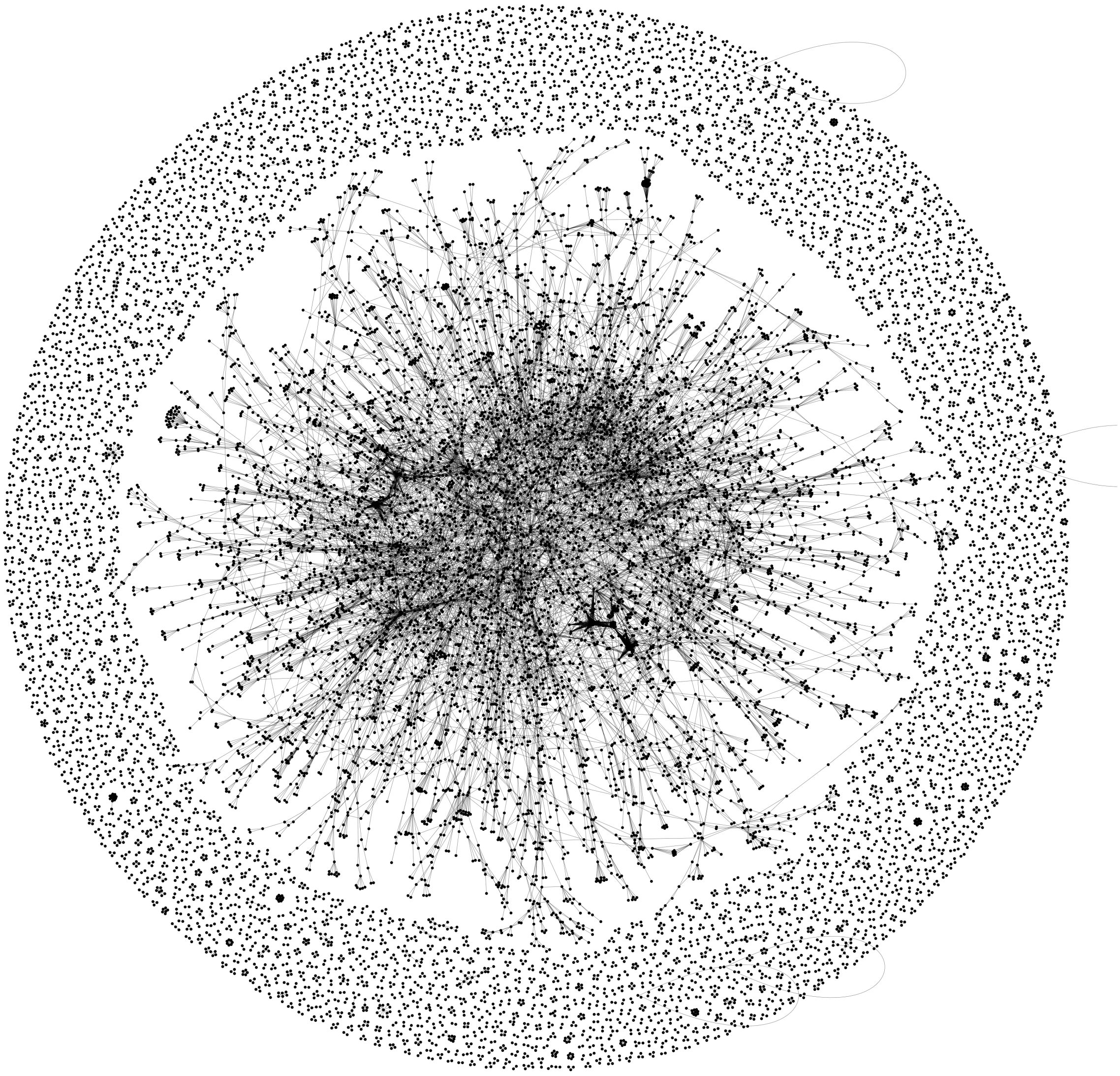

Figure 6 - Innovation Graph

Transitivity for the graph is 0.697, reflecting the fact that the graph is very clustered. In almost 70% of instances nodes *i* and *j* are connected when node *k* is connected to both nodes *i* and *j*. The graph is, however, very sparse. Only 0.0198% of the possible edges are present in the graph. Considering that there are as many as 2610 sub graphs in the data set, the fact that graph density is as low as it is, should not be a surprise. The graph therefore shows clear signs of the field cooperating rather little, but with tight cooperation between groups of researchers. As it is a bit difficult to see the details of the graph when it is so big we can take a closer look at certain aspects of the graph. Below is a part of the graph where I have focused on the small and disconnected sub graphs.

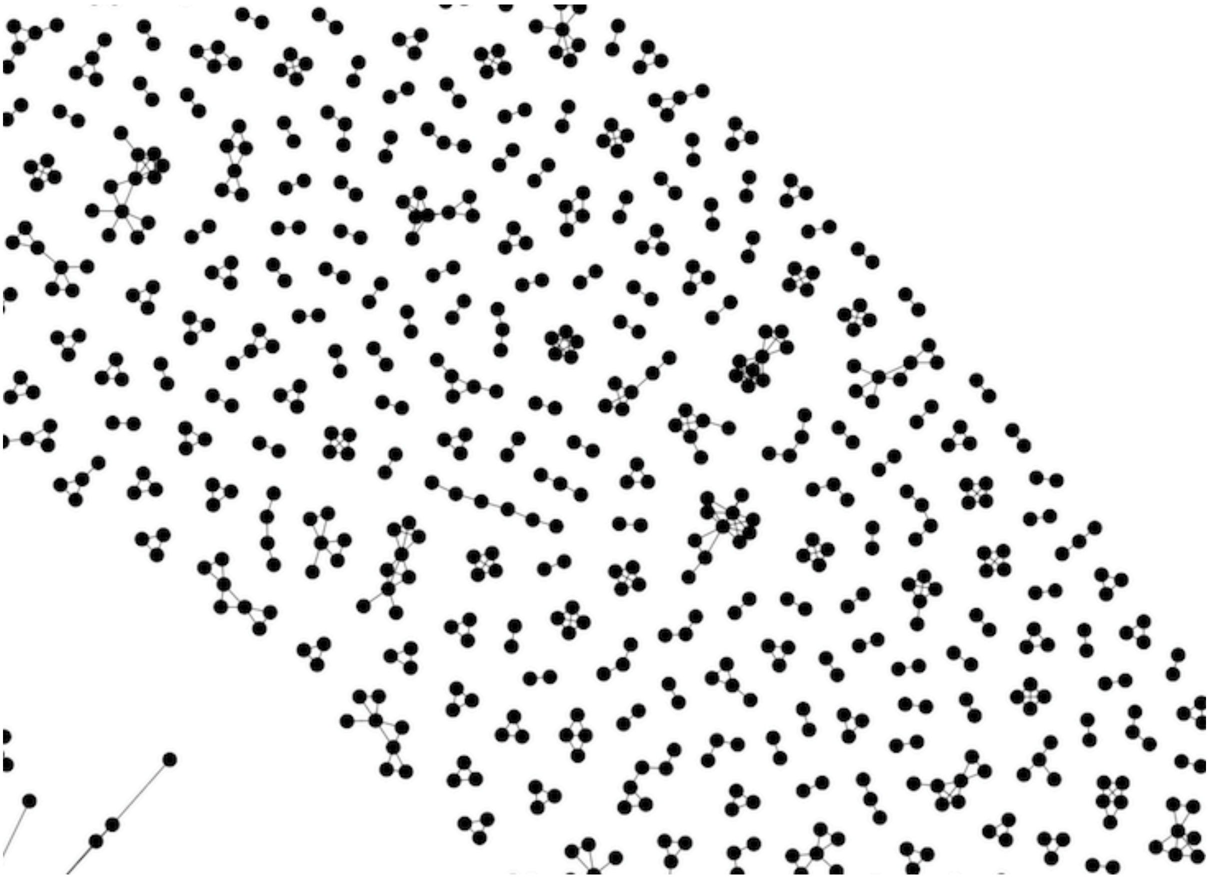

Figure 7 - Innovation Sub Graphs

Most of the sub graphs are cases of all the nodes in a cluster being connected to all the other, which is often a case of a group of researchers publishing a single article. These researchers might publish in a journal that they do not normally publish in, thus being disconnected from the rest of the graph. There are also several other interesting patterns in the illustration, such as straight lines of researchers and quite complex constellations that shows the intricacies of scientific cooperation. However, most are simply pairs of researchers having cooperated on a single paper. This tendency is the same for entrepreneurship and STS, as well as the graph covering all three fields.

There are a large amount of graphs that are not connected, while there is one central very large graph. The 16.768 authors have a total of 27.863 edges, while the diameter, meaning the largest distance between any two nodes, also called the geodesic, that are connected in a graph, is 28. When analyzing networks there is often one large component with several smaller components, and this is no exception. Many of the smaller graphs represent authors collaborating on a single article. As there is no collaboration, at least as reflected in the data set, between the different sub graphs, it is fruitful to delve deeper into the specifics of the largest component, as it is there most collaboration takes place.

I have therefore separated the data set to focus on the largest component of the sub graphs. The largest component consists of 8.541 nodes with 19.220 edges. While the diameter is still 28. 50.9% of the nodes in the innovation network are in the largest component, which can be seen as the large mass of connected nodes in the center of the illustration above. The transitivity of the largest component is 0.671 while the graph density is 0.000527. The transitivity remains about the same as the entire innovation network, while the density is significantly higher. The largest component is over twice as dense as the entire network. To get a clearer picture of the field we will again look closer at some of the characteristics of the graph.

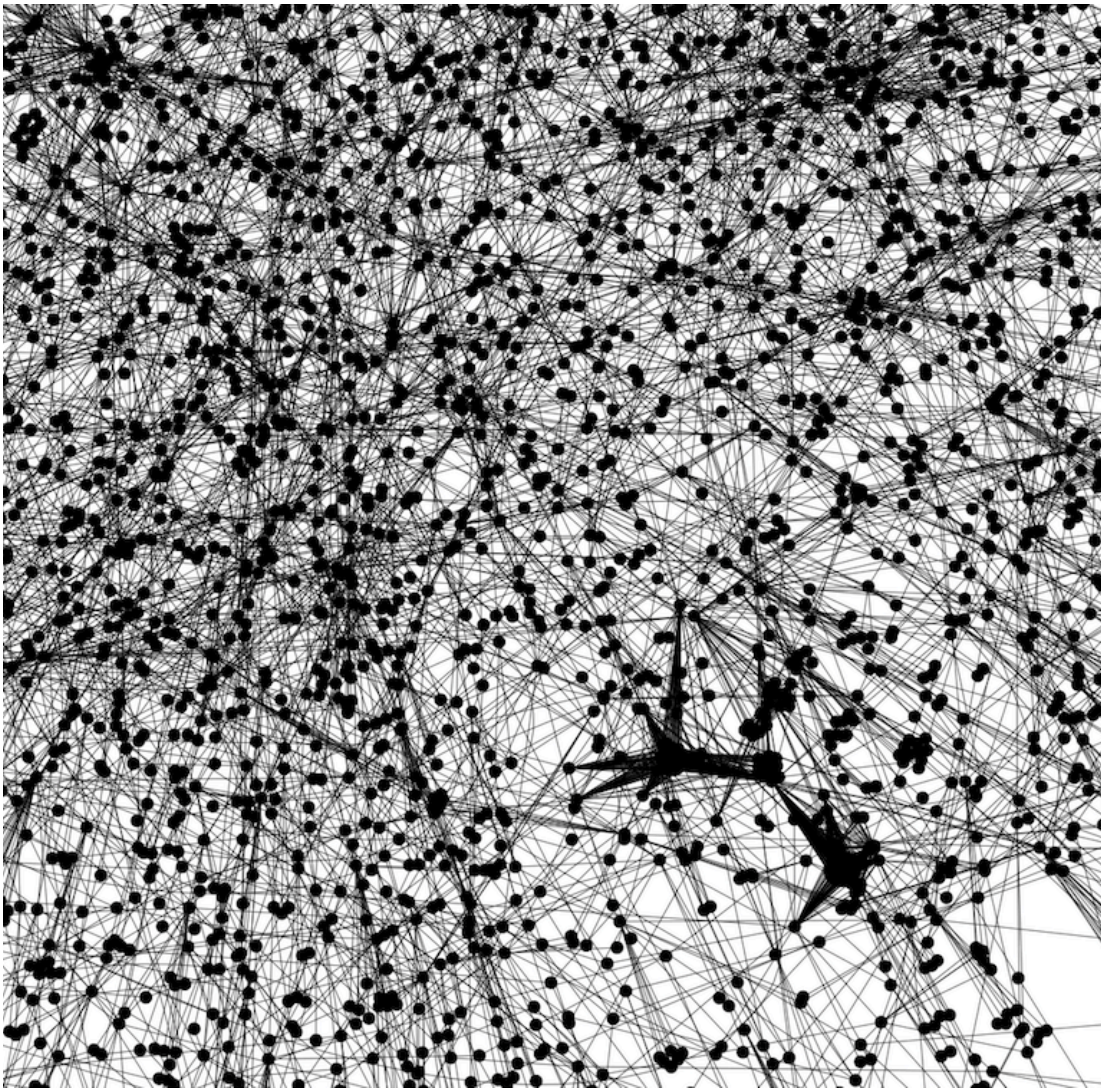

Figure 8 - Innovation Superclusters

The visual inspection of the largest component of the innovation network reveals a variety of edges, with some super-nodes with a high amount of edges. Much of the network, however branch outwards in a tree like fashion along the edges. We can also observe some very tightly knit clusters in the graph. Apart from this there seems to be a quite chaotic situation, which would reflect the sparseness of the network. In a highly structured or less fractured field, we would expect nodes that are connected to be closer together as well as a tendency for nodes to be connected to the other nodes close to them. This visualization clearly demonstrates the low level of density found in all the graphs studied in this paper. The same characteristics are to be found in all of them. When considering the relatively high transitivity scores it would seem that the field is characterized by many tightly knit clusters with a lot of interaction within them, and sparse collaboration between these clusters. A trait that is easily seen in the visualization presented above.

Still however, in networks as large as this visual analysis have some limitations, as the amount of information contained in the dataset is very large. Especially the degree distribution is important as it tells a lot about the level of cooperation in the field, yet difficult to get an overview of through graph visualization. The degree distribution below shows the number of nodes with the number of degrees associated with them as a percentage of the largest component.

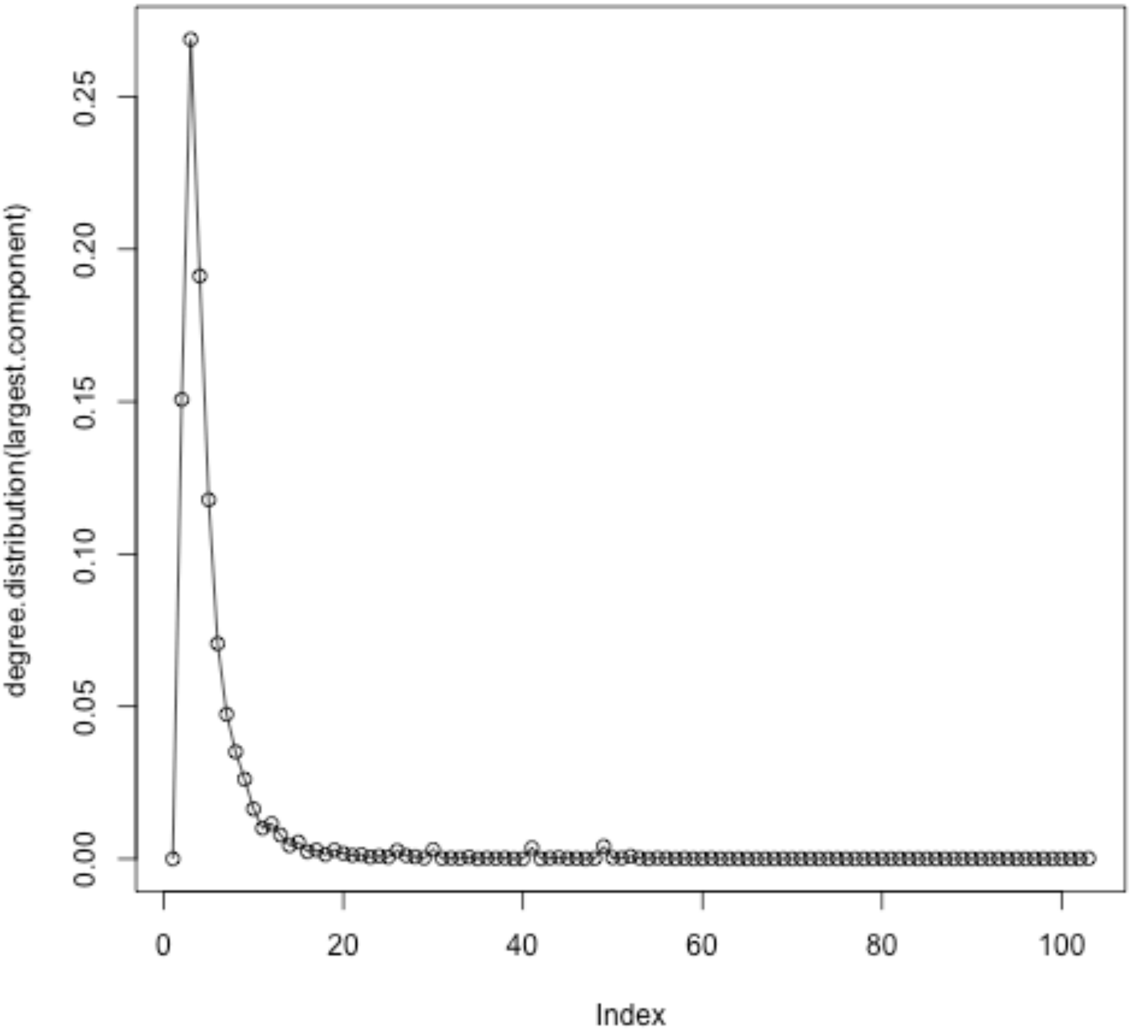

Figure 9 - Innovation Degree Distribution

As we can see from the degree distribution chart over 25% of the nodes have two adjacent nodes to them. Another interesting observation is that few nodes have only one adjacent node. The degree distribution gives us a good overview of the level of cooperation on the individual level; in terms of how many other researchers a specific researcher collaborate with.

5.2.2 Graphing the Entrepreneurship Network

The entrepreneurship network is the largest of the subsets where there are 19.699 articles in and 4940 isolates. After removing the isolates we are left with 14.759 articles, which is also larger than the innovation network. Below is an illustration of the entire entrepreneurship network, including all sub graphs, but not including isolates in the data set.

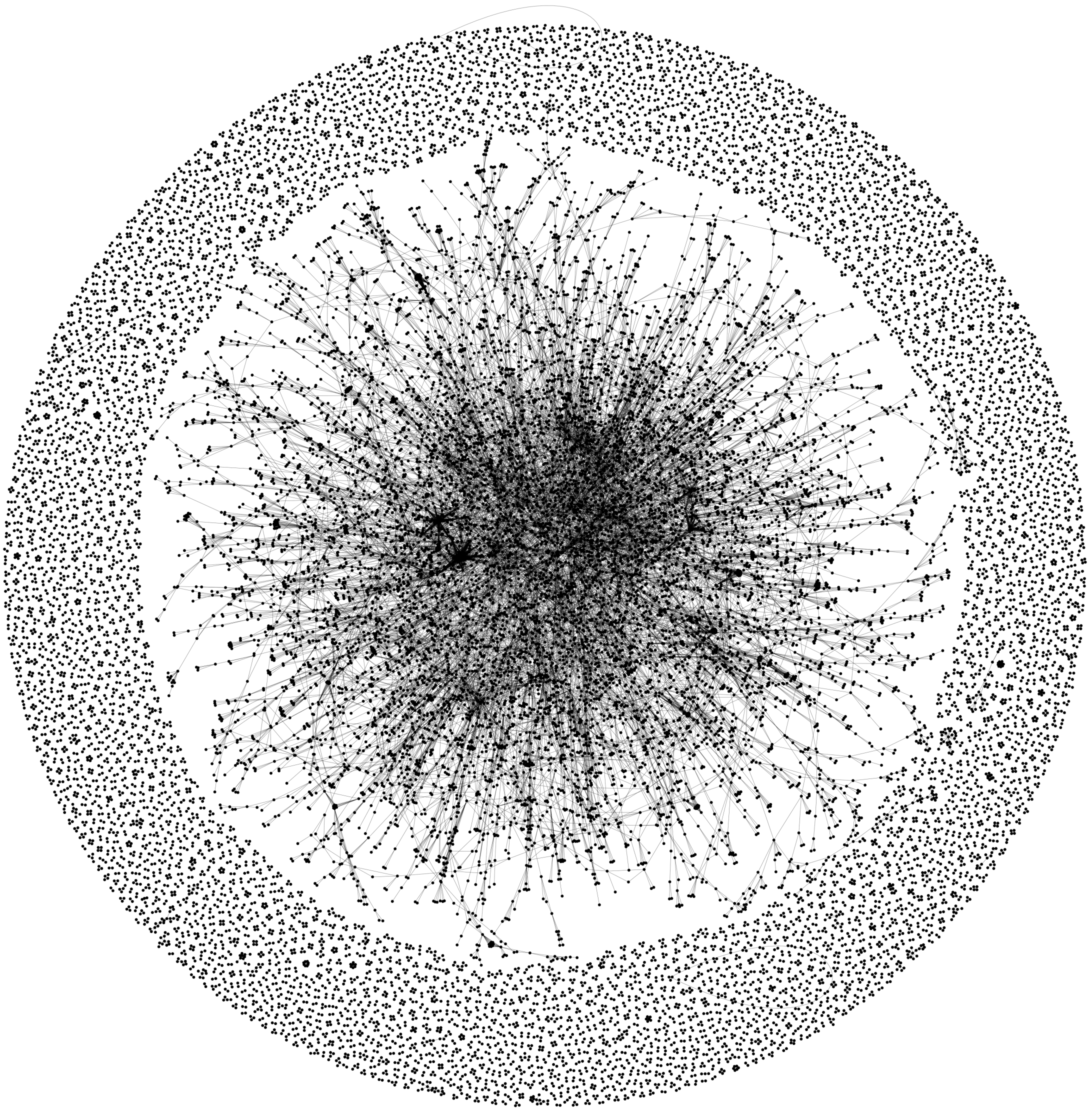

Figure 10 - Entrepreneurship Graph

The network contains 23.056 unique nodes with a total of 37.468 links, while the diameter is 29. The transitivity of the graph is 0.651, and the density is 0.00014. We are again dealing with a highly fractured field in terms of the amount of possible links present in the network, yet with a very high level of transitivity.

As with the innovation network there are a great many sub graphs in the data set. There are a total 3.652 sub graphs in the network, considerably more than the innovation network, but as expected considering that the entrepreneurship network is considerably larger. The large amount of tightly knit sub graphs, most of them very small with all the researchers in the graph tied together, and is as with the innovation network often linked to single papers in a journal. Perhaps indicative of a group of researchers publishing work in a journal on the fringe of their scientific field, so there would be little reason to suspect any previous collaboration with the core field which is represented by the largest component.

The largest component for the entrepreneurship network has 11.824 nodes with 26.544 edges, and the diameter remains 29. About half of the nodes in the entire network are present in the largest component, which can be seen in the center of the illustration above. The graph density for the largest component is as would be expected clearly higher than for the entire entrepreneurship network with 0.000379. The transitivity is about the same with 0.647. Also the largest component is similar to the innovation network in that it shows clear signs of being highly fractured, yet extremely tightly knit. If we were to zoom in on specific areas again, we would see the same telling characteristics as for the innovation graph. It is of course important to remember that there is a significant overlap in the journals included for innovation and entrepreneurship, which would automatically lead to somewhat similar results between the two. There density of information in the illustration in high, so we will need to look at the degree distribution of the authors, which you can see immediately below.

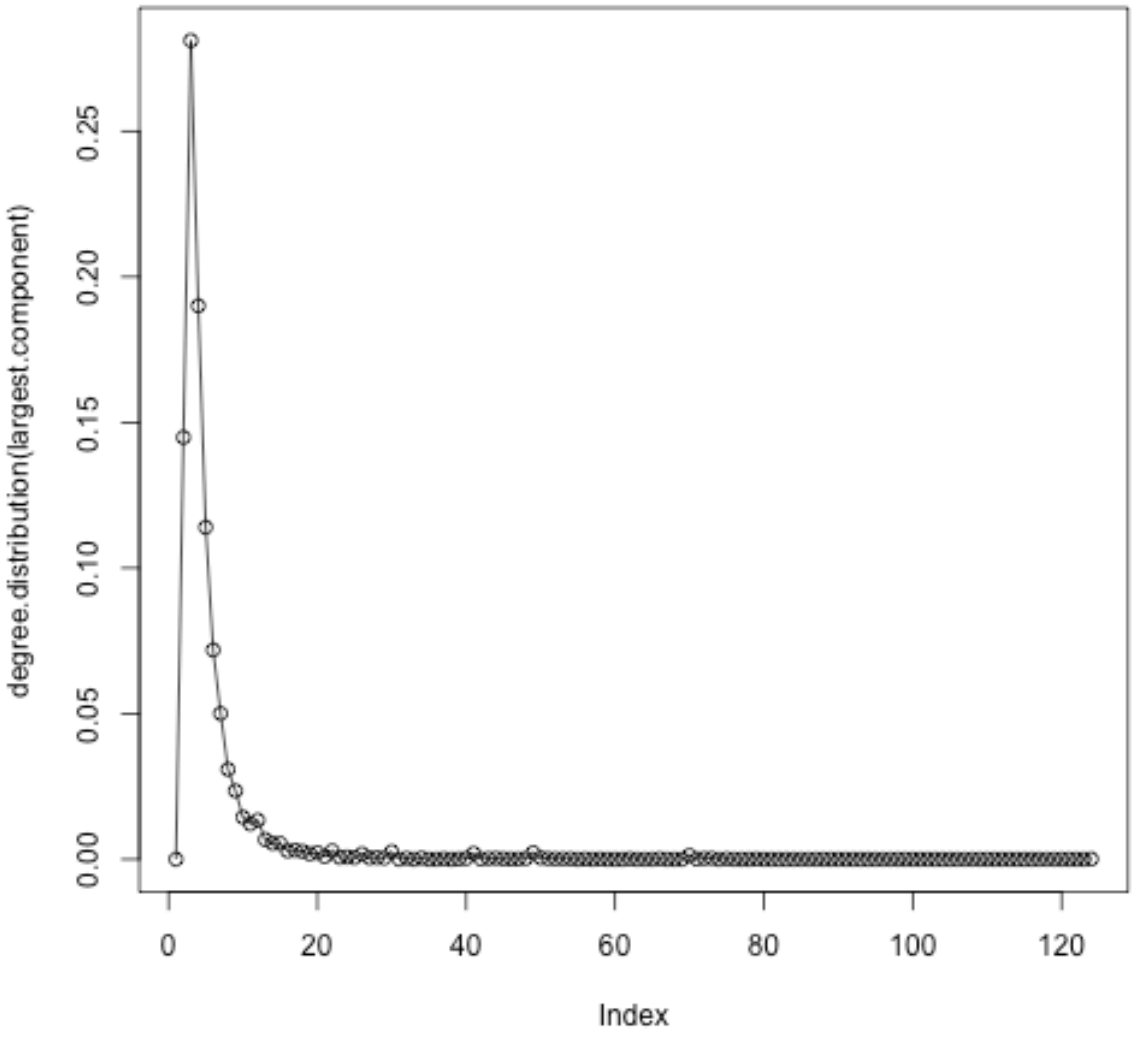

Figure 11 - Entrepreneurship Degree Distribution

The degree distribution is very similar to that of the innovation network, where two is the number of degrees most common. The distribution also shows the same steeply rising curve up to two degrees, then a rough power law distribution with a very long tail. Meaning that most authors chose to collaborate in small groups, while a few collaborate with many more on a particular publication.

5.2.3 Graphing the STS Network

There are a total of 17.505 articles in the STS corpus of articles. There are 5.856 isolates in the dataset, thus making the dataset to be analyzed to N=11.649. This is larger than the innovation network, yet smaller than the entrepreneurship network. There are 24.610 unique authors in the STS dataset with a total of 47.294 edges and a diameter of 30.

When the isolates are removed we are left with 3.851 sub graphs in the network, and again there is a clear largest component that contain a large amount of the authors. It´s noteworthy that the STS network contains more sub graphs than the other two fields, even though the amount of edges in the network is significantly larger. A graphic representation of the network can be seen below, were we can clearly see the large number of sub graphs.

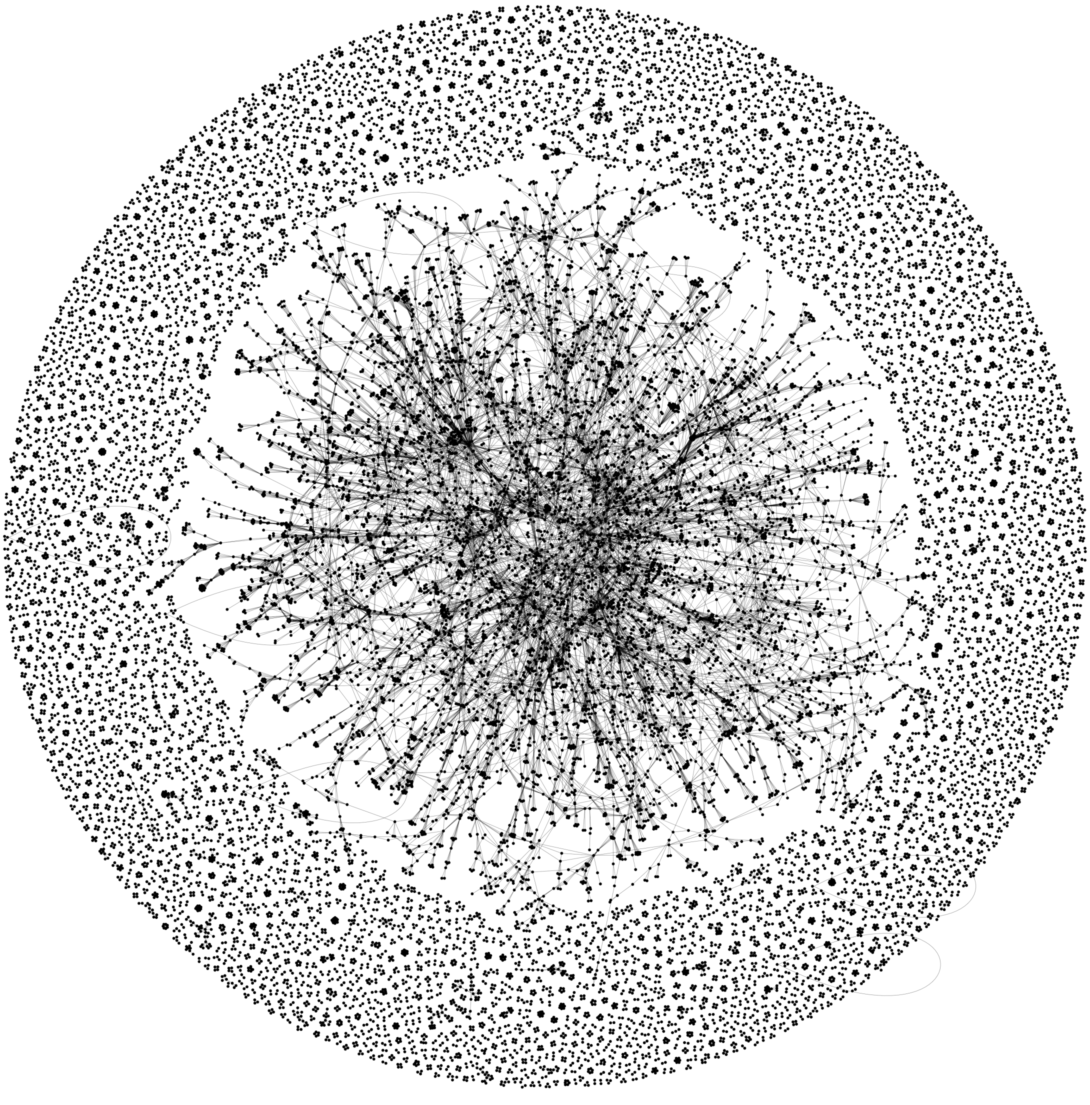

Figure 12 - STS Graph

The visualization of the STS field makes it apparent that the field is rather clustered as we can see by the large number of sub graphs populating the visualization. The characteristics of these clusters are the same as for innovation and entrepreneurship. However, we will again need to separate the largest component to go into further detail on how the field is connected. The largest component consists of 10.304 unique authors, which is 42% of the entire network. These nodes have 26.634 edges, while the diameter is still 30. 30 is the largest diameter among the three networks, but not by much. We can recognize some of the same super nodes as we have already observed in the other two networks, hinting at the three fields overlapping articles and journals, and thus also being similar to the characteristics displayed when looking closer at the innovation network. As we can see from the degree distribution of the nodes below it seems similar to that of the other two networks.

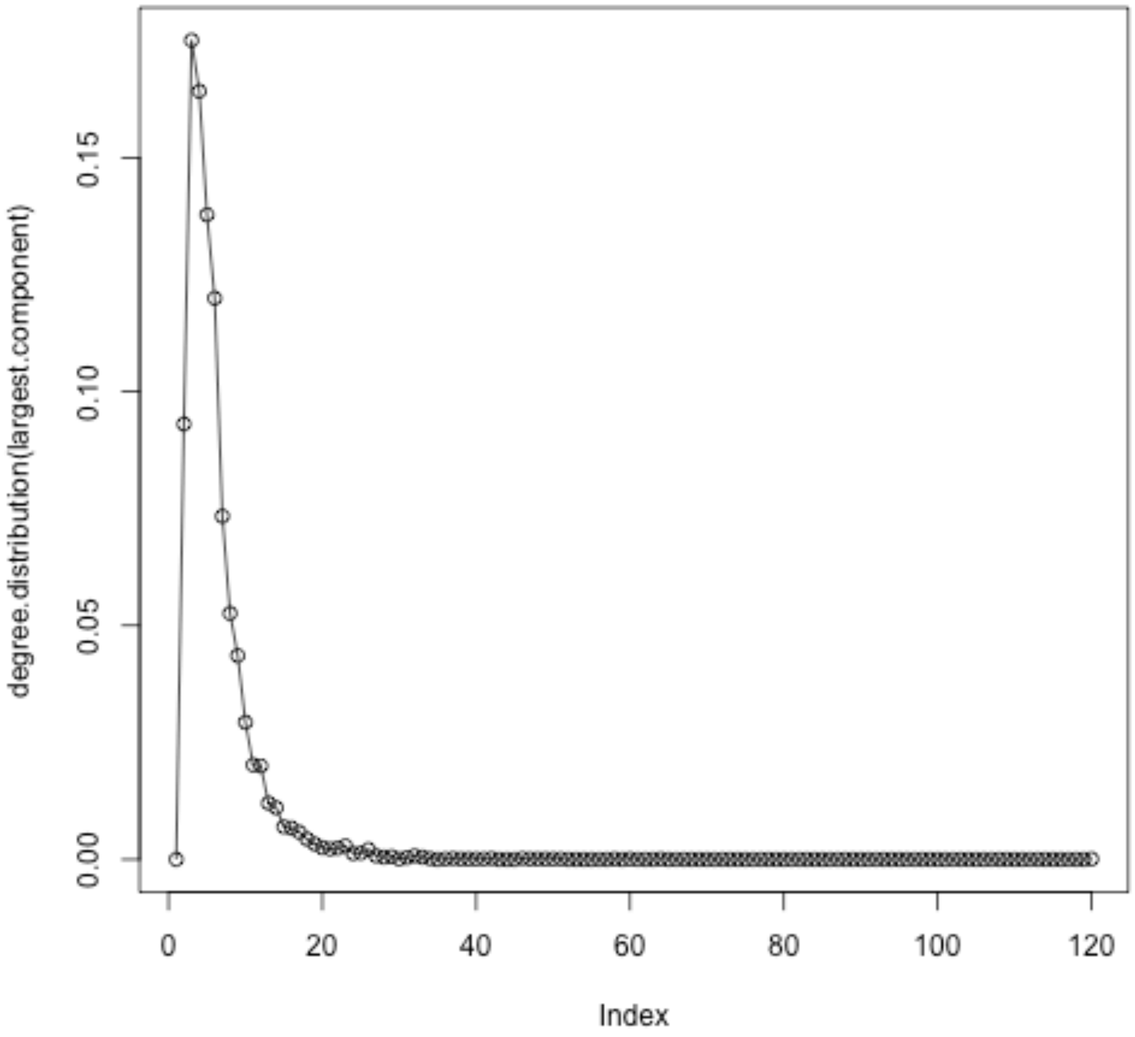

Figure 13 - STS Degree Distribution

When inspecting the graph of the degree distribution in the network we can see that the STS network shares the characteristic of having two as the amount of degrees most common among the authors, as well as following a rough power law distribution with a long tail. However the STS network have some peculiarities in terms of the percentage amount of authors with a degree of two, with only about 15%. The distribution of degrees among the authors in the STS network is tighter than that of both the innovation and entrepreneurship networks. Compared to innovation and entrepreneurship, STS have a lower share of authors with a high number of edges. Conversely there are a larger number of nodes with a low amount of edges.

5.3 Comparing the Three Networks

The three networks display similarities as well as certain dissimilarities. As there is a significant amount of overlap a degree of similarity is of course to be expected. In this section I will analyze the three fields when compared against each other on metrics and characteristics of the networks. In the table below these characteristics are summarized by field.

Table 13 - SNA results for Innovation, Entrepreneurship and STS

	Innovation	Entrepreneurship	STS	All fields
Articles	13653	19699	17505	37922
entire network nodes	16768	23056	24610	45335
entire network edges	27863	37468	47294	84455
sub graphs	2610	3652	3851	6159
entire network density	0.000198208	0.0001409748	0.0001552845	0.000082185
entire network transitivity	0.6791103	0.6514193	0.6046811	0.5771766
largest component nodes	8541	11824	10304	25022
largest component edges	19220	26544	26634	60130
largest component density	0.0005270075	0.0003797557	0.0005017609	0.0001920855
largest component transitivity	0.6711832	0.6476641	0.5311346	0.5503913
Diameter	28	29	30	27

Isolates	1384	4940	5856	11100
Articles without isolates	12269	14759	11649	26822
% of nodes in largest component	50,94%	51,28%	41,87%	55,19%
% of isolates of entire network	8,25%	21,43%	23,80%	24,48%

One of the first striking features is that innovation contains significantly fewer articles than the other fields, which again is represented by the fact that the network has comparably fewer nodes and edges compared to the other fields. It seems that STS have a significantly higher amount of edges than the other networks compared to its size in terms of number of nodes. This would lead us to believe that the density of the STS network is higher than that of the other fields, yet the graph density scores tells us that STS is in fact the second most dense field, with innovation being the densest.

Innovation is also the network with the highest transitivity, meaning that it is more tightly knit than the other two. Both of these results might of course be because of the lower amount of sub graphs, which follows the fact that the network is smaller than the other two. We must therefore compare the largest networks to rid some of these effects.

For both innovation and entrepreneurship over 50% of the nodes are within the largest component, while for STS only 41% is in the largest component. This indicates that STS is a more fractured field than the other two. We can also notice this tendency when looking at the number of isolates found in the networks; STS clearly has a larger number of isolates than the other fields, the percentage of isolates of all nodes is also clearly higher for STS. Innovation studies separate itself by having very few isolates compared to the number of nodes. All this further suggests that innovation and entrepreneurship is fairly similar and that STS is more distant and display characteristics that make it more fractured and a field where there is less cooperation among authors.

The transitivity of innovation and entrepreneurship in the largest component is about the same, and again STS display a much lower level of transitivity. The density of the STS network is not very low however. Entrepreneurship displays the lowest density of the fields. The interpretation of this is that entrepreneurship is highly fractured in that there is little cooperation among researchers, but there is a high degree of cooperation within the largest component among tightly knit group of researchers. Conversely STS has a relatively low transitivity and a relatively high density, meaning that the cooperation is fairly well distributed, and to a lesser extent done within tightly knit groups in the largest component.

The comparison of the three networks strongly suggests that the field of STS contains less cooperation than both innovation and entrepreneurship. Entrepreneurship and innovation is of course more similar in many of the measurements undertaken here because of the large overlap of most central journals for the fields. There are still many things that separate them. However, all the fields show clear signs of being highly fractured, yet extremely tightly knit. This might be because of the young age of the fields, and that the journals included in the data set are not all specific for the sub fields. The community of scholars active in one or more of the fields in the given time period might very well chose to collaborate with a few other who share their particular interest a lot, while sometimes collaborating with other scholars in adjacent fields. This would at least to some extent explain the highly fractured characteristics of the fields.

Considering the overlap between the fields, and the differences and similarities between them it is pertinent to analyze the fields seen as one entity. By doing this we can uncover to some extent if the fields should be seen as one overarching field rather than three separate ones by comparing it to its three sub components as well as other research on similar networks. In the next chapter I will analyze the entire field seen as one entity.

5.4 The Three Fields Seen as One

In this part I will analyze all the three fields seen as one, e.g. one large network of innovation, entrepreneurship and STS. By analyzing the network in this way, we can potentially find important scholars that bind the three together and are important for the diffusion of information. Centrality measures will therefore be important in this part, to look at what characterizes the nodes that are important for the diffusion of information. First, however, we will look at other aspects that characterize the network of researchers. It is important to remember that since there is significant overlap between the fields, this network will be similar to the others in terms of its structure.

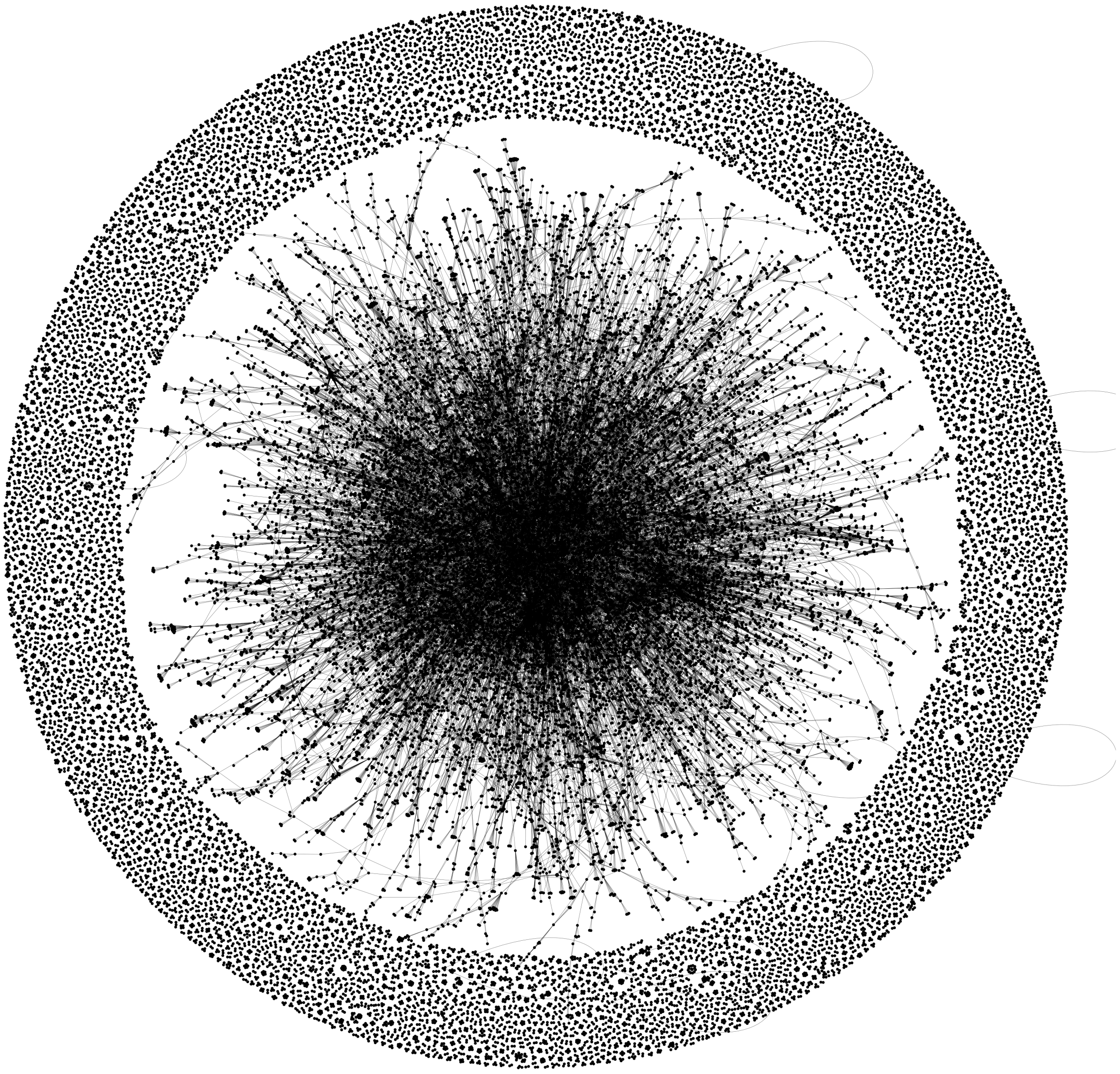

Figure 14 - All Fields as One Graph

The entire network for all three fields has 37.922 articles, when removing the 11.100 isolates we are left with 26.822 articles. The network consist of 45.355 nodes with 84.455 links, distributed between 6.159 sub graphs, and with a diameter of 27, which is a lower diameter than any of the other three fields analyzed separately. This might intuitively seem strange at the network is very much larger than either of the other networks, yet the network consists of the other networks, making it unlikely that the diameter will be any larger than the smallest of the other three. The transitivity is 0.577 and the density is 0.0000821. This network displays a significantly lower transitivity and density than the three fields separately, which is entirely as expected under the assumption that between field collaboration is lower than within-field collaboration, seeing as there will be more disconnected nodes, e.g. researchers who have never collaborated on a scientific article. The network has the highest percentage of isolates out of the entire network amount of nodes compared to the three separate fields, which seems to suggest that this network is sparser than the others. Yet the structure of the field is particularly difficult to see in this graph as a result of the vast amount of information contained within it. We will therefore inspect the degree distribution plot.

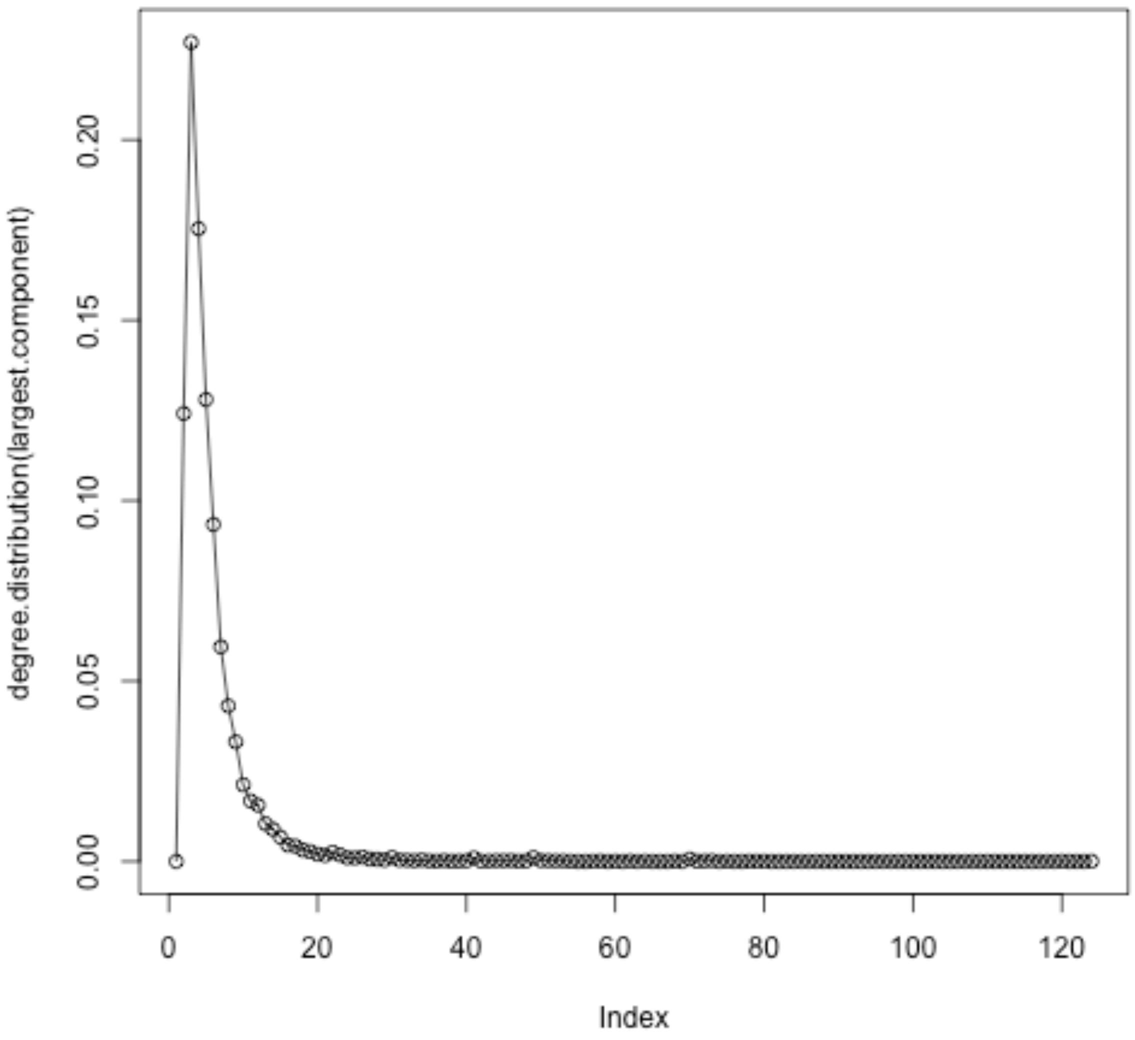

Figure 15 - All Fields Degree Distribution

When inspecting the degree distribution of the entire network we see a familiar shape to the other networks studied here, with the largest share of nodes having two vertices.
To further investigate the network we will again look at the largest component.

The largest component has 25.022 nodes with 60.130 edges. This network has a larger percentage of its nodes in its largest component with 55.19%, dramatically more than the STS network with only 41.87%. The largest component transitivity is 0.55 with a density of 0.000192. Interestingly the transitivity of this network is higher than the largest component transitivity of the STS network, and much smaller than that of both the innovation and entrepreneurship networks. This seems to suggest that the three fields are rather tightly knit as we would expect a lower density if they are more separate.

These researchers represents the largest connect sub graph in the dataset and it is within this subset that most of the information withheld in the research is diffused. Similar research have shown that the percentage of nodes present in the largest component is about 60 per cent (Nascimento, Sander, & Pound, 2003), whereas a study conducted by Newman of four co-authorship networks shows that the smallest largest component of the four contain 57.2 per cent of all the authors (Newman M., 2001), and a study of a network similar to that of this paper in the sense that the field is relatively new where they found that the largest component contains 38 per cent of the network (Liu et al.). The same studies reported clustering coefficients of 0.69, the highest clustering coefficient found by Newman was 0.726, and 0.89, respectively in the largest components. Compared to a clustering coefficient of 0.55 in the largest component in this network. In comparison to similar studies this network can not be said to be very fractured in terms of the size of its largest component, although it seems to be in the low end. The reason for the somewhat small largest component might be that it is a relatively new field or fields, or that there is in another way little cooperation in the field. However, considering that the graph contains data on three more or less separate fields, we would might expect low transitivity and density, yet it seems that this is not the case further pointing towards the fields being closely related.

5.4.1 Centrality and Diffusion of Information

The diffusion of information between the fields happens through this network. Or at least we can argue that much of the wide spread information diffuse through this network, as this is a sample of the journals that are popular to publish in for the fields. In this part we will analyze some aspects of how the diffusion of information happen through this network by looking at some network characteristics, and also analyzing the most important nodes in terms of the different centrality measures introduced in the method chapter. The results from the centrality analysis is presented in the table below, where the top 30 most central nodes according to the four different centrality measures used in this thesis is presented.

Table 14 – Centrality Measures for all fields seen as one

Degree Centrality		Betweenness Centrality		Normalized Closeness Centrality		Eigenvector Centrality	
Fu,PP	123	Kawachi,I	25759222,96	Li,Y	0,17	Fu,PP	1,00
Kawachi,I	119	Lee,J	19945287,13	Lee,H	0,17	Ralston,DA	0,97
Richards,M	79	Wright,M	19088597,17	Lee,J	0,17	Rossi,AM	0,96
Ralston,DA	78	Lee,H	16864697,71	Liu,J	0,17	Maignan,I	0,96
Liu,J	77	Li,Y	15592068,72	Wright,M	0,17	Richards,M	0,96
Rossi,AM	74	Leydesdorff,L	15151407,98	Kim,J	0,17	Terpstra-Tong,J	0,96
Lee,CH	72	Lee,S	11689167,29	Kawachi,I	0,17	Jesuino,JC	0,96
Terpstra-Tong,J	72	Kim,J	11572823,2	Li,J	0,17	Brock,DM	0,96
Jesuino,JC	72	Liu,J	11510281,05	Lee,C	0,17	Srinivasan,N	0,96
Brock,DM	72	Li,J	10212073,48	Park,Y	0,17	Lenartowicz,T	0,96
Srinivasan,N	72	Porter,AL	9358621,051	Lee,S	0,17	Palmer,I	0,96
Maignan,I	71	Chen,YY	9356210,557	Park,HW	0,17	Furrer,O	0,96
Palmer,I	71	Lee,C	9261164,301	Kim,H	0,17	Moon,YL	0,96
Lenartowicz,T	71	Park,HW	8500668,667	Park,J	0,17	Starkus,A	0,96
Wright,M	70	Zhang,J	7474910,38	Kim,K	0,17	Leon-Darder,F	0,96
Furrer,O	70	Lockett,A	7311744,211	Chen,YY	0,17	Chia,HB	0,96
Egri,CP	69	Park,Y	6805720,315	Leydesdorff,L	0,16	Egri,CP	0,96
Ramburuth,P	69	Kim,H	6707867,129	Zhang,Y	0,16	Ramburuth,P	0,96
Dabic,M	69	Autio,E	6688627,992	Moon,J	0,16	Dabic,M	0,96
Hallinger,P	69	Berkman,LF	6155253,299	Li,L	0,16	Castro,FB	0,96
Potocan,VV	69	Rousseau,R	6066859,681	Berkman,LF	0,16	Hallinger,P	0,96
Naoumova,I	69	Zhang,X	5915264,595	Zahra,SA	0,16	Thanh,HV	0,96
Casado,T	69	Zhang,Y	5896169,59	Kim,D	0,16	Potocan,VV	0,96
Castro,FB	69	Glanzel,W	5682752,966	Tong,TW	0,16	Naoumova,I	0,96
Ruiz-Gutierrez,J	69	Hitt,MA	5642018,208	Shin,J	0,16	Ruiz-Gutierrez,J	0,96
Starkus,A	69	Zahra,SA	5581934,843	Kim,Y	0,16	Molteni,M	0,96
Dalgic,T	69	Moon,J	5532012,012	Fu,PP	0,16	Dalgic,T	0,96
Leon-Darder,F	69	Fu,PP	5417510,586	Kim,C	0,16	Casado,T	0,96
Thanh,HV	69	Li,L	5410918,206	George,G	0,16	Lee,CH	0,75
Moon,YL	69	Robinson,J	5385284,247	Porter,AL	0,16	Perrewe,PL	0,73

Betweenness centrality is as discussed a measure on how important a node is in terms of connecting nodes throughout the network. And in this instance it can be seen as a measurement on how important certain nodes are for the diffusion of information between researchers. The highest scoring in terms of betweenness centrality is most important for the diffusion of information between established groups of researchers. These authors will often tie together entire communities, to the extent that their removal might often disconnect sub graph in the largest component from the component. It is important to note that the betweenness centrality measures produced a very different list of the top 30 most central nodes in the network.

Closeness is a measure of the length from a certain node to all other nodes in the network, and thus is a measure of how long it would take for information to spread from a particular node to the other nodes in the network. The nodes that score high on the closeness centrality measure are in the thick of if, centered in the network such that their average geodesic to other nodes is low compared to other nodes. We can note that the top 30 nodes measured in closeness are still very different from degree, yet quite similar to betweenness. There will tend to be a correlation between betweenness centrality and closeness centrality as the shortest path between any node *i* and *g* will often go through a node *k* that is central in terms of closeness. And as betweenness centrality is the probability of diffusion through a particular node, nodes with high closeness centrality will often be ascribed a probability of this happening.

The last centrality measure is the eigenvector centrality, which is a measure of a nodes importance in terms of the importance of the nodes it is adjacent to. The nodes that score high on this metric are therefore often central in the network in terms of their connectedness to other important nodes. However, this will often lead to high density areas of a graph to be over represented in terms of eigenvector centrality, which we can observe here in that the top 30 eigenvector centrality nodes is closer to those of degree, rather than betweenness and closeness.

5.4.2 Outlier Sensitivity in Centrality Measures

As discussed earlier in this study, the most intuitive and perhaps most used way to measure centrality in a network is to count a nodes number of links, also known as degree centrality and we can see that there are a few authors with a very high degree centrality. We also know that there are some outliers in the data in terms of articles that have a large amount of authors as we saw in table 3. In fact two of the articles have as many as 49 articles, immediately giving any author involved with this paper a degree centrality of 48 not accounting for any other articles the researcher might have. We can thus imagine that a researcher with low output who has contributed to one of these articles receives a high centrality score on account of a little bit of work, thus potentially skewing the results dramatically. Obviously a researcher who has co-authored 48 individual papers with 48 different researchers contributes more to the diffusion of ideas through a network than a marginal effort in a paper with a large amount of listed co-authors. In fact there is reason to question the very existence of papers with this amount of authors, as both the need and effectiveness of this mode of scientific production seem to not make intuitive sense. To account for this I will therefore remove the isolates and re-run the centrality calculation to see how this impacts the results.

There are many ways to set a cut-off for what should be included as an outlier, each with strength and weaknesses. I have chosen to set the cut of point at 20 authors. There are papers with authors ranging between 1-18, but not any with 19. I have thus chosen to set the cut-off point where gaps start appearing as we can observe in table 3. This includes 9 articles, and at this point we are at the 99.98th percentile, meaning that 0.02 percent of the articles included is cut. The results of the centrality measures excluding these articles can be seen below:

Table 14 – Centrality Measures for all fields seen as one without outliers

Degree Centrality		Betweenness Centrality		Normalized Closeness Centrality		Eigenvector Centrality	
Kawachi,I	119	Kawachi,I	25 921 225,19	Li,Y	0,17	**Chen,Y**	1,00
Wright,M	70	Lee,J	21 426 607,71	Lee,H	0,17	**Hill,J**	0,82
Lee,S	67	Wright,M	19 664 945,31	Wright,M	0,17	**Wilson,J**	0,82
Lee,H	67	Lee,H	16 798 005,28	Lee,J	0,17	**Yan,L**	0,81
Lee,J	63	Li,Y	16 034 710,17	Kim,J	0,17	**Cleaver,C**	0,81
Leydesdorff,L	63	Leydesdorff,L	15 782 509,98	Kawachi,I	0,17	**Acheson,J**	0,81
Subramanian,SV	57	Kim,J	11 983 531,40	Park,Y	0,17	**Kersula,M**	0,81
Chen,Y	52	Lee,S	11 946 328,63	Li,J	0,17	**Whitsel,L**	0,81
Hitt,MA	50	Li,J	10 117 753,77	Lee,S	0,17	**Wilson,CJ**	0,81
Williams,DR	50	Park,HW	9 068 571,52	Lee,C	0,17	**Congdon,C**	0,81
Marmot,M	50	Chen,YY	8 649 794,77	Kim,H	0,17	**Hayden,A**	0,81
Wang,Y	49	Lee,C	8 433 266,92	Park,HW	0,17	**Hayes,P**	0,81
Kumar,S	47	Zhang,J	7 889 301,22	Liu,J	0,17	**Johnson,T**	0,81
Zhang,J	47	Lockett,A	7 575 783,53	Park,J	0,17	**Turner,R**	0,81
Martikainen,P	46	Autio,E	7 000 433,61	Kim,K	0,16	**Wilson,CL**	0,81
Kim,J	45	Kim,H	6 921 939,97	Moon,J	0,16	**Morehead,G**	0,81
Li,Y	45	Park,Y	6 645 768,00	Chen,YY	0,16	**Steneck,R**	0,81
Chen,YY	45	Liu,J	6 618 136,92	Leydesdorff,L	0,16	**Vadas,R**	0,81
Rousseau,R	45	Berkman,LF	6 344 395,95	Berkman,LF	0,16	**Davis,W**	0,18
Li,J	44	Moon,J	6 065 130,81	Tong,TW	0,16	**Go,VF**	0,18
Liu,Y	44	Zahra,SA	5 893 107,51	Zhang,Y	0,16	**Celentano,DD**	0,18
Porter,AL	44	Rousseau,R	5 866 244,74	Zahra,SA	0,16	**Minh,NL**	0,17
Kim,H	44	Porter,AL	5 731 400,48	Shin,J	0,16	**Frangakis,C**	0,17
Lee,C	44	Zhang,X	5 664 910,83	Kim,Y	0,16	**Vu,PT**	0,17
Chen,HC	44	Zhang,Y	5 663 011,06	Kim,D	0,16	**Quan,VM**	0,17
Chen,J	43	Glanzel,W	5 601 593,52	Kim,C	0,16	**Ha,TV**	0,17
Lockett,A	41	**Kim,K**	5 462 629,25	**Kim,TY**	0,16	**Mo,TT**	0,17
Thelwall,M	41	Hitt,MA	5 357 918,43	Li,L	0,16	**Sripaipan,T**	0,17
Blakely,T	40	Robinson,J	5 324 016,34	**Lee,SH**	0,16	**Zelaya,C**	0,17
Sutter,M	39	**Chen,J**	5 316 589,01	Porter,AL	0,16	**Latkin,CA**	0,17

The researchers that are new to the top 30 list are written in red and bolded. As we can see there are some substantial differences after removing the outliers. First we notice that only two names in the degree centrality list are the same as they were earlier. As betweenness centrality measures is the most often used SNA measure and also widely used in several other bibliometric exercises such as ranking journals and researchers it is interesting to note that this particular measure has a strong bias towards certain characteristics of how a researcher chose to publish. However, apart from rendering the results from the degree centrality measure scores somewhat useless, it points to an interesting limitation with the metric itself when used in co-authorship studies. The results seem to imply that centrality is to an extent dependent on the number of authors who participate in one specific project, the more being the better. Thus being biased toward authors who have contributed to articles with many co-authors. While analyzing a weighted, as opposed to un-weighted, graph would somewhat remedy the issue (depending on the frequency of multiple co-authored articles among specific researchers), it would most likely still persist, albeit to a somewhat lesser extent.

Next we notice that for both betweenness centrality and closeness centrality there are only two new names on the top 30 lists. Further, these new names are situated at the bottom and the order of the rest of the list is to a large extent the same as it was when the outliers were included. This suggest that these metrics, in additional to providing the most useful interpretation of centrality, are the most robust metrics capable of withstanding potentially skewing outliers in an excellent way. The numbers associated with the authors for these two metrics are different, as would be expected, but the main message with these lists are to point out who are most central for the diffusion of ideas through the network.

The most striking difference is to be found with the eigenvector centrality measure. None of the researchers are still there after removing the outliers, suggesting that all of the authors listed previously have co-authored one of the outlier articles. When considering how the eigenvector centrality measure works, as detailed in the method chapter, it does not seem so strange that this might happen. Eigenvector centrality takes into account the number of 'in-links' (e.g. co-authors) when determining the 'strength' (e.g. centrality) of a node. Further, it takes into account the strength of the nodes that connect to it, when determining the strength of a specific node. As an example consider author *i* that has written one article with author *k* and one article with author *j* giving her an eigenvector score of *x*. Now imagine that author *k* writes an article with author *l*, thus improving the score of author *k*. As *k* links to *i*, the score of *i* will also increase as a consequence of this so that the score of *i* is now *x* plus some fraction of the score of *k*. Further, now that *i* has a higher score all of her links also get a higher score in the process, including author *k*. Thus increasing the score for everybody in the cluster but the effect is diminishing the further away the nodes are. It is therefore reasonable to assume that this effect, which is a results of tightly knit clusters, have taken place. They give each other 'strength' through having many links, and these links give each other a yet higher score.

Both the simplest and the most complex measure for centrality have thus proven to be non reliable when dealing with outliers, while the usefulness of betweenness centrality and closeness centrality has been proven to provide reliable and valid results. Interestingly, the most used measure is also badly equipped to handle statistical noise and outliers. We can also note that the transitivity, that is the propensity for researcher *i* and *j* to be connected when node *k* is connected to both *i* and *j*, has been reduced from 0.5503913 to 0.4127194. Which would make sense as we have removed the nine most tightly knit clusters in the network.

5.4.3 Gatekeepers in a Network

Studies have been done that show that nodes with a low eigenvector centrality and a high betweenness centrality are likely gate keepers in a network, whereas nodes with high eigenvector centrality and low betweenness are close to or in direct contact with the most important nodes in the network. Data scientist Drew Conway have studied and gives talks on this theme[1]. To investigate through whom the diffusion of information takes place in this network we will therefore plot betweenness and eigenvector centrality against each other, to identify these gate keeper nodes.

[1] Matt Bogard. "Using Twitter to Demonstrate Basic Concepts from Network Analysis" Jan. 2010. Available at: http://works.bepress.com/matt_bogard/9

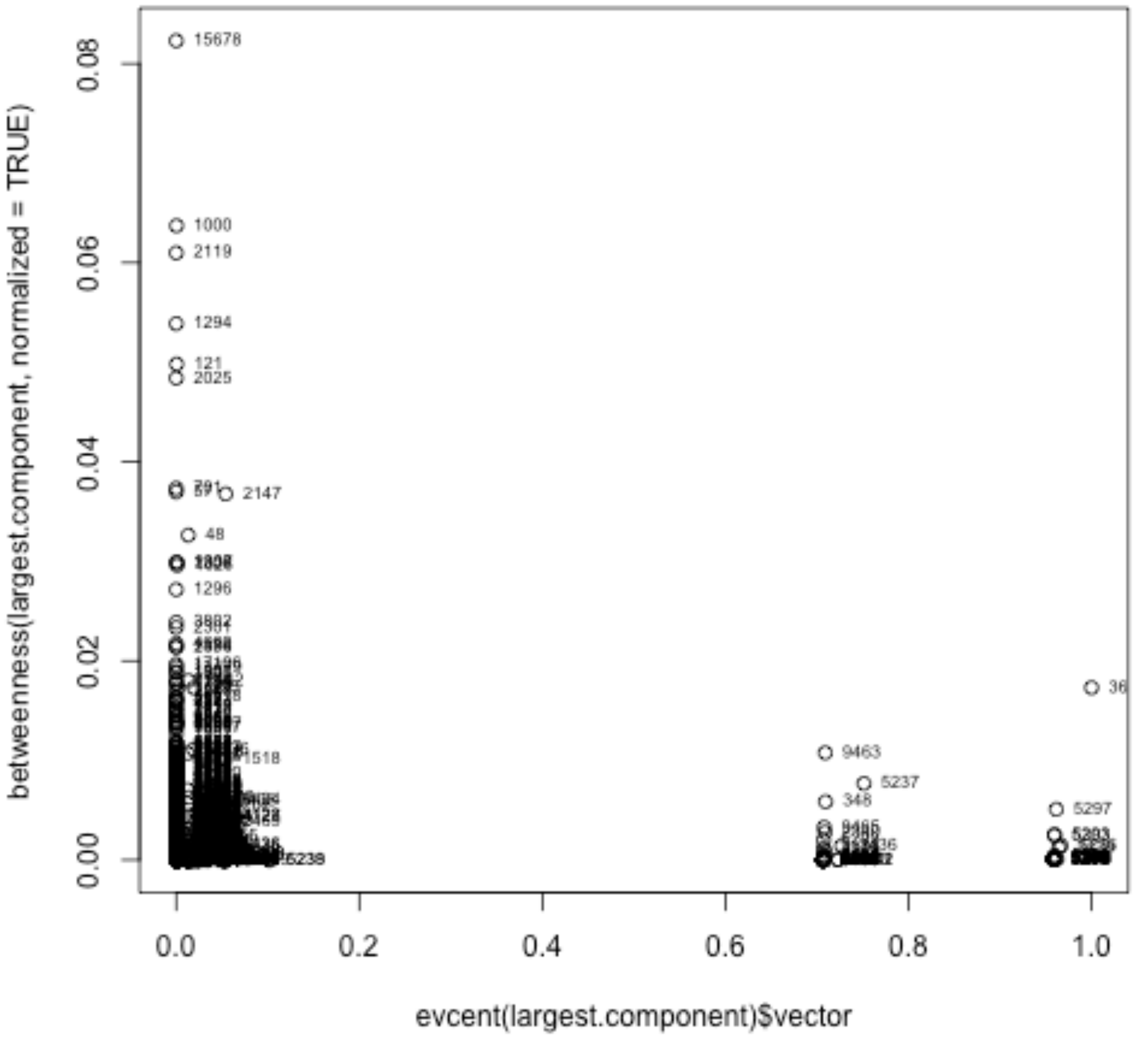

Figure 16 - Important Gatekeepers

Immediately we can see that no nodes obtain both high betweenness and high eigenvector centrality scores. This leads us into the reason behind why nodes with high eigenvector centrality scores and low betweenness centrality scores are central in the network in that they are connected to a lot of nodes who are in turn connected to a lot of nodes, thus representing a cluster of important nodes. While nodes with low eigenvector and high betweenness are important gatekeepers. These nodes lie on the geodesic of several pairs of nodes, yet are only connected to a few other nodes. Further, we can see that the distribution of eigenvector centrality scores is very clustered compared to the more continuous distribution of betweenness centrality scores. This suggests that there are some, specifically two, clusters of highly important nodes that are located very centrally in the network. We can also notice that the peculiarities of how the eigenvector centrality measure is calculated will lead to nodes connected to each other get high scores on account of their adjacent nodes high scores. However, the nodes that score highly on the betweenness centrality measure all have very low eigenvector scores. These nodes are the same nodes that are presented in the table above in the betweenness centrality column. This confirms the notion that the researchers listed there indeed are extremely important in connecting the network together and are important gatekeepers for the diffusion of information through the network. These authors are thus the most important in terms of diffusion in the network.

We can also note than the large majority of nodes have both low betweenness score and low eigenvector scores. Thus, there are some nodes that exhibit a significantly larger extent of centrality than most nodes. This suggests that there are certain nodes that are very central for the diffusion of information through the network. Yet a low betweenness centrality score does not necessarily mean that the nodes are unimportant in the network. We must remember that betweenness centrality is the probability for information being diffused through a network. Meaning that if the geodesic between any two nodes can take a large amount of paths, the betweenness centrality score would be reduced. We would therefore expect nodes with a low betweenness score to be connected to several nodes, and with a high degree centrality.

The next part will investigate the communities present in the network.

5.4.4 Community Detection

A common investigation in social network analysis is community detection. These are algorithms that are created to detect clusters of nodes that are close to each other. The algorithm used here is called the Fastgreedy community detection algorithms, which is based on modularity and maximize the within group similarity while maximize between group dissimilarity. The algorithm is used because of its fast nature. Other algorithms would be infeasible to use here considering the large data set.

Below is a representation of the network with its clusters. The amount of clusters are very high, and there are no signs of three separate and distinct sub-fields in it. This would seem to suggest that the fields are not far removed from each other. Yet we must remember that the overlap between the three fields in terms of their most central journals would naturally lead to difficulties in distinguishing the three fields clearly.

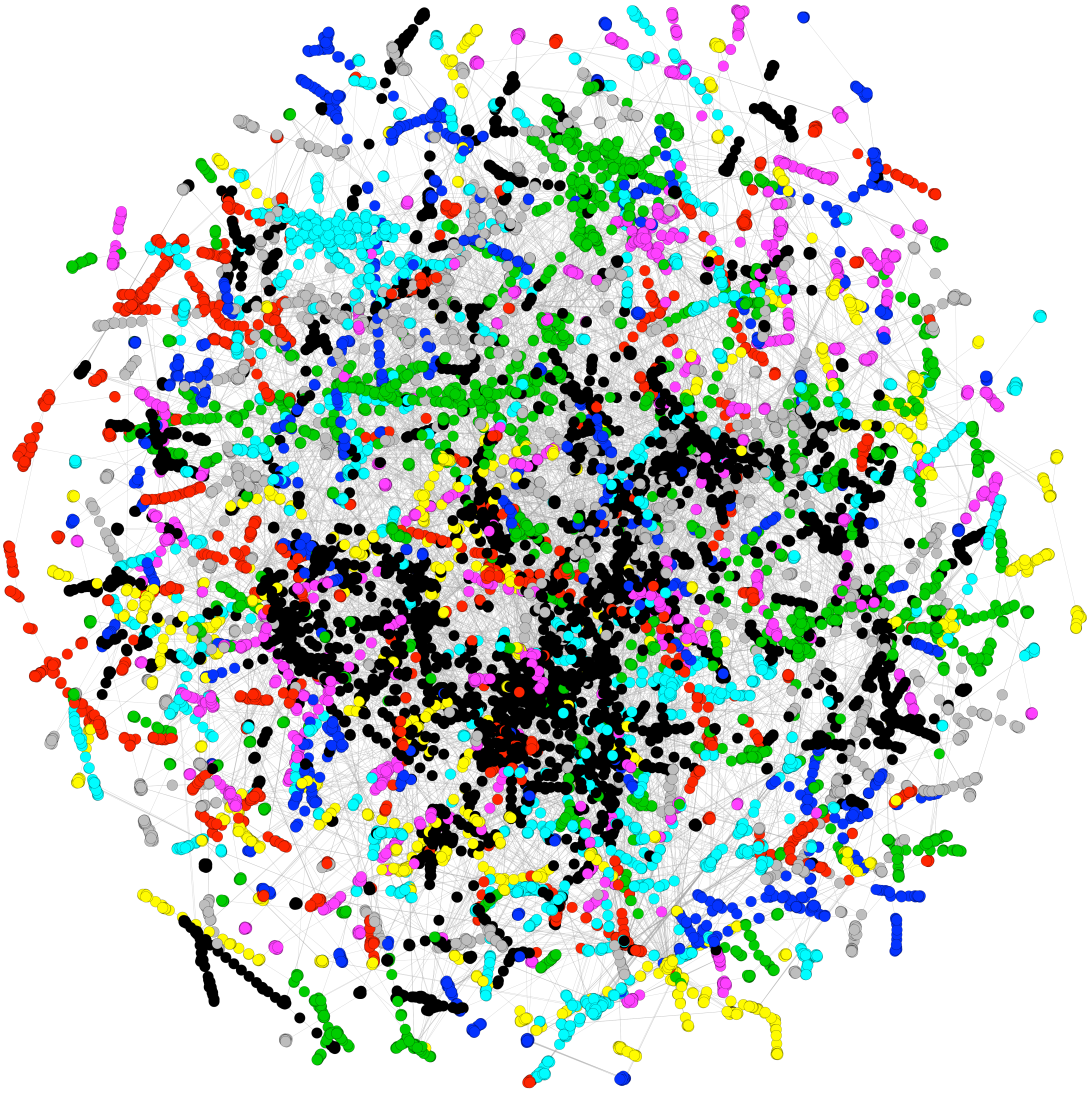

Figure 17 - Community Detection All Fields

Clearly, there are no communities that separate themselves significantly, and no three fields with clearly and organized divides. The community detection algorithms instead point to a field with several small communities of researchers that do not collaborate extensively with each other, as we have also seen from the global centrality measures. Where earlier research have shown that there is a clear distinction between the fields when looking at important contributions, there does not seem to be any such clear distinction when looking at the actual researchers that are a part of the fields.

The studies conducted earlier have found many characteristics of the fields, such as the most important handbooks, the most important journals, and the most important publications.

Yet none of them have tried to identify the most important scholars now, instead opting for identifying the historically most important researchers through using publications as nodes, something we are able to attempt through this study where the diffusion of information is important. Further I have identified the most central gatekeepers through whom diffusion of innovation is most likely to take place, a novel addition to the current research on the knowledge base of innovation, entrepreneurship and STS. Additionally the community detection provided inconclusive answers in terms of the relationship between the three fields. It did, however, provide some additional insights into the structure of the fields seen as one. In the next chapter I will discuss these findings and conclude.

6 Discussions and Conclusions

In this section I will discuss and draw my own conclusions, and compare findings with earlier research. In this thesis I have explored several research questions using different methodologies and approaches to gain important insight beyond the scope of earlier research. By utilizing a wide methodological range I am able to show the structure of cooperation between-field and within-field in a more granular and holistic way.

Earlier research have to a large extent answered how new scientific fields are created and how they develop in early stages. As innovation studies, entrepreneurship and STS alike are new in terms of disciplinary age, there is still much to be discovered. The question of how and under what circumstances a field of research is established and how it develops in its early stages have been theorized and discussed in this earlier research:

> New research fields in the social and natural sciences often originate at the interstices of established disciplines when researchers from neighboring disciplines, with differing disciplinary perspectives, realize they share a common interest. Over time, by working together, they may start to develop their own shared conceptual, methodological and analytical frameworks. This then allows them to move from publishing in journals of their 'parent' disciplines and to establish their own journals, professional associations, specialized university departments or units […] (Martin, Nightingale, & Yegros-Yegros, 2012).

Following this mode of thought we could assume that the fields would be fractured and disbanded in its early beginnings. Later on taking the form of a true academic discipline. This is reflected among other researchers as well:

> […] one important way in which social science renews itself is by responding to the emergence of new 'problems', pointing to the scarcity or lack of relevance of existing knowledge. Such challenges, especially when accompanied by new resources, may attract resources from a variety of backgrounds and may eventually lead to the creation of new research communities, with institutions and organisations designed to promote scientific progress in the area (Fagerberg, Fosaas, & Sapprasert, 2012)

The question is whether the three fields have progressed as outlined earlier, i.e. that scholars from different disciplines have met through their shared interest in a research theme, and from there begun a separation. Or if this process has been divided in three separate, yet somewhat related, instances. i.e. that innovation studies have attracted scholars from certain disciplines, entrepreneurship from others, and STS from yet other, so that there is in fact three separate fields?

6.1 How do the Fields Differ and What Separates Them?

As we have seen there is a significant overlap between the top journals between the fields. There is especially a large overlap between innovation studies and entrepreneurship, while STS is more secluded from the other two. Perhaps surprisingly, earlier research have concluded that entrepreneurship is far more removed from innovation that what seems to be the case when looking at the top publications:

> Despite the fact that entrepreneurship has borrowed theories from other fields and many scholars from other disciplines have migrated into entrepreneurship research, it has remained surprisingly disconnected from the neighboring field of innovation studies. Despite common roots in Schumpeter and some interrelated topics such as innovation management (corporate entrepreneurship) and an interest in technology-based firms, 'entrepreneurship' and 'innovation' have evolved over time as two largely separate research fields (Landström, Harirchi, & Åström, 2012).

In this study there is evidence that the two fields are largely connected, and that any discrepancies between them have either disappeared over time, or have in fact not existed. Because the time and methods differ from previous studies, it is difficult to conclude which interpretation is correct. It could also be both. However, overlap between journals is not the only way we have investigated how fields are connected. We must also investigate other ways in which fields are similar or differentiate themselves from each other.

To answer the question of whether the three fields overlap in terms of research themes I have examined the overlap in the most central journals for the three fields as well as a text-mining approach on the keywords for the articles in the journals. This method shows that there is a significant overlap in research themes between innovation and entrepreneurship, while STS seems to be more separate. Earlier studies using citations as nodes have to a large extent drawn the same conclusions:

> "moreover, we see that the INN field is positioned in the middle, corresponding to its higher tendency to be involved in between-field citations than the other two fields. INN and ENT are also generally closer than INN and STS, and there is actually a fair amount of overlap between INN and ENT on the vertical axis. The overlap between INN and STS is much weaker, represented by only a handful of documents […]. The

> impression that this gives is that the complete citation network indeed consists of subfields, which largely correspond to the three fields under study" (Bhupatiraju et al. 2012).

Earlier studies have sought to probe the historical origins of the fields using books and journals as nodes and citations as links. Here I have used the authors themselves as nodes and co-authorship as links. The conclusion is therefore that when looking at books, they are separated by fields in terms of what they are about. But when looking at authors, many of them seem to be involved in several of the fields simultaneously. This indicate that a specific researcher has a tendency to publish a work that fits snugly within one of the three fields and then citing canonical works within the same field, yet the same researcher may very likely publish other work that fits snugly into any one of the other fields as well in other work. The canonical works of the fields are therefore clearly in their own fields, but the researchers themselves are often not situated within one single field. Thus, a scholar is more likely to publish in innovation and entrepreneurship, rather than in the STS and innovation etc. Even though STS and innovation are more closely related in terms of historical origins, entrepreneurship and innovation are more related in terms of research field and researchers tendency to publish in specific journals, at least in the last ten years. This actually indicates a convergence, despite of their different origins and viewpoints.

When investigating differences in terms of structure and cooperation within the fields we have uncovered differences that further points to the conclusion that innovation and entrepreneurship are not as far removed as earlier research have argued, and that STS is more separate and thus more clearly constitutes a separate field. Of course, the results of the social network analysis and the metrics and indicators analyzed and reported above is to some extent influenced by the data set the analysis is based on. As innovation and entrepreneurship shares as many top journals as they do, it will obviously also tend to make the results similar, because there is significant overlap. However, there are also large differences between them. All the fields show clear signs of being highly fractured in terms of there being little cooperation as reflected in the density metrics. Still, all the fields exhibit signs of being very tightly knit in terms of transitivity. The reason for this is likely that there are small groups of researchers that often choose to collaborate with other researchers with common interests, educational background and/or share the same institutional affiliations. Thus, the fields are likely not as multi-disciplinary as it might seem on an individual level, but as fields they are of course highly multi-disciplinary.

Despite of the partially overlapping data, the fields exhibit many signs of having different structures and levels of cooperation. Especially scholars in STS seem to prefer to work more alone, than compared to innovation and entrepreneurship. Innovation and entrepreneurship are more similar, although innovation shows signs of being more prone to cooperation, especially as the percentage of isolates in the dataset is significantly lower than the other fields.

Despite of these differences the fields are all fractured in terms of the level of cooperation in them. As discussed in the theory chapter all the fields are new additions to academia and we might therefore suspect that that the fields are not highly cooperative. Also, we discussed the possibility that such fields would consist of several tightly knit clusters of research as measured by the transitivity scores. This has been shown to be the case, and all the fields, as well as the fields seen as one large field, all show clear signs of this phenomenon. To sum up the structure and cooperative activities between researchers we can conclude that the fields are both highly clustered, yet extremely sparse in their network of co-authorship collaboration. Even though they have many differences and peculiarities, they still seem to be part of an over-arching field. Additionally they show clear signs of still being young fields, not yet having fully transformed into separate ones. For this to happen, more collaboration and tighter communities of researchers across disciplines and sub fields would have to be established. However, this study do not show the historical development of collaboration and further research will have to be done to find more evidence.

6.2 The Diffusion of Innovation Through Networks

Many of the measures presented in the analysis are measures of collaboration between researchers. By comparing these metrics we can show how the three fields differ in terms of diffusion of tacit knowledge through the network of related researchers. As discussed in the paragraph above the global centrality measures of the network have shown that all the fields are simultaneously sparse and highly clustered, in this section I will discuss the local centrality measures.

The list of the most central authors provide us with insight into who are the most important researchers in the field in terms of being situated at important points in the network, where much of the information travel through. I have especially looked at betweenness centrality as this measure specifically investigates whether a specific researcher is important for the diffusion of information. Most researchers are not particularly important, while a few exhibit signs of tying large and/or many smaller clusters together. The overview of the most central authors acts as a snap-shot of who the most important scholars in the field are today. Further, we looked at how we can identify the most important gate-keepers in the network through plotting betweenness centrality against eigenvector centrality. This investigation further showed us that there are a few extremely important researchers that are crucial for the effective diffusion of information in the network. At least important for the tacit knowledge often shared when cooperating on research activities. Most, however, are not particularly important for this diffusion, which is not surprising when considering the extreme scarcity of the networks.

The last method used in this thesis is a cluster analysis. As information is likely to be distributed to a larger extent within clusters than between clusters, we could expect some clusters to show signs of belonging to a specific sub-field. This is not the case. The cluster analysis shows no clear signs of any three main fields. This finding therefore support the discussion above about the fields being separate in terms of canonical works and the historical trail of ideas, not necessarily in terms of the individual researcher. The fields do collaborate with each other quite a lot when looking at individual researchers. Researchers are likely to publish work that is clearly viewed as belonging to a specific discipline while also publishing work that is clearly viewed as belonging to another one. The collaboration between the fields when analyzing other data, such as books or handbooks, would therefore lead us to believe that there are few attempts at bridging the differences between them. This study indicates that the fields are closer together than previously when observing and analyzing individual researcher behavior. A more fruitful approach, seeing as fields are not separate entities by themselves, but rather built and defined by the individuals that inhabit them.

6.3 Observations and Further Research

I have uncovered some methodological weaknesses related to centrality measures. Especially degree centrality, the most intuitive and most commonly used indicator as well as eigenvector centrality, are highly susceptible to interference from outliers. By removing outliers from the dataset, the view of the field becomes more reliable and valid. This view is supported by the fact that the best indicator for determining important nodes in terms of diffusion, betweenness centrality, is not affected considerably by the removal of the outliers.

This study gives a holistic and the most thorough investigation into the differences and similarities between the three fields done to date. However, there are further steps that could be taken to gain an even better understanding of the development of the fields. A way forward is to run the same analyses that I have done in this study for different periods of time and look at the development of the fields. By calculating transitivity, clustering, diameter etc. for every ten year period we could see if the fields are indeed converging or diverging in terms of collaborative activities, and thus whether the fields are indeed developing into steadily more clear academic fields. More research is still needed to fully grasp the intricacies of new academic fields in general, and innovation, entrepreneurship and STS in particular.

Bibliography

R and R packages

Adrian A. Dragulescu (2014). xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files. R package version 0.5.7. http://CRAN.R-project.org/package=xlsx

Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org

Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

Ian Fellows (2013). wordcloud: Word Clouds. R package version 2.4. http://CRAN.R-project.org/package=wordcloud

Ingo Feinerer and Kurt Hornik (2014). tm: Text Mining Package. R package version 0.5-10. http://CRAN.R-project.org/package=tm

Paradis E., Claude J. & Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289-290.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Articles and books

Acedo, F. J., Barroso, C., Christóbal, C., & Galán, J. L. (2006). Co-Authorship in Management and Organizational Studies: An Empirical and Network Analysis. *Journal of Management Studies* (43), 957-983.

Barbási, A.-L. (2002). *Linked - The New Science of Networks.* Cambridge: Perseus Publishing.

Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The Architecture of Complex Weighted Networks. *Proc. Natl. Acad. Sci. USA 101, 3747* .

Bhupatiraju, S., Nomaler, Ö., Triulzi, G., & Verspagen, B. (2012). Knowledge Flows - Analyzing the core literature of innovation, entrepreneurship and science and technology studies. *Research Policy* (41), 1205-1218.

Bonachic, P. (1987). Power and Centrality: A Family of Measures. *American Journal of Sociology* (92), 1170-1182.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network anlysis in the social sciences. *Science* (323), 892-895.

Brin, S., & Page, L. The PageRank Citation Ranking: Bringing Order to the Web. *Stanford InfoLab , 1988*.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structures in very large networks. *http://arxiv.org/pdf/cond-mat/0408187v2.pdf* .

Collins, H. M. (1974). The Tea Set: Tacit Knowledge and Scientific Networks. *Science Studies* , 165-186.

Cunningham, S. J., & Dillon, S. M. (1997). Authorship Patterns in Information Systems. *Scientometrics* (39), 19-27.

Crane, D. (1972). Invisible Colleges*; Diffusion of knowledge in scientific communities*. Chicago, University of Chicago Press.

de Solla Price, D. (1963). *Little Science, Big Science.* New York: Columbia University Press.

Edge, D. (1979). Quantitative measures of communication in science: A critical review. *History of science* (17), ss. 102-134.

Elo, S., & Kyngäs, H. (2007). The qualitative content analysis process . *Journal of Advanced Nursing* (62), 107-115.

Fagerberg, J. (2005). Innovation: A Guide to the Literature. I J. Fagerberg, D. C. Mowevry, & R. R. Nelson, *The Oxford Handbook of Innovation* (ss. 1-26). Oxford: Oxford University Press.

Fagerberg, J., & Verspagen, B. (2009). Innovation studies - The emerging structure of a new scientific field. *Research Policy* (38), ss. 218-233.

Fagerberg, J., Fosaas, M., & Sapprasert, K. (2012). Innovation: Exploring the knowledge base. *Research Policy* (41), 1132-1153.

Fagerberg, J., Martin, B. R., & Andersen, E. S. (2013). *Innovation Studies: Evolution and Future Challenges.* Oxford: Oxford University Press.

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge: Cambridge University Press.

Fortunato, S. (2010). Community Detection in Graphs. *http://arxiv.org/pdf/0906.0612v2.pdf* .

Frame, J., & Carpenter, M. (1979). International research collaboration. *Science* (9), ss. 481-487.

Freeman, L. C. (1979). Centrality in Social Networks: Conceptual Clarification. *Social Networks* (1), 215-239.

Garfield, E., Sher, I. H., & Torpie, R. J. (1964). *The Use of Citation Data in Writing the History of Science.* Philadelphia: The Institute for Scientific Information.

Gauthier, É. (1998). *Bibliometrics Applied to Public Policy: Methods and Examples.* Science and Technology Redesign Project. Statistics Canada.

Hearst, M. (2003, October). What is Text Mining. *SIMS, UC Berkeley* .

Hoffman, D. L., & Holbrook, M. B. (1993). The Intellectual Structure of Consumer Research: A Bibliometric Study of Author Cocitations in the First 15 Years of the Journal of Consumer Research. *Journal of Consumer Research* (4), 505-517.

Jackson, M. O. (2008). *Social and Economic Networks.* Princeton: Princeton University Press.

Kariv, D. (2011). *Entrepreneruship: An International Introduction.* New York: Routledge.

Katz, S., & Martin, B. R. (1997). What is Research Collaboration? *Research Policy* (26), 1-18.

Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology* (Vol. 1984). California: Sage Publications Ltd.

Landström, H., Harirchi, G., & Åström, F. (2012). Enrepreneurship: Exploring the knowledge base. *Research Policy* (41), 1154-1181.

Leopold, E., May, M., & Paaß, G. (2005). Data Mining and Text Mining for Science & Technology Research. I H. F. Moed, W. Glänzel, & U. Schmoch, *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems* (ss. 187-2014). Dordrecht : Springer Science + Business Media, Inc. .

Liu, X., Bollen, J., Nelson, M. L., & Ven de Sompel, H. Co-authorship Networks in the Digital Library Research Community. *Prepaper submitted to Elsevier Science , 2008.*

Lundvall, B.-Å., & Borrás, S. (2005). Science, Technology and Innovation Policy. I J. Fagerberg, D. C. Mowery, & R. R. Nelson, *The Oxford Handbook of Innovation* (ss. 599-631). Oxford: Oxford University Press.

Martin, B. R., Nightingale, P., & Yegros-Yegros, A. (2012). Science and technology studies: Exploring the knowledge base. *Research Policy* (41), 1182-1204.

Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics* (3), ss. 363-377.

Nascimento, M. A., Sander, J., & Pound, J. (2003). Analysis of SIGMOD's coauthorship graph. *SIGMOD Record , 32* (3).

Newman, M. J. (2001). The structure of scientific collaboration networks. *PNAS* , 404-409.

Newman, M. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E* (64), 016132.

Newman, M. (2001). Scientific Collaboration Networks: I. Network construction and fundamental results. *Physical Review E.* (64), 016131.

Pao, M. L. (1992). Global and local collaborators: a study of scientific collaboration. *Information processing & management* (28), 99-109.

Polanyi, M. (1956). *Personal Knowledge - Towards a Post-Critical Philosophy.* Chicago: University of Chicago Press.

Porter, M. (1980). An algorithm for suffix stripping. *Program* (3), 130-137.

Powell, W. W., & Grodal, S. (2005). Networks of Innovators. I J. Fagerberg, D. C. Mowery, & R. R. Nelson, *The Oxford Handbook of Innovation* (ss. 56-85). Oxford: Oxford University PRess.

Price, D. J. (1987). *Little Science, Big Science ... and Beyond.* New York: Columbia University Press.

Ramos-Rodríguez, A.-R., & Ruíz-Navarro, J. (2004). Changes in the Intellectual Structure of Strategic Management Research: A Bibliometric Study of the Strategic Management Journal, 1980-2000. *Strategic Management Journal* (25), 981-1004.

Rogers, E. M. (2005). *Diffusion of Innovations.* New York: Free Press.

Scott, J. (2013). *Social Network analyisis.* London: SAGE Publications Inc.

Scott, J. (1988). Social Network Analysis. *Sociology* (22), 109-127.

Sismondo, S. (2010). *An Introduction to Science and Technology Studies.* West Sussex: Blackwell Publishing Ltd.

Small, H. (1973). Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. *Journal of the American Society for Information Science* , 265-269.

Small, H., & Griffith, B. c. (1974). The Structures of Scientific Literatures I: Identifying and Graphing Specialties. *Science Studies* , 17-40.

Smeaton, A. F., Keogh, G., Gurrin, C., McDonald, K., & Sødring, T. (2002). Analysis of Papers from Twenty-Five Years of SIGIR Conferences: What Have We Been Doing for the Last Quarter of a Century ? *SIGIR Forum* (36).

Smith, M. (1958). The trend toward multiple authorship in psychology. *American Psychologists* (13), 596-599.

Tan, A.-H. (u.d.). Text Mining: THe state of the art and the challenges. *white paper* .

Tiegland, R. (2003). 'Knowledge Networking: Structure and Performance in Networks of Practice' Dissertation for the degree of Doctor of Philosophy, Stockholm School of Economics, Stockholm

Uddin, S., Hossain, L., Abbasi, A., & Rasmussen, K. (2012). Trend and efficiency of co-authorship network. *Scientometrics* (90), 687-699.

Wasserman, S., & Faus, K. (1994). *Social Network Analysis: Methods and application.* Cambridge University Press.

Watts, D. J. (2001). *Small Worlds: The Dynamic of Networks Between Order and Randomness.* Princeton: Princeton University Press.

Weiguo Fan, L., & Wallace, S. R. (2006, September). Tapping the power of text mining. *Communication of the ACM* (49), ss. 77-82.