A Network Analysis of Knowledge Production in Transportation: Exploring Co-evolution of Path- and Place-dependency

Master Thesis Report

Tjapko Struik
Student number: 3847713
E-mail address: T.Struik@students.uu.nl

Science and Innovation Management
University of Utrecht
Master Thesis (45 ECTS)
Supervisor: dr. Gaston Heimeriks

February 28, 2014

Abstract

This research presents a network analysis of knowledge production. We analyse a 2-mode network of cities and topics, obtained by a bibliometric case study of publication record of the scientific field of transportation. The main objective is to explore knowledge diffusion by path- and place-dependent patterns. Based on cognitive and geographic proximity of cities and topics we design prediction models to measure this exploratory efforts. The topics we study are based on the title words of articles, written by scientists that work in a certain city while publishing their knowledge. The 2-mode network we study has been transformed in both a city and topic network for further analysis. In these networks we explore the concepts of geographic and cognitive proximity. We linked the optimisation of proximities to the notion of absorptive capacity, which brings the network of cities and topics together. Although this link has been explored before, no quantitative support was found so far. This research shows a high significance level of χ^2-statistics for path- and place-dependency. Since a prediction should be as precise and specific as possible, we also evaluate the prediction models with the more intuitive F-score, which is determined by a weighted combination of a prediction's precision and recall. The precision of our path-dependent prediction model is high for small prediction sets, while the place-dependent model is not very precise. The recall is for both models relatively low. When we optimise the topic proximity with a co-evolutionary approach of the behaviour of cities and topics, the recall of the path-dependent model increases. However, this is at the cost of its precision. A concrete preference for the relative importance of precision and recall is required to determine an optimal design of the prediction model. This design can be either path-dependent, place-dependent or based on a co-evolution of cities and topics.

Contents

Preface

This master thesis is a report of the final research project of my master 'Science and Innovation Management'. The research brought together some of the innovation theories that were taught during the first year of the master. The type of research is based on my interest in graph theory and network analysis, which I followed several courses about in my bachelor 'Mathematics'.

By means of a network of cities and topics, I aimed to represent the current state of scientific developments of a technological field. Based on innovation theories about cognitive and geographic proximity, I designed prediction models of path- and place-dependency. By testing these models, I followed some research patterns from my supervisor dr. Gaston Heimeriks. He explores path- and place-dependency to find correlations between proximity of topics and cities, and developments of them in a scientific field.

The design of the models required several choices with respect to both city and topic networks. At various moments a cut-off in data must be made to keep the analysis feasible, which can have meaningful consequences. Furthermore, optimisation of similarity in topics and cities needed to be tested. Argumentation for those choices required involvement of an additional innovation theory.

Therefore, I finally explored the concept of absorptive capacity, which is mostly used for firms. For this research, I have adapted it to a property of cities. Absorptive capacity reflects the learning ability of scientists in a city and therefore brings together the role of cities topics. While for prediction purposes two separate networks are used, the city network and topic network are practically dual to each other, this final theory will bring them back together again.

Chapter 1

Introduction

The research presented here aims to explain evolution of knowledge production in the field of transportation. For this purpose, we have explore two concepts, path-dependency and place-dependency, and integrate them to design prediction models for diffusion of knowledge. In the background section (1.1) we explain the nature of these concepts and why there is need to extend previous research. In the second section (1.2) we introduce a research question, based on the need to integrate existing explanations of knowledge production. Subsequently, we describe an approach that complements results from innovation literature. This approach involves a network analysis, applied to a bibliometric case study. The case we study is knowledge production in the scientific field of transportation. We analyse 9451 article records that were published during the period from 1998 until 2011 in 'Transportation Research Record', which is quantitatively the core journal in transport literature.

1.1 Background

Knowledge production is a process where scientists generate knowledge within a communication system (Beal et al., 1986). This scientific knowledge is measured by the publications of scientists in journals. It is important to study what scientific research has been performed in a scientific field to understand technological developments (Cooke, 2007; Asheim and Coenen, 2006; Nelson and Winter, 1977). A common approach in innovation studies with concern to this examination, is to build an innovation system. Such a system brings all aspects of a technological field together (Tuominen and Himanen, 2007). The role of scientists has been stressed to be a relevant part in the context of these innovation systems (Hekkert et al., 2007; Lundvall, 2007). Scientists are the key factor in the changing configuration of topics that are studied in a scientific field. To capture the essence of this change, we should focus on diffusion of scientific knowledge to explain knowledge production (Shapin, 1998).

Knowledge diffusion is defined as travel of topics across the world (Shapin, 1998; Leydesdorff and Rafols, 2011). Theoretically, the objective of understanding this behaviour of topics is part of building an innovation system of knowledge production. From a more practical point of view, knowledge diffusion is a concept that policy makers need to understand as well. More particular, policy makers face the problem how to influence the diversification and specialisation of knowledge in their own region (Boschma, 2013). To address that problem, they need to find patterns in knowledge production. In this research, patterns in knowledge production are explored by a case study, as it varies per technological field what factors determine technological developments (Schartinger et al., 2002). As a result of new developments in intelligent transport systems, transportation is currently a popular field for innovation scholars (European Union, 2010; Giannopoulos, 2009; Quddus et al., 2007; Sai et al., 2013; Yan et al., 2012; Zhang, 2008). Due to this popularity, there is a sufficient amount of recent articles published to evaluate knowledge production in this field. We research patterns in knowledge production by designing prediction models for knowledge diffusion. These models contribute to an explanation of current knowledge production as a result of diffusion.

To accomplish their policy targets, it is an advantage for policy makers to be aware of the current knowledge diffusion (Tuominen and Himanen, 2007). In innovation studies, this diffusion is expressed by topics that are new in a scientific field; or topics that are new to a certain city or region (Van den Besselaar and Heimeriks, 2006; Lei and Xin, 2011; Boschma et al., 2013). In a system that describes evolutionary knowledge production, knowledge diffusion is characterised by the emergence of new combinations of topics and cities. By keeping track of these new combinations, we can capture the effect of diffusion on scientific knowledge production. This approach results in an evolutionary network that describes the current knowledge and enables forecasting evolution of developments in transportation. A network that combines cities and topics is part of an encompassing innovation system that highlights the role of institutional structures and the importance of actors for the emergence of technological innovations (Musiolik et al., 2012, p. 1032-1033). Our research focuses on the role of scientists, whom represent one type of actors. To understand how their part fits into an innovation system of transportation, Tuominen and Ahlqvist (2010, p. 120) emphasise the importance to integrate the technology developments better with societal developments and transport policy design. In their paper, the objective is to create a method that meets these conditions. Such a method can function as a tool for policy makers to synchronise transport policy with developments in both society and technology. Constructing an accurate method to explain knowledge production is a particularly challenging problem as the technological frontier evolves rapidly. Knowing what factors influence evolutionary behaviour of knowledge diffusion has improved our view on the nature of technological developments. An evaluation of their effects brings us closer to solving the problem of understanding knowledge production (Urbina and Wolshon, 2003; Tuominen and Ahlqvist, 2010).

The design of our prediction models is based on innovation theories about proximity. These models will be refined by delving into the ability of scientists to acquire new knowledge knowledge, which is referred to by absorptive capacity. The diffusion of knowledge in a scientific field is often explained and predicted by geographic and cognitive proximity of scientists, which influence the likeliness of scientists sharing knowledge (Boschma, 2005; Nooteboom et al., 2007). By combining geographic and cognitive perspectives, we are able to study activities of scientists not only on an individual level, but also at the level of cities.

Geographic proximity is determined by distance between cities where scientists carry out their research; while cognitive proximity is based on the relatedness of topics that scientists study. Boschma and Fornahl (2011) argue that there is a relatively high interaction between scientists in the same region. Since face to face contact is quite important in knowledge sharing, universities located close to each other will influence each other more than universities that are far apart (Katz, 1994). This is based on a geographic analyses of topics, in which their content is not taken into consideration (Almeida and Kogut, 1999; Boschma, 2004; Hoekman et al., 2009). We regard this prediction of evolution of topics, based on geographic history, as place-dependency.

Cognitive proximity, however, is increasingly influential due to globalisation, which has caused geographic proximity of universities to be less important than in the past (Balland et al., 2011; Lei and Xin, 2011; Boschma et al., 2013; Wu, 2013). If we would would not take the geographical aspect into account, scientists would expand their present knowledge by building on topics they studied before; also referred to as knowledge accumulation. Present knowledge is a cognitive state of scientists that can be described by the topics they published about in scientific journals. Scientists expand their knowledge base by studying similar topics. These topics can be entirely new, but they can also be a result of scientists adopting each other's topics. This similarity in topics scientists publish on, determines cognitive proximity of scientists. Likewise, we regard cognitive proximity of cities by topic proximity (Leydesdorff and Welbers, 2011; Boschma et al., 2013). An optimal cognitive proximity, that obtains the most fruitful collaboration, is difficult to determine. Reason for this is that both high proximity and low proximity have advantages (Boschma, 2005). To understand this dilemma of optimising proximity, we focus on required tacit knowledge to share scientific knowledge. Having tacit knowledge in place means that also the key organisational practises are developed, which enables realisation of ones capacity to absorb knowledge.

Hence, necessary for knowledge production are absorptive capacity and the presence of some kind of knowledge infrastructure to facilitate diffusion (Nooteboom et al., 2007; Heimeriks and Leydesdorff, 2012). However, the degree of cognitive proximity between cities has been shown to influence the likeliness to share scientific knowledge in various fields of research (Balconi et al., 2004; Leydesdorff and Welbers, 2011). The cognitive state of scientists in a city is determined by the path a city has chosen. We regard prediction of evolution of cities, based on previous studied topics, as path-dependency.

1.2 Research Question

Understanding knowledge production is a matter of knowing *what* knowledge is studied *where*. To examine knowledge production over time coherently, which involves both geography and cognition, is an evolutionary problem (Heimeriks and Boschma, 2012). Therefore, the puzzle of evolutionary knowledge production questions how knowledge production changes over time. Models to solve this puzzle generally choose either path-dependency or place-dependency as a perspective, while various studies have been conducted that stress the need for both perspectives (Neffke et al., 2011; Valdaliso et al., 2011).

In a recent study by Heimeriks and Boschma (2012), the influence of both path-dependency and place-dependency have been shown in the field of biotechnology. We start our research similarly by testing the extent of the influence of these perspectives in the field of transportation. Additionally, to find an interdependency between both perspectives, further empirical evidence is required that measures accuracy of forecasting. Furthermore, the empirical evidence needs to be complemented with an approach that enables a proper merge of both perspectives. Therefore, both perspectives need to be supported by models that are comparable as well as compatible.

A comparison of path- and place-dependency enables us to measure how both effects relate to each other. Additionally, their compatibility enables us to join these models of path- and place-dependency when they have sufficient predictive value. In a practical sense, the objective of this research is to create a model that merges both explanations, but also takes their interdependencies into account. Therefore, evolutionary knowledge production is modelled by a co-evolution of indicators of both path- and place-dependency. This is a model that aims to predict diffusion and supports us to answer our main research question.

Research Question:
How can a model of path- and place-dependency explain evolutionary patterns of knowledge production in the field of transportation?

Path- and place-dependency are evolutionary explanations of the behaviour of respectively cities and topics; a co-evolutionary model intends to explain both. A co-evolution simultaneously merges two evolutions and processes their interdependencies. The relation between topics and cities can be visualised in a network (Balland, 2009). In such a knowledge network, a city and topic are associated when a scientist from that city has published about that particular topic. This 2-mode network forms the basis of analysis in this research, which leads us to answer our research question.

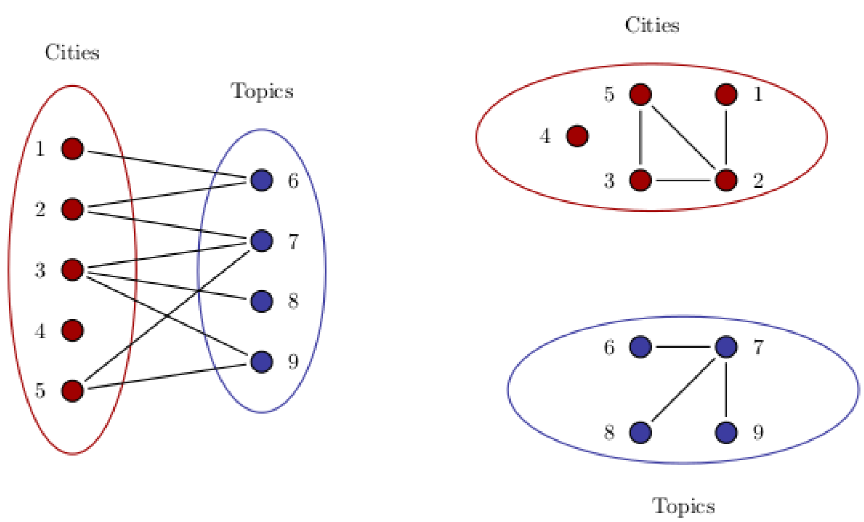

Figure 1.1: Separating a 2-mode network in a city and topic network

In the network analysis, we distinguish two separate networks; a city network and a topic network. Either cities or topics can be subject of a network; the other will function as attribute of the subject. Since our analysis will make use of the duality of these two networks, we include an example of their theoretical coherence in figure 1.1. In city 1 scientists study topic 6; in city 2 scientists study topic 6 and 7; in city 3 scientists study topics 7, 8, and 9; in city 4 no topics are studied at the moment; and in city 5 scientists study topics 7 and 9. From this 2-mode network, we can derive a city network in which two cities are adjacent when they share at least one topic. The same derivation can be performed for topics. We can vary in the threshold for the determining adjacencies at the right, but the concept remains unchanged.

Although the city and topic networks have interdependencies (Gao et al., 2011), there are two important reasons for this separation. First, separation of cities and topics increases simplification of modelling. Second, transforming the underlying bipartite graph of the network into two smaller graphs allows us to analyse cities and topics by their relation. To understand the behaviour of cities, we can study the topic network, and for topic evolution, we can study city networks.

A previous network analysis in the field of 'Transportation science and technology' that does acknowledges interdependencies between cities and topics, is provided by Wu (2013). He evaluates spatial diversity of knowledge, but refrains from forecasting by some kind of prediction model. In this research, we first design two one-dimensional models, for path- and place-dependent knowledge evolution, separately. Subsequently, we would like to merge them to provide a combined evolutionary model. Due to low predictive value of these one-dimensional models, this turns out to be not possible yet. When those models are further refined, merging them might lead to a weighted model, based on an empirically optimised ratio of path- and place-dependency. Such a combined evolutionary network will enable unification of both perspectives, but since they are merged after interpretation of the separate evolutions, it would still lack an important part of interdependency between cities and topics. This interdependency comes down to geographic and cognitive proximity of both cities and topics (Boschma, 2005; Heimeriks and Leydesdorff, 2012).

Geographic and cognitive proximity as tools for determination of an evolutionary explanation, make use of similarity in networks. McPherson et al. (2001) have stressed the relevance of homophily in social networks; they claim that similarity increases the likeliness to form ties in a network. For this process of tie formation in knowledge networks, the relevance of similarity between attributes of actors has been studied by Balland et al. (2011). Due to interdependency of cities and topics in knowledge production, this possible tendency of tie formation by homophily should be applied to both cities and topics as attributes of each other. In our final model, we examine different similarity ranges of topics to predict city adjacencies. We aimed to find an optimal proximity for topics and cities, that corresponds to the highest accuracy of prediction. The approach of optimising these proximities includes both one-dimensional evolutions simultaneously. For evolution of cities and topics this means that co-adaption of each other's dependencies gets involved in one model. The result is a co-evolution of 'path-dependency of cities' and 'place-dependency of topics'. To obtain an optimised design, we try to link optimisation of proximity to absorptive capacity of cities. First, we test proximity ranges based on a quantitative intersection of novelty value and absorptive capacity, based on an idea of Nooteboom et al. (2007). This concept has some success in terms of the prediction's precision. Second, we reverse the approach and experimentally optimise the proximity range. This latter addition enables to increase the prediction's recall as well, which determines the sensitivity of the model.

Chapter 2

Theoretical Framework

The objective of our research is to construct a prediction model that explains how knowledge production evolves. We develop models based on results of multiple studies that have been performed in the past and apply it to the field of transportation. To embed our research, we will start with defining knowledge production (section 2.1), explaining the case study (section 2.1.1) and showing the value of a network analysis (section 2.1.2). This will give a brief insight in the efforts to understand knowledge production, made in transport literature. After this introduction, we will provide a more general approach in network analyses that involve proximity theories (section 2.2). By researching different dimensions of proximity, we are able to reveal evolutionary patterns of knowledge production. For our prediction models, we will focus on two dimensions in evolutionary networks; those are path-dependency (section 2.2.1) and place-dependency (section 2.2.1). These dependencies can be linked, which is explained as a concept of co-evolution (section 2.2.3). The additional innovation theory about absorptive capacity (section 2.3) will provide more details about how to integrate the concept of co-evolution in a prediction model. Finally, we will describe how those steps form a coherent analysis, supported by a conceptual model (section 2.4). With this model, we show how we test our integrated designs of prediction models on a data set from the field of transportation.

2.1 Knowledge Production

In an innovation system, as in the innovation system of transportation, knowledge must be generated and transferred. We will omit the part of knowledge transfer that aims to translate scientific knowledge to a sector, and rather focus on the dynamics of generating and diffusion between scientists. Generating scientific knowledge is a process within a communication system carried out by scientists (Beal et al., 1986), and is regarded as knowledge production (Carter, 2008; Ernst and Kim, 2002). To evaluate knowledge, an analysis of its developments is required. One approach that can reveal a large part of the production, is focusing on knowledge diffusion, rather than searching for the origin of new knowledge (Amin and Cohendent, 2000; Gray and Schubert, 2012; Slavtchev, 2013). In innovation literature, the concept diffusion often includes knowledge transfer as well. However, in a communication system that studies scientific knowledge, knowledge transfer is left out. Knowledge diffusion is interpreted as travel of topics across the world (Shapin, 1998; Leydesdorff and Rafols, 2011), which is generally a result of knowledge sharing by scientists (Boschma et al., 2013; Wu, 2013).

2.1.1 Bibliometric Case Study

The analysis of our research starts with a bibliometric case study. It is a case study, as we focus on the developments in one particular scientific field. We only regard publication records of this field, which makes it a bibliometric study. The purpose of analysing these data is to design models that predict diffusion of topics across cities. These topics are derived from title words of publications. Consequently, we evaluate those models by their predictive strength. For the design of prediction models we make use of innovation theories about proximity and absorptive capacity.

Within the field of transportation, attempts to describe need for knowledge production focus often on the demand for transport policy (Tuominen et al., 2008). Additionally, policy makers adapt societal developments, as technological developments are judged by increase of mobility, economic growth and effect on the environment (Pettersson, 2013). Therefore, technology, society and policy gets assembled, which describes an encompassing innovation system. Tuominen and Ahlqvist (2010) stress that relationships between evolving markets, products and technologies can be explored and communicated in various ways. Structuring these options is what they regard as 'roadmapping'. Roadmaps can be categorised in *'science and technology'*, *'industry technology'*, *'corporate or product-technology'*, and *'product/portfolio management'* (Kostoff and Schaller, 2001). Tuominen and Ahlqvist (2010) explore roadmaps by means of a case study, designing descriptions of future developments that are plausible. This qualitative approach highlights the variety of roles that actors fulfil. With quantitative support, based on those different roles of actors in the roadmaps, a more detailed forecast can be created. Based on these findings, a detailed forecast of knowledge production requires a quantitative research to a *'science and technology'* roadmap.

Consistent with the results from roadmaps in transportation, Musiolik et al. (2012) have stressed that in an innovation system the role of the various actors is quite important to grasp developments. Actors that are concerned with knowledge production are scientists. Therefore, understanding of the transportation system can be refined by elaboration on the role of scientists. Wu (2013) has made a citation rank of cities and countries based on spatial diversity of scientific articles in the field of transportation. He attempts to unravel "spatial properties of citing distances, citation patterns and spatial diversity", in order to understand geographical knowledge diffusion. However, citation networks give no clear view on knowledge itself, making it harder to measure transfer of tacit knowledge. A co-authorship network performs better in capturing what knowledge is shared and whether tacit knowledge is present (Wu, 2013). Other types of networks place scientists in a central position and show other patterns, like *who* are key members in sharing knowledge (Lei and Xin, 2011). For a geographic understanding, an examination of cities or countries is a suitable choice. However, clustering scientists to cities or countries excludes cognitive elements, which should therefore be involved in another way.

2.1.2 Network Analysis

With a network analysis we research the evolutionary patterns in of knowledge production. We choose to use networks of cities and topics and analyse them with the basic concepts of geographic and cognitive proximity, which will be illustrated in the next section. A network of topics and cities focuses on the relation between knowledge and its producers, which juxtaposes scientific knowledge with its geographic origin. This has advantages over network analyses of co-authorship, which suffer from differences between countries when it comes to the habits of collaboration and minimal required effort for co-authorship (Glänzel, 2001; Glanzel and Schubert, 2001; Hoekman et al., 2009).

To create a network of cities and topics, we depend on the bibliometric data, obtained by categorising documents at city level and codification of knowledge. The latter transformation, codification of knowledge into topics, has disadvantages. It is important to note that knowledge is more than just information about a topic, it requires skills to understand a topic as well (Amin and Cohendent, 2000). Commercialising of topics, but also proper study of topics at universities, require tacit knowledge (Zucker et al., 2002). Although codification traces just parts of a city's entire knowledge base, scientific topics seem proper indicators for knowledge and for identifying relations between topics (Boschma et al., 2013). A minimum of tacit knowledge in a city about topics is, naturally, resembled by the ability of a scientist to publish his document in a journal. Furthermore, journals, which show a certain quality of work, are regarded as a proper indicator for the production of scientific knowledge (Leydesdorff and Cozzens, 1993).

In conclusion, codified knowledge is derived from a substantial set of article records, brought together in a network. Based on codified knowledge we can describe topic portfolios of cities. Finally, changes in these portfolios represent knowledge diffusion of cities.

2.2 Proximity

Proximity is relevant in the way it has impact on interactive learning and innovative performance (Boschma, 2005; Nooteboom et al., 2007). Interactive learning requires scientists to adopt topics from each other, while innovative performance highlights the need for some degree of novelty. Proximity enables to distinguish the behaviour of scientists concerning knowledge diffusion. It is therefore a suitable measure to compare scientists. There are five types of proximity distinguished in literature for collaboration and knowledge diffusion; cognitive, geographic, social, institutional and organisational proximity (Boschma, 2005). These five dimensions of proximity form a solid basis to find patterns in a network of cities and topics that can explain knowledge production (Slavtchev, 2013). Cities can be studied in such knowledge networks at document level and by address. Therefore, two dimensions of proximity imply a dependency on the evolution of these networks; those are cognitive and geographic proximity.

Cognitive proximity is the first dimension we study. To reduce uncertainty in knowledge production, people attempt to deal with the cognitive constraints that make it impossible to act optimally. To overcome that risk, they will try to rely on routinised behaviour for their learning objective. Further improvement in knowledge production will therefore be created within close proximity of the existing knowledge base (Boschma, 2005; Nooteboom et al., 2007; Balland, 2009). This phenomenon of learning from people that already share the same knowledge base is what we will consider as cognitive proximity. Cognitive proximity is a measure suitable for topics, but also for topic portfolios, as an extension of comparing cities.

Geographic proximity is the second dimension we study. Its interpretation is straightforward, as it takes into consideration the spatial or physical distance between cities. The belief that a small distance increases the probability of learning from each other, is what is considered as geographic proximity. The distance can be regarded as an absolute or relative measure, which provides several possibilities to describe the similarity of cities (Katz, 1994; Boschma, 2005; Broström, 2010). Scientists' choice of topics influences the increase or decrease of similarity of cities; thus geography fulfils an indirect role in the developments in topic choices of scientists in a city. We use geographic proximity as indicator to describe city similarities and we use aggregation at city level to exclude the effect of geographic proximity to some extent.

The notion of topics being closely related to each other is often regarded as relatedness by cognitive proximity (Boschma, 2005; Nooteboom et al., 2007). The cognitive proximity can be complemented with geographic proximity. That is a logical approach for path-dependent analyses, which start with an evaluation of topics. The reversed strategy, starting with a geographic analysis of cities, followed by adding a cognitive analysis of their topics, is a place-dependent approach. An overview of our research toward these dependencies will be provided first, after which we will explore combinations of these two patterns.

2.2.1 Path-dependency

The changing composition of a topic portfolio is what we regard as the path of a city. A city's path is therefore determined by the topics that are studied over time by scientists in that city. When we examine evolution of knowledge production, we study topic portfolios of all cities individually. Predicting or explaining these paths by searching for related topics is, in a nutshell, the implication of path-dependency; "it provides a kind of map of possible new scientific topics that may occur from the current set of topics" (Boschma et al., 2013, p. 6).

Previous research on path-dependency in innovation literature has been performed by analysis of citation transitions, where the path is defined by the references in articles that are produced by a city (Lucio-Arias and Leydesdorff, 2008). Citation analysis can be very useful to describe relations and to bring coherence in a network of scientists (Shinn, 2002; Leydesdorff and Vaughan, 2006). However, there are also drawbacks to the use of citations for analysis, due to large databases, ambiguity in content, and fraudulent use of citations (Martin and Irvine, 1983; Amin and Mabe, 2003). Uncertainty of the relatedness between documents can complicate forecasting of knowledge production.

Prediction of developments in a scientific field are often based on path-dependency. Attempts for predictions are made by gathering previously studied topics in a city. Those are linked to new topics that have co-occurred frequently in scientific documents of other scientists (Van den Besselaar and Heimeriks, 2006; Leydesdorff and Vaughan, 2006; Leydesdorff and Welbers, 2011). The result is a forecast of new combinations of scientists and topics (Ronde, 2003; Yuan et al., 2010). An important difficulty seems to be how to examine the whole history of a city. However, it is also possible to analyse the present state of a city, with regard to its portfolio, instead of involving a city's full cognitive history (Lucio-Arias and Leydesdorff, 2008; Boschma et al., 2013). Models have been designed by that approach before and make use of the Markov property. A concept that refers to the assumption that the next development of a city, formally the conditional probability distribution, does *only* depend on a city's present state (Snijders et al., 2010; Balland et al., 2011).

The notion of dependency should not be interpret too strictly. Full path-dependency implies that all new knowledge can be deduced from the present state of knowledge (Kauffman, 1995). In the previous examples and the analysis performed during the research, a weakened version of path-dependency is used. This means that a large cognitive proximity increases the probability of relatedness (Boschma et al., 2013). When we assume that the chosen path has effect on future knowledge production for all cities, then evolution is to some extent path-dependent (Van den Besselaar and Heimeriks, 2006; Heimeriks and Boschma, 2012). Hence, we are interested whether path-dependency can be interpreted as a evolutionary pattern in knowledge production.

Presence of path-dependency in scientific knowledge production has been suspected in various studies (Lucio-Arias and Leydesdorff, 2008; Neffke et al., 2011; Boschma et al., 2013). It has been observed that the "collective [of technology] evolves by a process of self-creation" (Arthur, 2007, p. 167). Self-creation implies that the process deviates from its current state of knowledge. An explanation of that origin can be found in theories of cognitive proximity, as co-occurrence of words indicates relatedness (Boschma et al., 2013). By clustering words new hypotheses in the same field could be generated. Consequently, these can be stepping stones for further research in science and technology, followed by new publications (Stegmann and Grohmann, 2003; Arthur and Polak, 2006). Occurrence of such deviations in scientists' use of topics are evidence of path-dependency.

Therefore, our first attempt in the research is to explain evolution of cities and topics as path-dependency, in order to obtain an indication to what extent such an explanation is valid. A reason for restricted effect of path-dependency is that it lacks involvement of communities of practices that are essential to the role of learning (Amin and Cohendent, 2000; Larsen, 2008). However, it is still relevant to know what part of evolution is path-dependent, and how crucial that part is. For a better understanding of learning, it is important to know how to adopt topics that are not studied in a city yet; which requires a geographic approach. For that purpose, interdependency with place-dependency of topics has to be studied (Neffke et al., 2011; Valdaliso et al., 2011).

2.2.2 Place-dependency

While path-dependency builds upon topic portfolios of cities, place-dependency reverses the approach. It focuses on topics and is concerned with the question *where* these are studied over time. In other words, for each topic there is an amount of cities where it is studied. The change in compositions of city portfolios is often based on local knowledge spillovers (Almeida and Kogut, 1999; Zitt et al., 2003; Boschma, 2004; Furman and MacGarvie, 2007). We will use this phenomenon, a place-dependent evolution of knowledge production, as definition of place-dependency.

First, the expansion of the knowledge base of a city depends on existing knowledge. This does not necessarily requires to be knowledge from that city. Local spillovers are mentioned as a cause of change in portfolio for a good reason. Diffusion of knowledge requires transfer of tacit knowledge. Intensive communication simplifies this process, which is easier with a small geographic proximity.

Second, we will focus on place-dependency as a complement to the already explained occurrence of path-dependency. In industries, for example, an increase in absorptive capacity comes from a combination of these two dependencies (Valdaliso et al., 2011). We will elaborate the role of tacit knowledge and absorptive capacity in section 2.3.

When we reflect on knowledge production by path-dependency, developments follow from topics that have a large cognitive proximity with other topics that have been studied before in the same city. For place-dependency, a similar effect is detected by means of geographic proximity (Katz, 1994; Slavtchev, 2013). Furthermore, the adoption of new topics in a city, due to the expansion of the capacity, is implicitly a deviation from a city's path (Lazaric et al., 2008); and thus cannot be deduced from that same path. It is therefore a place-dependent factor that influences this phenomenon. Place-dependency of topics creates new clusters that will not be discovered by the path of individual cities (Ernst and Kim, 2002; Boschma and Fornahl, 2011). This deviation illustrates how place-dependency can complement path-dependency.

From this illustration we can derive that knowledge evolution can be partly explained by place-dependency. However, as Neffke et al. (2011) argue, each region seems to have a profile that provides coherence and hence influences evolutionary behaviour of that region. Therefore, just like the effect of path-dependency, the extent of effect should be examined for place-dependency as well.

2.2.3 Co-evolution

Path- and place-dependency are two evolutionary patterns that have been observed in knowledge production (sections 2.2.1, 2.2.2). As we already pointed out, there is interdependence between these patterns due to overlap in their corresponding cognitive and geographic perspectives (Ernst and Kim, 2002; Neffke et al., 2011; Boschma and Fornahl, 2011; Valdaliso et al., 2011). Both perspectives have to be brought together in order to comprehend the full process of knowledge production. Therefore, we aim to provide an insight of innovation studies that have tried to give a coherent overview of co-evolutionary processes in knowledge production. It has been argued by Heimeriks and Leydesdorff (2012) that in a broad sense, scientific fields can be studied in three dimensions; knowledge, geography and socio-economy. In which case the co-evolutionary nature of research, science, and societal dynamics gives rise to the emergence of search regimes in a scientific field. As we focus on knowledge diffusion, we are interested to find out more about knowledge infrastructure, which is designed by joining dimensions of knowledge and geography. In this approach of research, dynamics are a subject to co-evolution. Another logical approach is to search for patterns that enable a co-evolution (Xing and Stroulia, 2006). Path- and place-dependency can be such patterns and are therefore explored in this research.

Some concrete findings about co-evolution that involve path- and place-dependency are found in a recent study of Heimeriks and Boschma (2012). Path- and place-dependency were detected as separate processes that influence the scientific knowledge production. Empirical evidence in the field of transportation is provided in a study by Wu (2013), who evaluated spatial diversity of knowledge and the effect on its geographic diffusion, based on citation ranks. Another example of bringing various dimensions of proximity together, is provided by Torre and Rallet (2005). They compared geographic and organisational proximities and explain how these dimensions complement each other. They argue that geographic proximity facilitates practical advantages for collaboration, while organisational proximity is a powerful mechanism for long-distance collaboration. Although cognitive and geographic proximity complement each other in a different way, there are similarities as well. In a knowledge network, topics and cities interact. This has as a result that path- and place-dependent evolutions influence each other (Heimeriks and Boschma, 2012). In our prediction models of future developments in knowledge production we had to take that conclusion into account. How we design such models, will be explained in section 2.4. Due to co-evolution of path- and place-dependency, part of the design involves refinement of the integration of proximities. Central in that refinement is the ability of scientists to acquire knowledge. We explore this perspective by delving deeper into innovation studies of absorptive capacity.

2.3 Absorptive Capacity

The concept of absorptive capacity is introduced as a characteristic of a company, describing how "to recognize the value of the knowledge, assimilate, and exploit it" (Cohen and Levinthal, 1990, p. 140). From this perspective, it was concluded that "incentives for investing in absorptive capacity are themselves driven by three industry-wide effects: demand, appropriability, and technological opportunity" (qtd. in Lane and Lubatkin, 1998, p. 463). For our research, we are interested merely in the value of knowledge, which covers the *know-what* of the learning abilities of a city. Since we research only one dimension of this concept by restricting ourselves to scientific knowledge, we will use a more modest concept of absorptive capacity as well. We will define it as a property of a city, where the city represents the scientist that publish from that city.

Absorptive capacity of a city is the "capacity of a [city] to absorb, diffuse and exploit extra-[city] knowledge" (qtd. in Lazaric et al., 2008, p. 829). In the context of scientific knowledge production it resembles the ability of scientists to transform acquired knowledge to new knowledge. To combine topics and cities, we need to study path- and place-dependency as a co-evolution. Studying them simultaneously creates an opportunity to find optimal proximities of cities, based on similarity in topics, for prediction of knowledge production (Ernst and Kim, 2002). To understand optimisation of absorptive capacity of a city, we need to study the difference between potential and realised absorptive capacity of a city (Zahra and George, 2002; Lazaric et al., 2008). Each city pursuits to realise the most potential absorptive capacity. This strategy is an expansion of a co-evolutionary attempt to increase absorptive capacity (Valdaliso et al., 2011). By changing organisational circumstances, one can influence various dimensions of proximity; and therefore realise more potential absorptive capacity. Those characteristics of cities and topics might vary per region as well as over time. In a forecast of an evolutionary network, one needs to optimise these proximities and control them for deviations over time.

For theoretical application, absorptive capacity plays a crucial role in knowledge diffusion (Giuliani and Bell, 2005). Though, it should be compared to the novelty value of the topics involved. Nooteboom et al. (2007) provided an approach for firms to optimise the balance between absorptive capacity and novelty value within an industry. In science this should be balanced as well (Heimeriks and Boschma, 2012). The extent to which those two factors apply, vary over the proximity of topics and cities. There are five dimensions pointed out by Boschma (2005) that could be useful for this matter. For the network of topics and cities it is mostly important to distinguish spatial and non-spatial factors (Agarwal, 2004; Mattes, 2012). These are also described by Valdaliso et al. (2011) as intra-cluster and extra-cluster knowledge systems. For an optimisation of this balance Nooteboom et al. (2007) argued to use cognitive proximity. They claim that the intersection of absorptive capacity and novelty value determines the optimal learning opportunities. In their work, they show by means of a graph Nooteboom et al. (2007, fig. 1, p. 1018) what the shape of the associated learning curve would be. In figure 2.1 we quantify this notion. Results of the application of this concept can be found in the evaluation of the learning curve (section 4.4.1) and when we provide the results of experimental optimisation with different cognitive proximities (section 4.5).

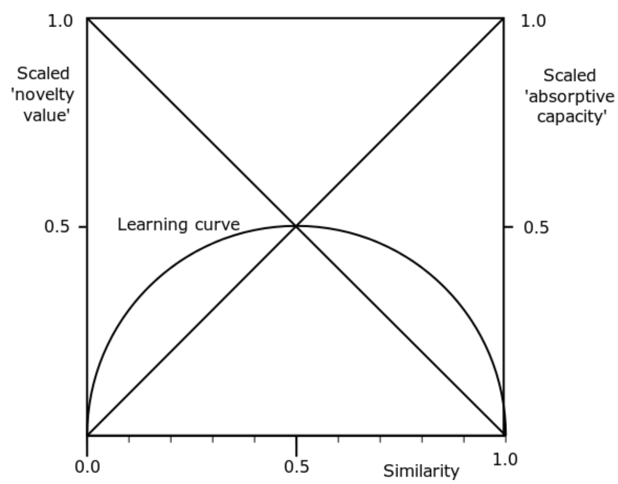

Figure 2.1: Learning curve optimisation

To apply the use of absorptive capacity in society, we need a better understanding of learning as a part of knowledge production. Therefore, Lane and Lubatkin (1998) introduced a concept of relative absorptive capacity, in addition to absolute absorptive capacity. This concept makes it easier to distinguish one-way and two-way learning. For one-way learning there is, for every chunk of knowledge, a teacher and a student. A high relative absorptive capacity shows that a city's published knowledge is easily understandable. The lack of such a difference on the other hand, makes the presence of two-way learning more likely. For operationalisation of this research, this concept has no added value, but in conjuncture with more specific knowledge of the transport sector it reveals the structure of knowledge production. Furthermore, within this sector, the role of institutions of open science is stressed to be important. Cockburn and Henderson (1998) argue that a close partnership with them is necessary for learning of the key organisational practises. However, embedding all knowledge production within firms is, just like policies that are weakening institutions of open science, counterproductive on the long run. In conclusion, the realisation of absorptive capacity depends heavily on the connectedness with scientists that produce the necessary knowledge.

2.4 Conceptual Model

To operationalise our research, we first design individual models of path-dependency and place-dependency, as these two concepts are the variables of interest in the research. Both dependencies have the ability to forecast knowledge diffusion. The result of that forecast determines whether evolutionary patterns of knowledge production in transportation can be explained by a co-evolutionary model. Hence, it determines whether a combination of path- and place-dependency can actually be regarded as a co-evolution.

Similar models have been designed for topic predictions in a network of knowledge production. In another scientific field, Boschma et al. (2013) carried out a study to forecast the coming and going of specific topics. However, grasping diffusion requires an analysis of already studied topics that do not disappear. It excludes those topics who enter or leave the scientific field.

In the field of 'Transportation of Science and Technology' a previous study has been conducted that showed empirical evidence of the effect of spatial diversity on geographic diffusion (Wu, 2013). In that same scientific field we now search for empirical evidence of interdependency between evolutions of cities and topics. We use path- and place-dependency as the basis for a conceptual model (2.2). In the upper half of the model it is shown that knowledge production in transportation performs as a case study for our innovation study. It provides data for, and therefore determines, the innovation processes of co-evolution of path- and place-dependency *(a)*. On an individual basis, path-dependency describes the evolution of cities *(b)*, while place-dependency describes the evolution of topics *(c)*. The objective of modelling co-evolutionary processes is to predict diffusion by means of these two dependencies, which is visualised by an empirical study in the lower half of the conceptual model. How these processes are carried out, has already been elaborated when we clarified the research question (section 1.2). At this stage, we make the approach more concrete, as we distinguish five steps. These steps will be guided by sub questions, each resulting in a potential prediction model of knowledge diffusion.

1. Can a model of path-dependency explain evolutionary patterns knowledge production in the field of transportation?
2. Can a model of place-dependency explain evolutionary patterns of knowledge production in the field of transportation?
3. Can an optimised ratio of path- and place-dependency provide a combined model that explains evolutionary patterns of knowledge production in the field of transportation?
4. Can an optimal choice for absorptive capacity determine a co-evolutionary model that explains evolutionary patterns of knowledge production in the field of transportation?
5. Can an experimentally optimised proximity improve other models to explain evolutionary patterns of knowledge production in the field of transportation?

These five questions form the empirical study that complements the theoretical concept in figure 2.2.

The first two question investigate whether path- and place-dependency are proper evolutionary explanations on their own. This is partly a reproduction of earlier innovation studies. It has, however, two additional components; it is applied to a particular field and both explanations are modelled in a similar way. Subsequently, the concepts have to be merged in a suitable way, dealing with the overlap and the influence of both explanations. This is covered in as well the third as the fourth sub question. The first objective after combining path- and place-dependency is to find an optimal ratio between the two concepts; see *(d)* and *(e)*. This is only possible if both one-dimensional models predict well. Second, we investigate the effect of optimising proximities by absorptive capacity on the design of a prediction model *(f)*, questioning the prediction strength of this refinement by theory about proximities. The last sub question is a second attempt to optimise proximity without direct use of absorptive capacity, as this concept is difficult to measure in an exact way and link it to proximity. Therefore, we perform an experimental analysis, to omit those problems. Finally, we can use the prediction results as feedback mechanism to reflect on the developments in transportation *(g)*.

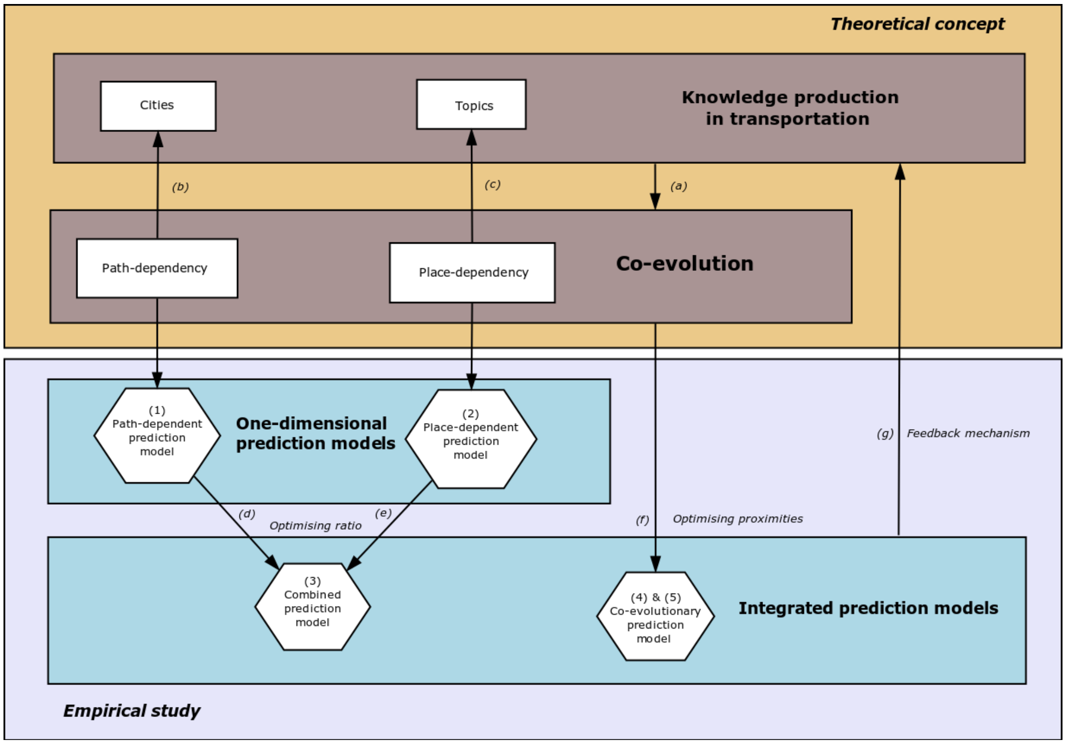

Figure 2.2: Conceptual model

In the operationalisation we present a more detailed overview of network levels, that will provide a better understanding of the duality of topic and city networks. We will distinguish two types of networks. A network in which topics are placed centrally and cities imply relations; and a network in which these roles are reversed. This duality complicates the relation between path- and place-dependency, but is simultaneously the essence of their co-evolution. How we will delve into the interdependency of cognitive and geographic proximity, which are the underlying perspectives of this co-evolution, will be explained in more detail in the methodology (chapter 3).

Chapter 3

Methodology

This chapter elaborates what data we use and how we process them. First, we explain how we gather our data and what tools we use. Then, we will elaborate knowledge networks, which are the basis of the method of analysis we apply on our data set. Finally, we operationalise the approach of analysis by specifying dependent and independent variables and we explain how to measure these variables.

3.1 Data and Tools

The data that we analyse for this research are publication records. These records include a variety of information, of which we use the titles of the documents and the addresses of their authors. Titles induce a set of topics that form codified knowledge of the 'Science and Technology of Transportation' field. Addresses of the authors provide a framework for the analysis, as the topics can be regarded as characteristics or attributes of the cities. For each city, a portfolio can be composed of topics that are studied by authors in that city. We use the topic portfolios to compare the cities and to evaluate the evolution of knowledge.

The source of data is the ISI Thomson Reuters Web of Science. Within this source we choose the category 'Transportation Science and Technology' as field of study. From the records in this category, the journal 'Transportation Research Record' published most documents. This journal is therefore regarded as the core journal of the field, and the analysis assumes that it fulfils a proper indication of the knowledge production. Since we are interested in the dynamics of knowledge production rather than the effect of its content, we choose a journal with a quantitative majority share over other journals, that could be selected based on different indicators like citation counts and impact factors. The total number of records from 'Transportation Research Record' is 9451; these records form the entire data set of this research. For a proper description of its evolution, the time span must be sufficiently large. We consider the period from 1998 until 2011 for this analysis, as the core journal publishes articles since 1998.

After the records of these articles were downloaded from the Web of Science, we pre-process the data to extract locations and topics of the articles more easily. We perform this extraction by the executable file isi.exe, which is freely available for academic usage (Leydesdorff, 1989). One of the documents provided by the extraction, is a list of all addresses of authors that have published in the core journal during the period from 1998 until 2011. In figure 3.1 a map of cities is shown, where the scientists were working when they published. This map represents the geographic origin of knowledge production (R Development Core Team, 2008; Becker and Wilks, 2012; South, 2012).

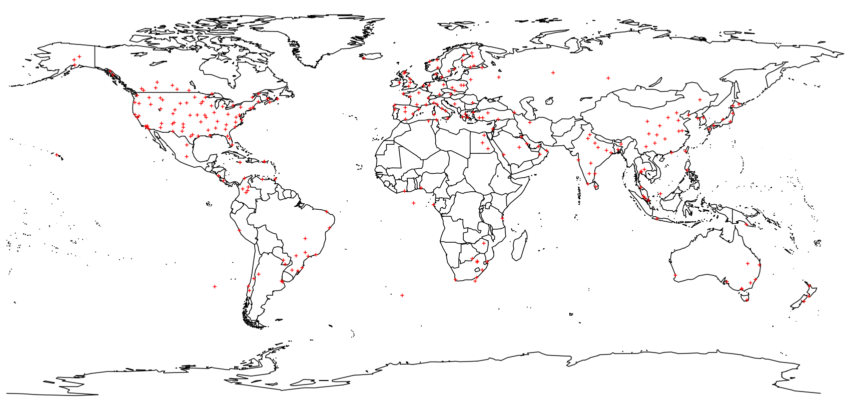

Figure 3.1: Cities (266) of knowledge production after aggregation

Subsequently, we aggregate the addresses on city level. This enables us to carry out a geographic analysis, which is not evidently clear when we would map individual scientists. With portfolios of the cities, we construct a network with ties between cities that are similar in terms of co-occurrence of topics in their portfolio.

Furthermore, we use of Yahoo! Geocoder (Yahoo Geocoder, 2013), after we customised data from 'ISI Thomson Reuters Web of Science' with specific ISI software of Leydesdorff (1989). This goecoder is a visualisation tool that runs in a web browser and transforms addresses into GPS coordinates. Further analyses we perform by the free software R (R Development Core Team, 2008). Within this statistical software environment we can use several packages to simplify analysis. For the visualisation of figure 3.1 we used the packages 'mapdata' (Becker and Wilks, 2012) and 'rworldmap' (South, 2012).

For the study of topics, we perform a semantic mapping of words that makes use of word-document matrices (Leydesdorff and Welbers, 2011). The package in R that supports this is the 'tm' package (Feinerer et al., 2008; Feinerer and Hornik, 2013) for general text mining and the 'lsa' package (Wild, 2011), which provides all types of tools of latent semantic analysis.

3.2 Knowledge Production as Network Analysis

To picture knowledge as evolutionary, we use networks to present its current state per year. As nodes for our networks we use cities. We include the neighbourhoods of cities as well, which makes that it does not really differs from analyses of regions. The key point is to aggregate at geographic level. Scientific knowledge is determined by the development of groups (Lei and Xin, 2011), and regions or cities are a natural for competition (Boschma, 2004; Heimeriks and Boschma, 2012).

In our research we perform aggregation at city level, because comparison by geographic proximity would need to use a central point in a region to perform analyses. Therefore, we prefer to use cities for analysis, in the sense that we consider them geographically as coordinates; as latitude and longitude. The cohesion of cities is based on shared knowledge, resulting in a knowledge network with cities as nodes or actors. Adjacency in knowledge network of cities depends on the amount of shared knowledge (Balland, 2009; Balland et al., 2011; Heimeriks and Boschma, 2012).

The dual representation of a city network is a topic network, where topics are related when they are sufficiently studied in the same documents or cities. This approach describes similarity of cities by the similarity of topics that they studied, which is comparable to an industry that is examined by similarity of products. Hausmann et al. (2011) have developed a so-called Product Space to link countries to their industrial activities. In the knowledge network, the level of comparison of topics determines the dynamic of the search regime of knowledge (Heimeriks and Leydesdorff, 2012). We distinguish different levels by making comparisons of topics at document level, which indicates cognitive proximity; and comparison of topics at city level, which indicates geographic proximity.

Essentially, both the city network and the topic network are transformations from a 2-mode network, containing two disjoint sets; those sets are cities and topics. As we would like to make a clear distinction between path- and place-dependency, and simultaneously reduce the complexity of analysis, we will treat the disjoint sets as separate networks. In these networks, adjacencies are present when topics or cities are two steps separated in the original 2-mode network (Leydesdorff and Vaughan, 2006). These two network transformations improve the overview and reduce the complexity and amount of calculation.

We analyse the networks, with regard to path- and place-dependency, on three different levels. These levels are concerned with endogeneity, heterogeneity, and proximity. At those levels we study respectively the network structure, actor characteristics, and actor similarity within the network (Balland et al., 2011). The following sections will explain the details of the levels of analysis with regard to a knowledge network. Either a city or topic network will be used as example to clarify the variety of interpretations.

3.2.1 Endogeneity level

Endogeneity of the city network is concerned with the internal properties. It emphasises how cities are related to each other. From the properties of a network it is possible to make some predictions for future developments of the network (Balland et al., 2011). This is generally referred to as preferential attachment

(Leydesdorff and Rafols, 2011) and is based on cognitive and geographic proximity analysis in this research. Despite the similarity in terminology, this analysis differs from the analysis at the level of proximity, as it involves only the quantitative component of the topic records. Analysis at proximity level is concerned with a qualitative study of the origin of the relation between cities. That requires an analysis of the topic network; as will be explained in section 3.2.3.

Cognitive city proximity is used directly to construct a knowledge network of cities by setting a threshold for a minimum of topics that two cities have in common. Reaching this threshold is visualised by a connection between the two cities (Heimeriks and Boschma, 2012). This construction can be compared to another cognitive approach, that compares cities at document level with each other. That approach studies a co-authorship relation that shows only interest in co-authorship between cities and leaves out co-authorship within cities. Geographic city proximity is a more straightforward measure. It examines the distance between addresses of cities (Wu, 2013). Additional to this measurement, a transformation is used to create a proximity measure instead of a distance measure. This transformation makes that all analyses are uniformly carried out and therefore provide comparable results.

Both approaches reflect on endogenous properties of the network. The cognitive approach of a customised co-authorship measure is very similar to the original construction of the network and the geographic approach is very static. Therefore, a more sophisticated complement is required to derive information from current developments. The following section, 3.2.2, illustrates how to search for optimal city proximity to improve information retrieval for prediction models.

3.2.2 Heterogeneity level

The endogenous analysis (section 3.2.1) is suitable for prediction if one assumes a correlation between proximity and increase of similarity. However, too much proximity might have a negative effect on collaboration (Nooteboom et al., 2007) and therefore also on the probability of topic adoption that increases similarity between cities.

For the method design in this section, we take into consideration that an optimisation is necessary before linking cognitive or geographic proximity to the probability of increased similarity. Heterogeneity of actors in the knowledge network of cities will be involved for that purpose. Based on the position in the network of individual actors, cities or topics, characteristics can be assigned to them (Boschma and Fornahl, 2011; Leydesdorff, 2004a). Also connectedness of clusters in the network can be studied by an evaluation of articulation points (Leydesdorff, 2004b). These kinds of network properties have an influence on the probability of a city to adopt certain topics. Therefore, they have the ability to refine proximity based prediction models for knowledge production.

While the previous example emphasises city properties, presence of particular sets of topics can make a difference as well. For that reason, we return to the dual network, where topics play a central role. Based on characteristics of a topic, its novelty value can be determined. The degree of novelty of a topic varies per city. Therefore, characteristics of cities and topics have to be combined. It is the intersection of absorptive capacity and degree of novelty that indicates the right cognitive distance (Nooteboom et al., 2007). This is a second example of how to design prediction models for knowledge production.

3.2.3 Proximity level

In section 2.2, the meaning of proximity is already explained. It is also demonstrated that we use two dimensions of proximity for analyses in this research; cognitive and geographic proximity. Proximity, however, also indicates a level of analysis in network studies.

Optimisation of proximity can be a useful tool, as shown in section 3.2.2. However, it requires the ability of cities to realise their potential absorptive capacity (Lazaric et al., 2008). As that is not always possible, or not possible to estimate, a plainer approach is required. We will construct an empirical simulation of the evolution of topics, and use that as a starting point for a plainer approach. To make a comparison of cities possible, each city will be examined by its topic portfolio. Proximity is a measure for dyadic variables; it evaluates the tie formation in a knowledge network of cities (Balland et al., 2011). Ties in a city network are actors in its dual network representation, which is a topic network. In section 3.2.1, an endogenous analysis was introduced for the city network. A very similar analysis can be performed on the topic network. The results can be translated back to the city network as an analysis of dyadic variables, due to duality of the two networks. This concatenation of transformations is the proximity analysis that determines whether tie formation is likely. This distinctive nature of analysis highlights the significant expansion with respect to innovation theory of proximity. Proximity, as introduced earlier on (in section 2.2), is restricted in its use when it comes to forecasting, as it is a continuous similarity measure.

These possibilities for proximity analysis enable two angles for comparison. Topic similarity can be operationalised at document level, when cognitive proximity is the measure of similarity; or by aggregation at city level, when geographic proximity is the measure of similarity (Heimeriks and Leydesdorff, 2012; Heimeriks and Boschma, 2012). For our research, we describe topic adoption by cities at proximity level and design a method that calculates probable future topic adoption. Through application of this method, to all cities, we gain predictions of increased similarity of cities in terms of their topic portfolios.

3.3 Operationalisation of Variables

Following the steps of the conceptual model, the research question is subdivided into five questions (section 2.4). These questions aim to distinguish the roles of path- and place-dependency. More specific, they attempt to find out whether they can be combined or that they co-evolve and therefore must be studied simultaneously. So, after studying both patterns separately, the first objective is to find an optimal ratio between the two concepts. Then, we will investigate the effect of optimising proximities before applying path- and place-dependency on knowledge evolution, questioning the accuracy of this refinement by theory about proximities.

The evolution of knowledge production is concerned with cities and topics. The 2-mode network of topics and cities can be separated into a city network and a topic network. We study both networks at three different levels, which are the original indicators for this research; endogeneity, proximity, and heterogeneity. The application of these indicators is adapted to the nature of underlying network (see section 3.2). For place-dependency of topics, we have study the dual network of cities, and for path-dependency of cities, we also have to study the reversed perspective, which is the topic network. The operationalisation of our research is shown in figure 3.2. Here, we also provide their indicators (A-F) and how we measure them.

Variables	**Dimensions**	**Indicators**	**Measures**
Dependent			
Knowledge production	City network (CN)	Network ties	Adjacency matrix
	Topic network (TN)	Network ties	Adjacency matrix
Independent			
Place-dependency	Endogeneity of CN	(A) Geographic city proximity	Perceived/scaled similarity
	Proximity of CN	(B) Cognitive city proximity	Cosine similarity
	Heterogeneity of CN	(C) Absorptive capacity	Descriptive
Path-dependency	Endogeneity of TN	(D) Cognitive topic proximity	Cosine similarity
	Proximity of TN	(E) Geographic topic proximity	Cosine similarity
	Heterogeneity of TN	(F) Novelty degree	Descriptive

Figure 3.2: Operationalisation

Since cities determine the main perspective of our research, we start with the city network. Because the duality we just explained, we thus operationalise place-dependency first. From this operationalisation table, most of the dimensions, indicators and measures are explained in the theoretical framework (chapter 2) or in the methodology on knowledge networks (section 3.2). The measurement of indicators for independent variables is the only part that is left to be elucidated. The tenor of indicators of our dependent variable is quite straightforward. A more detailed description of their measures is provided in section 2.2. Accuracy of prediction of these indicators is the objective of our research and has been studied by emphasising the independent variables. Therefore, we will provide a more thorough annotation of the measures of indicators for these independent variables.

For the geographic analysis of cities, part of the place-dependency perspective, a binary matrix with cities and addresses will coincide at entries close to each other. There is, by definition, no overlap in geographic position of cities, which results in zero similarity between cities for any similarity measure. Because that does not provide any information, we use an intuitive dissimilarity measure, which is the Euclidean distance between addresses of cities. For comparison with the similarity measures, it is convenient to transform this dissimilarity measure to a similarity measure as well. We test two derivative measures for this purpose. First, we use the perceived similarity measure. This measure is directly related to the Euclidean distance d by an exponential function: for cities A and B the perceived similarity coefficient is given by $s(A,B) = e^{-d(A,B)}$. This exponential function is used at the level of cities, although a city in this context is actually a set of actors. Those actors are inclined to act by psychological rules, as stated in the universal law of generalisation (Shepard, 1987). Second, we use a simple scaling function, for it is difficult to estimate how geographic distance relates to the measures of other indicators, which are all calculated similarly. Scaling similarity is also based on the Euclidean distance d and depends on the maximum distance max_d between two cities in a network: for cities A and B, the function $s(A,B) = 1 - \frac{d(A,B)}{max_d}$ provides a scaled similarity coefficient (Tan et al., 2006).

For the other indicators, which are geographic topic proximity and cognitive proximity of both cities and topics, we use another measurement. A binary matrix of cities or topics with scientific documents or addresses is a well-founded basis to calculate their similarities. When we compare cities or topics, it does not contribute to similarity when both data objects are not present in a document or at an address. Hence, it would be convenient to ignore entries when both data objects are zero. This modification is especially important when the matrix is very sparse, as in this analysis. The cosine similarity measure is not sensitive to zeros (Leydesdorff, 2005). Therefore, a cosine similarity measure would be the best choice for this binary data comparison.

So far, we use a combination of endogeneity and proximity analyses, but analysis at heterogeneity level is not taken into account. Heterogeneity can be measured by characteristics of an actor, which is a city or topic, depending on the network. In a city network, heterogeneity can be measured by absorptive capacity. For a topic network, the degree of novelty of a topic is an important factor, which might be a relative measure, as it depends on the involved cities. Because these nuances need further exploring first, we base the application of heterogeneity on intermediate results and the theoretical positioning from section 3.2.2. By including this perspective, we can illustrate the difference between combining path-dependency and place-dependency on the one hand, and exploring a co-evolutionary model of both perspectives on the other hand.

Now we have operationalised our research, the six indicators we have introduced and explained need to be implemented. An overview of their implementation is explicated in a flowchart in figure 3.3. At the right side of every block we start with the data and end with the necessary indicators for our prediction models. At the left side of every block we denoted the elements that are required in the process. These latter elements are the building blocks of our analysis that connects the raw data (introduced in section 3.1), with the obtained results (chapter 4). The building block contains lists and matrices that provide intermediary results, which have been described throughout this methodology in more comprehensible terminology.

The method we use to test the presence of evolutionary patterns, is based on similarity measures. With these measures we can determine alleged adjacencies between cities and topics. The approach we use to transform these similarities of indicators into a prediction model, consists of four steps. First, we choose a data set that determines the current situation, which functions as input for the prediction model. Second,

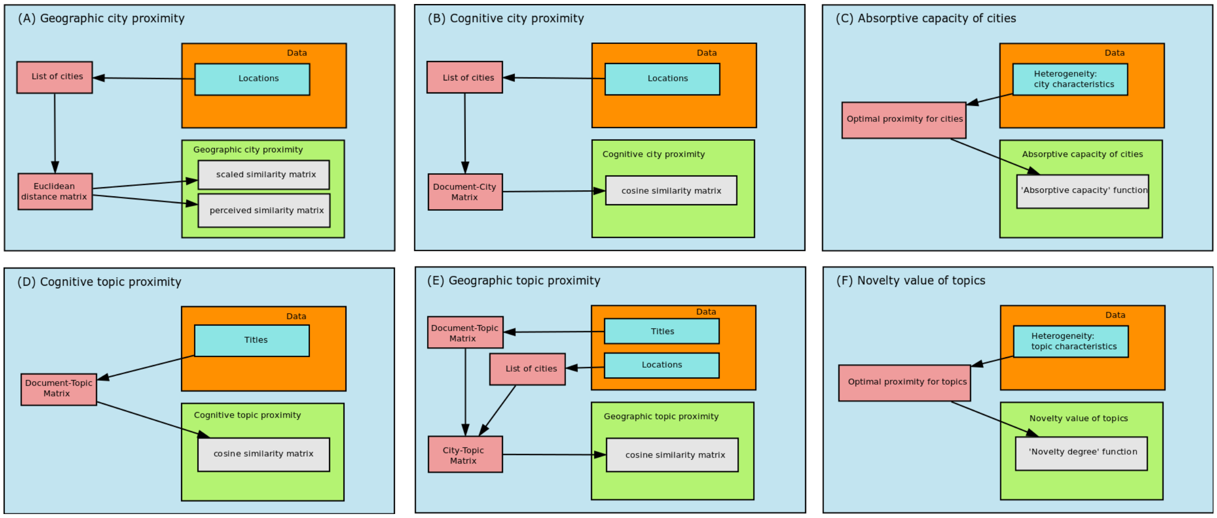

Figure 3.3: Indicator approaches

we determine the city and topic network, that represents the current situation. Third, we calculate the similarities of the four basic indicators for the chosen data set. Finally, we determine the most likely new adjacencies between pairs of cities (and pairs of topics), based on the similarity between them.

To provide a chronology in the research, we have split the data set into 14 consecutive parts. Each year marks a fragment of the total set of documents and we have constructed batches of consecutive fragments. We start with the first 7 fragments, which provide input for a prediction of the new situation in year 8. Second, we repeat this process with fragments from year 2 towards 8 for a prediction of the new situation in year 9. We carry out this process 7 times successively to generate more data to make claims about the accuracy of the method.

In chapter 4 we will provide results of this process, in which we make a distinction in prediction model designs. These different designs enable us to answer the five sub questions that are necessary to grasp the encompassing nature of the research problem.

Chapter 4

Results

In this chapter, we present the results of the data analyses of all indicators that are discussed in the methodology (chapter 3). The first section starts with an introduction that motivates statistical measures we have used. We have tested the significance level with a χ^2-test and their predictive value in terms of an F-score, that takes both precision and recall of a prediction into account. By addressing both components of the F-score individually, we address the core problem of interdependencies between path- and place-dependency, which are based on evolutionary patterns from section 2.2 and explored during this research. After this introduction, sections 4.2 and 4.3 provide direct results from measuring indicators. Since the data can be regarded from a city and topic perspective, results of their analysis will be presented accordingly. Then, section 4.4 takes the remaining indicators into account. The absorptive capacity is linked to a learning curve that that determines optimal proximity, which bring city and topic networks together. Finally, we apply an experimental optimisation of proximity, which concludes our results.

4.1 Statistical Introduction

There are several measures to render the predictive value of our models for knowledge diffusion. For the results of our indicators, we use some abbreviations. We will discuss these, and their meaning, now. In the result tables we will start with providing the number of actual adjacencies between cities in a particular year, based on co-occurrence of topics in their portfolio. We regard two cities adjacent when they have a topic in common that is among the top 500 most popular topics of that year. Taking more topics into account requires too much calculation. Second, we construct a prediction model, based on the particular indicator, which predicts adjacencies. For each year, that number is given in the second row. The geographic position of cities is static, therefore the number of predicted adjacencies here is the same every year. Later in the research, this number will vary per year. Next, we compare the predicted adjacencies to the actual adjacencies. As an example of the adjacency matrix of cities, we visualised the data of the year 2011 in figure 4.1.

When we compare the prediction with the actual adjacencies, we obtain four values; these are true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). The sum of those four values is N. The values are usually represented by a contingency table, as shown in figure 4.2.

Based on this distinction, we have calculated whether the indicator has a significant effect with a χ^2-test and given its predictive strength as a prediction model by means of an F-score.

4.1.1 χ^2-test

As a statistical method for showing the significance level of our predictions, we use the χ^2-test (Fienberg, 2011). Following the standard representation for the χ^2-test, the test statistic is given by the formula

$$\chi^2 = \sum_i \frac{(O_i - E_i)}{E_i}, \tag{4.1}$$

where our predictions are the observations O_i and actual adjacencies are represented as the expected values E_i. For this this test statistic there is just one degree of freedom and we choose a significance level of $\alpha = 0.05$, which leaves us with a critical value of 3.84 for the χ-distribution. This enables us to evaluate an indicator, by testing whether or not to reject the null-hypothesis H_0 which claims that the predicted adjacencies, based on that indicators, and the actual adjacencies are independent. Sometimes the significance level is very high, which can be recognised by a very large test statistic. When the associated p-value is smaller than $2.2e - 16$, then the R software will automatically round this value to 0.

Since we expect a correlation with high significance level, we need to evaluate the prediction model with a more biased indicator. We have used the F-score for that purpose. In the next section we will elaborate the relation with the more common χ^2-test and also motivate our choice for this biased variation of measure.

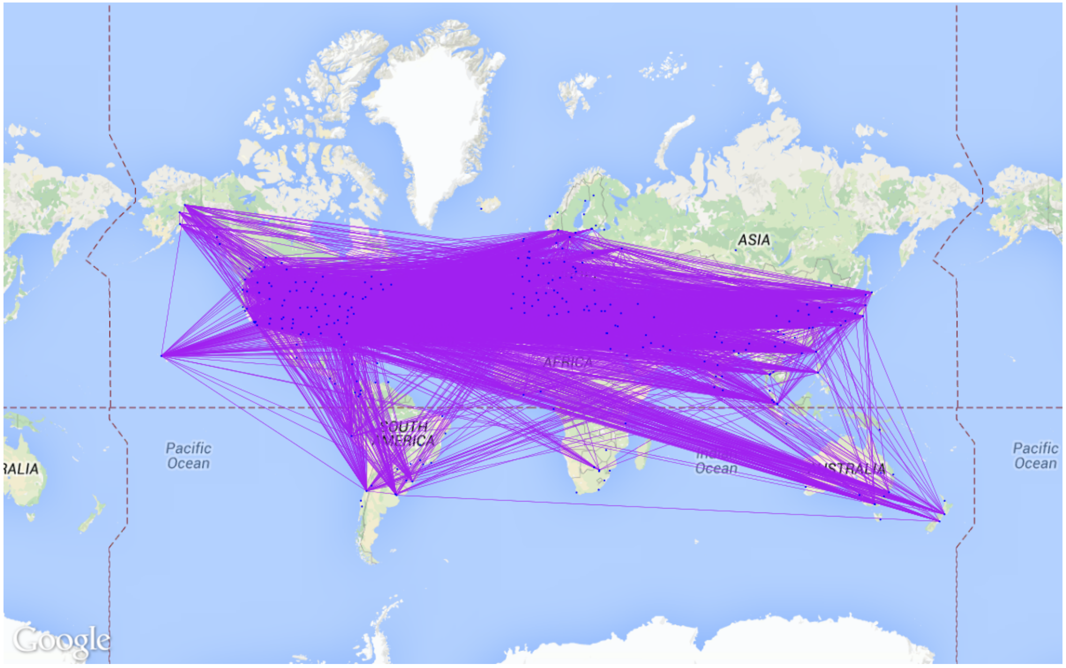

Figure 4.1: Cities (266 blue vertices) with adjacencies (purple edges) in 2011

Test outcome	Positive Condition (Adjacent)	Negative condition (Non-adjacent)
Adjacent	TP	FP
Non-adjacent	FN	TN

Figure 4.2: Contingengy table

4.1.2 F-score

To link the strength of predictions models to practical applications, we need to involve two concrete objectives. Those are the precision and recall of a prediction. Precision of a prediction indicates how much of predicted adjacencies actually occur and is given by

$$\text{precision} = \frac{TP}{TP + FP}. \tag{4.2}$$

The recall of a prediction model indicates how much of the actual adjacencies is predicted by a model and is given by

$$\text{recall} = \frac{TP}{TP + FN}. \tag{4.3}$$

Both indicators neglect the number of true negative predictions. This compensates the characteristic of the skew distribution, caused by high sparsity of the underlying data. These two indicators can be combined, each with a certain weight. This allows interpretation of predictive value by varying the importance of precision and recall. The formula for this advanced version of the F-score, is the F_β score, given by

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}. \tag{4.4}$$

For now, we will work with the harmonic mean, where precision and recall are equally important ($\beta = 1$) and denote the F_1-score simply as F-score.

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{4.5}$$

As pointed out, this score neglects true negatives. A more direct approach to do this, is to use the accuracy measure, accuracy $= \frac{TP}{TP+FP+FN+TN}$, and simply leave the true negatives out. This gives us the Jaccard index, $J = \frac{TP}{TP+FP+FN}$, which has been used in similar research as well (Boschma et al., 2013). However, this is a direct transformation of the F-score, as $J = \frac{F}{2-F}$.

To compare their interpretations, a more comprehensible denotation of the F-score would be

$$F = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}. \tag{4.6}$$

Whereas the Jaccard index will always be lower than both precision and recall, the F-score lies between them. The F-score assumes there is a natural number of true negatives, choosing it equal to the number of true positives. A non-sparse data set is simulated by this approach.

When we would like to link these measures to the χ^2-test, we need to return to the 2x2 contingency table. For the χ^2-test, we can retrieve our test statistic via the ϕ-coefficient, which we will introduce first. The ϕ-coefficient, also known as Matthews correlation coefficient, is the result of the Pearson correlation coefficient with two binary variables. It can be calculated by

$$\phi = \frac{\frac{TP}{N} - \text{recall} \cdot \text{precision}}{\sqrt{\text{precision} \cdot \text{recall} \cdot (1 - \text{recall}) \cdot (1 - \text{precision})}}. \tag{4.7}$$

Now we obtained the formula for the ϕ-coefficient, with a contingency table of c columns and r rows, we can continue with the relation between ϕ and χ, which provides us with the test statistic and its degrees of freedom (*df*) (Fawcett, 2006):

$$\chi^2 = N \cdot \phi^2, \text{ with } df = (c-1) \cdot (r-1). \tag{4.8}$$

Although the χ^2-test controls for the true negative predictions, and is therefore a necessary starting point to check the significance level, the approach of the F-score is most intuitive due to its straightforward composition of precision and recall. For that reason we test the prediction models in more detail based on precision, recall and F-score.

4.2 Place-dependency: A City Network Analysis

The innovation theory of proximity has introduced geographical and cognitive aspects of knowledge evolution. Both aspects are dimensions of proximity, that form a basis for evolutionary patterns in knowledge production. Path- and place-dependency, as well as interdependencies between them, are patterns that have been elaborated in section 2.2.

We have analysed these patterns empirically through six indicators, which are introduced in figure 3.3. Three of those indicators are related to the city network. From the city network of transportation, we can derive both a geographic and cognitive indicator. The third indicator that can be derived from it is absorptive capacity of a city. This latter indicator will be discussed later on (in section 4.4.1), as it relates strongly to the novelty degree of topics, an indicator derived from the associated topic network.

Geographic city proximity (indicator A)

The analysis of the geographic city proximity starts with a set of all addresses where is published about the scientific field of 'Transportation Science and Technology'. The number of addresses is 16856, which is a too large number for analysis due to computability restrictions. To gain a better overview we remove duplicates and cluster the addresses to smaller regions. For the addresses mostly represent universities, the names of cities will cover a clear set of regions. The addresses are therefore clustered at city level, as has been shown in figure 3.1. As the size of cities differ, and because we want to examine the geographic similarity of addresses, we regard addresses as the same when their latitude and longitude are equal after they are round to multiples of four. The number of four has been chosen experimentally, as it reduces the number of cities to a number of 266. A number of regions or cities between 200 and 300 has been found a reasonable amount in previous analysis (Heimeriks and Boschma, 2012).

We perform a geographic analysis of these 266 cities by taking pairwise distances between addresses, which determine geographic city proximity. Subsequently, we transform the resulting distance matrix of cities to a similarity matrix. This can be done by several concepts of similarity. We analysed two of them, which are perceived and scaled similarity measures. As a result of a too steep decrease of dissimilarity when geographic distance increases, the perceived similarity measure provided a very small set of distinctive city pairs to be linked together. Therefore we restrict ourselves to scaled similarity, which gives better results for proper interpretation. In the map of figure 4.3 the purple edges show the predicted adjacencies based on geographic proximity. The results for this prediction are shown in table 4.1.

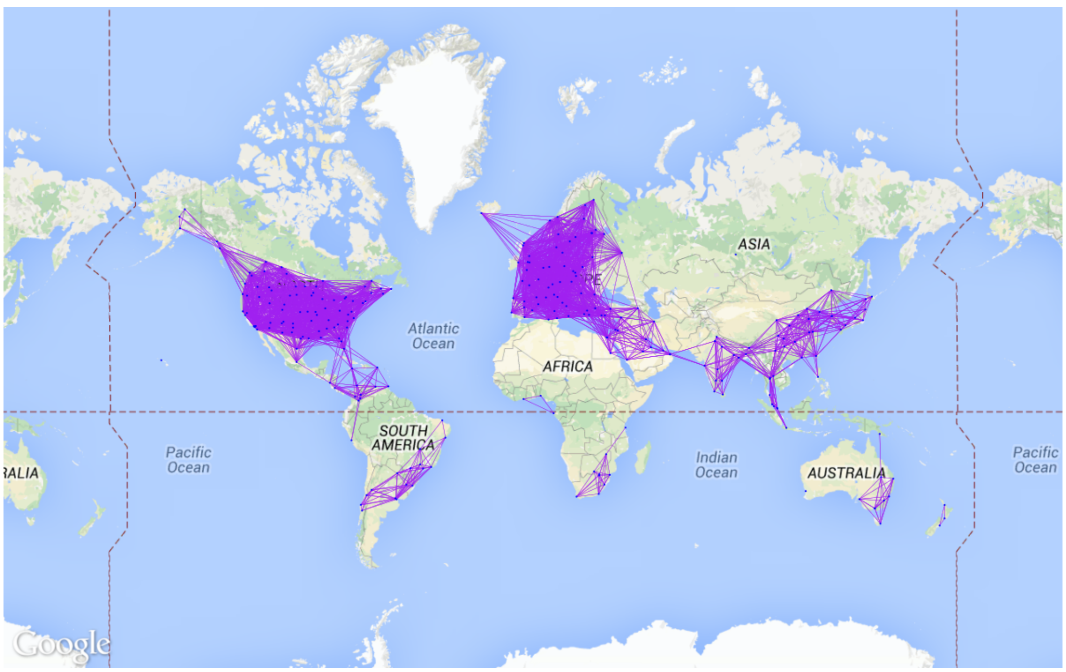

Figure 4.3: Cities (266 blue vertices) with predicted adjacencies (purple edges) based on geographic proximity

As the table shows, the p-values of geographic proximity are very small (0 after rounding). So the presence of a significant effect (for $\alpha = 0.05$) is clear. As the effect recurs each year, we can even consider it be a trend. This trend we will find again with the other indicators. It is mainly because it is a very plausible assumption to expect a high accuracy in sparse data sets, when one has a strong indication of the density of adjacencies.

This problem, of sparse data, makes the predictive use less clear in this research. This brings us to the final part of the results table, more suitable for measuring predictions in sparse data. These indicators are precision, recall and their harmonic mean, which is called the F-score.

After application of these enhanced indicators for prediction models, the results are less supporting with regard to the expectations of the research. Both precision and recall are in the range $0.15 - 0.26$. As we will see in further results, the recall remains in this area, but is acceptable when one would like to have at least some predictions for future adjacencies. From the relatively small part of predictions, with respect to the

Geographic City Proximity	2005	2006	2007	2008	2009	2010	2011
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	3706	3706	3706	3706	3706	3706	3706
TP	943	538	758	728	924	839	925
FP	2763	3168	2948	2978	2782	2867	2781
FN	4743	2172	2543	2071	3991	3229	3808
TN	26796	29367	28996	29468	27548	28310	27731
Test statistic χ^2	264.7	271.0	598.3	774.0	415.6	498.3	472.5
p-value	0	0	0	0	0	0	0
Significant (alpha=0.05) effect	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Precision	0.25	0.15	0.20	0.20	0.25	0.23	0.25
Recall	0.17	0.20	0.23	0.26	0.19	0.21	0.20
F-score	0.20	0.17	0.22	0.22	0.21	0.22	0.22

Table 4.1: Results: Indicator A (for similarity ≥ 0.9)

number of potential adjacencies between cities, a higher score can be expected. We will see a clear distinction when we proceed with our results.

So, in despite of an acceptable accuracy, a model based on this indicator lacks precision. Therefore, as a solitary indicator, geographic city proximity has no high predictive value.

Cognitive city proximity (indicator B)

In the analysis of cognitive city proximity we make an attempt to focus on relatedness between cities, based on the co-publishing of scientific documents. This is similar to a co-authorship relation analysis, except the cities are now the unit of analysis. The result of that choice is that the intra-city of co-authorships are neglected, and the focus is only on the inter-city co-authorships.

Starting with the list of cities, from figure 3.1, we construct a matrix in which for each document the cities of the authors are marked. This matrix is called a document-city matrix (see figure 4.4), and formed the base of the data for this analysis.

	city 1	city 2	city 3	...
document 1	0	0	1	...
document 2	1	1		
document 3	0		$\ddots$	
$\vdots$	$\vdots$			

Figure 4.4: Document-city matrix

The co-occurrence of documents in a city infers the cognitive city proximity. The calculation of the cosine similarity of documents makes use of these co-occurrences. Although we make use of the cosine measure, which omits sensitivity for zeros in the matrix, this method is unable to provide any useful results. The differences in similarity are too small to make a clear distinction between various city combinations.

Hence, cognitive city proximity provides no explanation for evolution in knowledge production in the scientific field of transportation. This indicator will therefore neither be used for further predictive purposes.

4.3 Path-dependency: A Topic Network Analysis

Based on the topic network we will design two additional indicators. First, we study topics by their relatedness due to co-occurrence in the documents of our data set. Second, we reverse the perspective, and shift from topics to cities as unit of analysis by aggregating these same documents at city level.

Cognitive topic proximity (indicator D)

The cognitive topic proximity can be used as basis for a path-dependent prediction model. For this analysis we design a corpus of all words used in the titles of our data set of transportation articles. A matrix of documents versus topics represents their relatedness. Then we make a selection in topics, based on a minimal frequency occurrence that still provides at least 500 topics for analysis. We have plotted some figures to determine how to choose these occurrences. Concerning their shape, all figures are identical. Therefore we only provide three plots, to give an indication of their ratios.

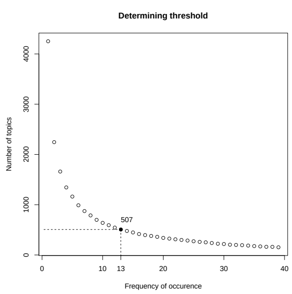

Figure 4.5: Frequency versus number of topics 1998-2004

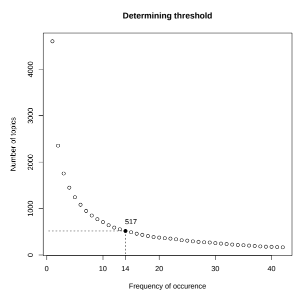

Figure 4.6: Frequency versus number of topics 2001-2007

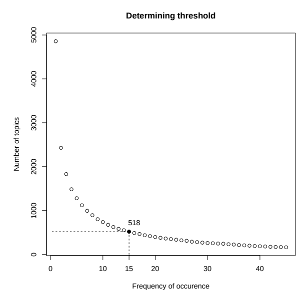

Figure 4.7: Frequency versus number of topics 2004-2010

After we restricted ourselves to the more frequently used words, analysis was reduced considerably. With a reduced matrix of documents and topics, we can calculate a cosine matrix of topics or documents, as explained in more detail in the operationalisation (section 3.3).

The first scenario, placing the behaviour of topics centrally, might be most fruitful approach. This builds a basis for a dynamic analysis, as the configuration of topics changes more rapidly than those of cities over the years. This however creates a new basis as well for testing prediction models, as we have to model the current topic network as well. This dynamic approaches involve more constraints for implementation. Therefore, we prefer to restrict ourselves to the same type of prediction as for the other indicators, which is the prediction of alleged adjacencies between cities.

When we restrict the models to predicting city adjacencies, we have two options. We can either use this indicator to predict document adjacencies first and then aggregate at city level, or reversed. The first option will give very little information, since there is very little differentiation within title words of documents. Therefore, we will have a closer look at using topic similarity by adding the geographical component in an earlier stage. This geographic topic proximity is the essence of the next indicator and is an expansion of cognitive topic proximity.

Geographic topic proximity (indicator E)

For the indicator of geographic topic proximity we aggregate documents by their geographic location, based on their coordinates of the global positioning system (gps). For each year, we design a prediction of adjacencies between cities, based on similarity of their topic portfolios of the past 7 years. We compare those predictions with the actual adjacencies for that year. We regard two cities adjacent when they have n topics in common. We have tested the prediction method for $n = 1, 2, 3$. However, these variation in n provided barely different values, so we only show results for $n = 1$ in this chapter. Similarity of topics has been suspected to influence adjacency, in the sense of positive correlation. Therefore, we choose to set a threshold for prediction of city adjacencies at 0.9. When two topics have a similarity higher than that threshold, cities who both studied one of those topics recently, will be adjacent next year according to our prediction model.

Geographic Topic Proximity	2005	2006	2007	2008	2009	2010	2011
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	5731	5487	5646	5437	5164	5121	5072
TP	511	239	195	79	347	56	169
FP	5220	5248	5451	5358	4817	5065	4903
FN	5175	2471	3106	2720	4568	4012	4564
TN	24339	27287	26493	27088	25513	26112	25609
Test statistic χ^2	262.8	101.2	276.0	369.2	262.5	639.5	518.5
p-value	0	0	0	0	0	0	0
Significant (alpha=0.05) effect	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Precision	0.089	0.044	0.035	0.015	0.067	0.011	0.033
Recall	0.09	0.088	0.059	0.028	0.071	0.014	0.036
F-score	0.09	0.06	0.04	0.02	0.07	0.01	0.03

Table 4.2: E: Results per year (for $0.9 \leq$ similarity ≤ 1.0)

Based on these results, it seems that the model provides us with a good prediction. The precision of the prediction is very low, and even the recall of adjacency prediction is lower than in the prediction model of indicator A. So, the pairs with highest proximity are not suitable for a prediction model as planned in this research. Therefore, we take a closer look at the distribution of the similarity of pairs, based on this proximity. This gives us a distribution as in figure 4.8. The exact shape varies per year, but the main tendency is a normal distribution with a peak close to similarity value 1.

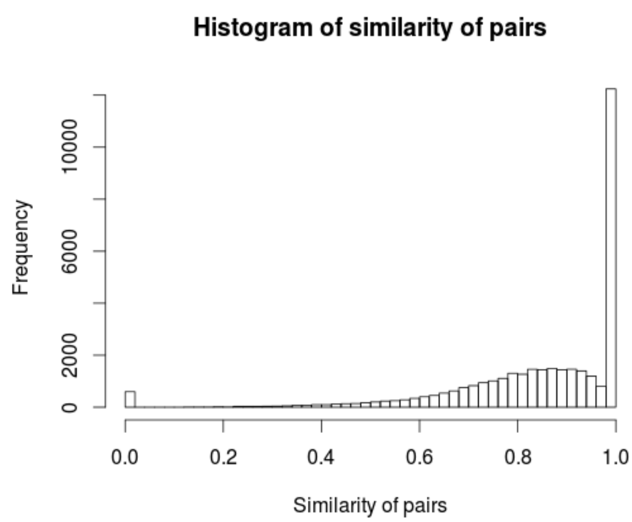

Figure 4.8: Distribution of similarity pairs

Because of the large number of pairs, in comparison to the total amount of pairs, we cannot regard them as outliers. However, this peak raises the suspicion that this part of the data should be treated differently from the rest of the pairs. Therefore, we exclude this exceptional peak from our prediction model. Instead, we choose to include the next range of 0.1 in similarity (so from 0.8 to 0.9) to use for a prediction model, while we leave the notion of positive correlation between topic similarity and city adjacency in tact. Results of this adjustment are shown in table 4.3.

Geographic Topic Proximity	2005	2006	2007	2008	2009	2010	2011
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	6091	6511	7218	7217	7016	7260	7080
TP	978	338	399	248	760	411	610
FP	5113	6173	6819	6969	6256	6849	6470
FN	4708	2372	2902	2551	4155	3657	4123
TN	24446	26362	25125	25477	24074	24328	24042
Test statistic χ^2	0.025	69.8	156.9	251.2	70.4	309.0	176.0
p-value	0.87	0	0	0	0	0	0
Significant (alpha=0.05) effect	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Precision	0.16	0.052	0.055	0.034	0.11	0.057	0.086
Recall	0.17	0.12	0.12	0.089	0.15	0.10	0.13
F-score	0.17	0.07	0.08	0.05	0.13	0.07	0.10

Table 4.3: E: Results per year (for $0.8 \leq$ similarity ≤ 0.9)

From these results, that still show the same low values of precision and recall, it is obviously clear that the model has either a low predictive value or it needs some variables to be chosen with more care. Two variables that are involved so far, are the threshold for city adjacency n and topic similarity. The threshold for city adjacency is tested for three values, without additional information after increasing the threshold. Topic adjacency has been regarded as a it would be positively correlated to the tendency for city adjacency. A more elaborate version of the same indicator is discussed in the next section (4.4).

4.4 Co-evolution: A Dual Network Analysis

Instead of studying the knowledge network by using the associated city and topic network separately, we now take a look at an approach that includes these dual networks simultaneously.

Absorptive capacity of cities (indicator C) and novelty degree of topics (indicator F)

The absorptive capacity of cities (section 2.3) and is linked to novelty degree of topics by the analysis of the heterogenitic network level (section 3.2.2). Since they already are theoretically introduced, we will focus here on the practical implementation. The problem is to fit these two concepts in meaningful functions, to find the optimum for the learning ability and likeliness to share knowledge of scientists, which are discussed in the proximity section (2.2). Nooteboom et al. (2007, p. 1018) has made a qualitative visualisation of this concept by means of a learning curve (figure 2.1). He claims that the optimum of that learning curve can be found at the intersection of absorptive capacity and novelty value of topics. Although the context of his claims is very different, we try to design suitable function for those two concepts as well in the context of knowledge production.

For absorptive capacity we use the similarity of topic portfolios. Even though it might not be directly lead to increased adjacency of cities, it increases theoretically the feasibility of adapting knowledge. The practical motivation, to actually realise that potential absorptive capacity, depends on the novelty value. The obvious function for that concept would be the reversed similarity function, which makes it easy to optimise the learning ability. For now, there are no theoretical clues for designing a weighted version of those functions.

4.4.1 Learning Curve

Based on the functions of absorptive capacity and novelty degree, we attempt to quantify the learning curve from the work of Nooteboom et al. (2007), as shown by the graph in section 2.3, figure 2.1.

After scaling absorptive capacity and novelty degree, we can regard values of similarity of topics and cities instead. Now, we only have to read the graph. Because of the scaling transformation, our functions automatically intersect at similarity 0.5. We will include city and topic pairs that have similarity 0.5 ± 0.05.

Geographic Topic Proximity & Learning Curve	2005	2006	2007	2008	2009	2010	2011
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	719	847	1031	953	898	992	933
TP	608	338	548	477	617	715	654
FP	111	509	483	476	281	277	279
FN	5078	2372	2753	2322	4298	3353	4079
TN	29448	32026	31461	31970	30049	30900	30233
Test statistic χ^2	2535.0	1264.4	2393.5	2369.9	2298.1	3657.4	2642.2
p-value	0	0	0	0	0	0	0
Significant (alpha=0.05) effect	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Precision	0.85	0.40	0.53	0.50	0.69	0.72	0.70
Recall	0.11	0.12	0.17	0.17	0.13	0.18	0.14
F-score	0.19	0.19	0.25	0.25	0.21	0.28	0.23

Table 4.4: Indicator C/F: Results per year (for $0.45 \leq$ similarity ≤ 0.55)

These results are based on a more specific motivated choice for topic similarity. It turns out that this similarity range for prediction of city adjacencies provides better results (see table 4.4). In particular precision has an extremely increased value. This improves prediction for small batches of city adjacencies. However, when one would like to clarify general tendency for emerging of adjacencies, the recall of a prediction is more important. Also these values are increased in comparison to tables 4.2 and 4.3, though they are still relatively low. Therefore, we included a final sub question that requires evaluation of the similarity variable experimentally as well for a larger set of similarity intervals.

4.5 Experimental Optimisation

After we have tested a few different ranges of similarity, based on theoretical concepts, we will now provide a small overview of experimental results of prediction models by varying that same range. Since there is no specific trend to test for, except for the three ranges as evaluated before, there is a risk of overfitting on our data. Findings from this evaluation are therefore only applicable to this transportation data set. They should be analysed on a larger data set to state general conclusions, but can confirm suspicions regarding the need for an approach to optimise the topic proximity range.

Also, a broad evaluation can lead to indications for more plausible choices for the functions of absorptive capacity and novelty degree of topics. For that purpose, we analyse the indicator that can benefit from optimisation of similarity range of topics, which is geographic topic proximity (indicator E). We distinguish 20 smaller intervals of 0.05 as proximity indicator function for prediction. This indicator function determines whether two cities are considered as predicted adjacencies. Extensive results of this analysis can be found in appendix A (p. 37). For now we are only interested in the resulting precision, recall, and F-score of those intervals.

The mean values over the period 2005-2011 are captured by discrete plots (figures 4.9, 4.10, 4.11). Before we draw conclusions from this plots, it should be noted that these values are mean values. The distribution

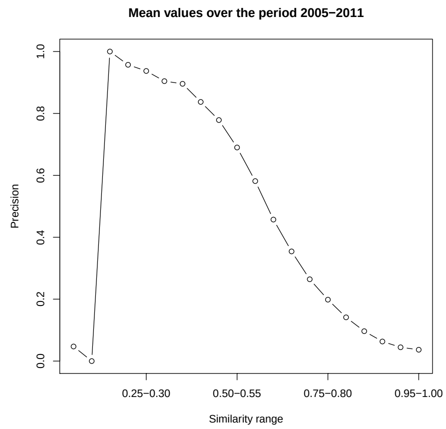

Figure 4.9: Evaluation plot of 'Precision'

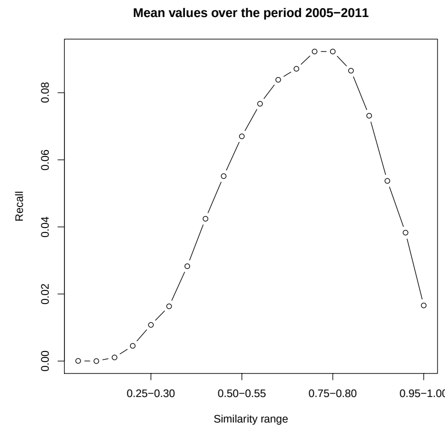

Figure 4.10: Evaluation plot of 'Recall'

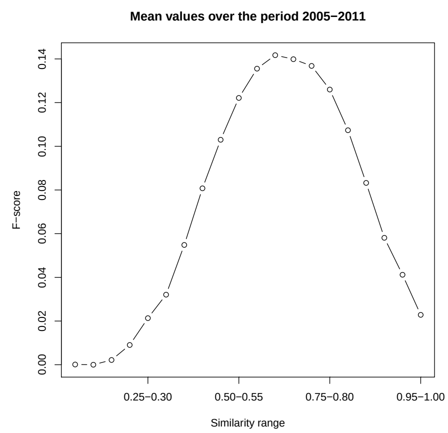

Figure 4.11: Evaluation plot of 'F-score'

over the various years determines the reliability of such conclusions. Separate plots per year can be found in appendix B (p. 42). These separate plots indicate very different values of precision in the tails of the distribution, which make them unreliable. However, the ranges of topic similarity that provide extreme values will be neglected in the design of prediction models. Roughly, we can regard the range $0.2 - 0.8$ as useful for interpretation.

We recognise an increasing function at this interval of $0.2 - 0.8$ in the plot of mean values of the recall of our latest prediction model and a decreasing function in the plot of the mean values of the precision. Although the exact values of these functions depend on the particular data set of transportation, the constructions might indicate a representation for respectively absorptive capacity and novelty value. Whether the recall and precision plots will generally show this pattern, should be tested for larger data sets. For now, we only have a solid basis for raising this suspicion for correlation between them.

Since the recall of these predictions is relatively low, we can choose wider intervals for topic similarity to increase the number of predicted adjacencies. Although this will probably negatively influence the precision, it might provide better results concerning the overall measure, reflected by the F-score. Based on a balanced success of both indicators, shown in the plots in 4.9 and 4.10, we test a final prediction model based on topic similarity on the interval $0.60 - 0.75$. These results are shown in figure 4.5.

Geographic Topic Proximity & Learning Curve	2005	2006	2007	2008	2009	2010	2011
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	3342	3905	4594	4686	4554	5021	5041
TP	1582	480	889	811	1258	1338	1439
FP	1760	3425	3705	3875	3296	3683	3602
FN	4104	2230	2412	1988	3657	2730	3294
TN	27799	29110	28239	28571	27034	27494	26910
Test statistic χ^2	2654.5	130.4	619.1	646.9	814.1	1306.9	1154.7
p-value	0	0	0	0	0	0	0
Significant (alpha=0.05) effect	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Precision	0.47	0.12	0.19	0.17	0.28	0.27	0.29
Recall	0.28	0.18	0.27	0.29	0.26	0.33	0.30
F-score	0.35	0.15	0.23	0.22	0.27	0.29	0.29

Table 4.5: Indicator C/F: Results per year (for $0.60 \leq$ similarity ≤ 0.75)

Although the presumed increase of recall indeed occurs, the precision of those predictions has strongly decreased. Concerning the reliability of this last prediction, we need to be careful. These results are obtained by a very rough indication of a suitable topic similarity. An exact analysis is required to balance the influence of precision and recall. However, what these results do show, is that the recall of the prediction indeed has increased.

Chapter 5

Discussion

We will start with the placing our results in the context of other findings in innovation and transport literature that studied knowledge production. Both the place-dependent and the path-dependent model, refined by a co-evolutionary approach, show a high significance level (section 4.2, table 4.1 and section 4.3, table 4.5). This is in line with the empirical findings in a different scientific field of Heimeriks and Boschma (2012), who detected both path- and place-dependency in knowledge production. The separation into three levels per network type, which was useful in a study of Balland (2009), was not very well applicable in this research. It made clear the differences in predictive strength, but in our findings the different network levels did not work harmonious together that well.

Our findings concerning the field of transportation are complementary to the stressed need for better roadmapping of Tuominen and Ahlqvist (2010). However, comparing the results to such studies is difficult, for these are purely qualitative concepts. At a quantitative level, this research is most closely related to the efforts of Wu (2013), who focused on the role of scientists by their citation rank and spatial diversity. The results of our research confirm his findings in the sense that nowadays cognitive elements, strongly related to the tacit knowledge in specific regions, enable more precise prediction than spatial patterns due to the increase of globalisation.

We will continue with some limitations of our research. Due to some arbitrary choices in methodology, data were transformed. Those choices were motivated throughout the report, but might nevertheless have consequences.

To reduce the size of topic data, we made a cut-off based on a minimum frequency and remained about 500 most common topics. On one hand, such a threshold for occurrence increases the reliability of the research, because anomalies are neglected through this precaution. On the other hand, these uncommon topics might conceal small developments in the scientific field. So this same precaution will occasionally destroy information. This observation is supported by the strong discrepancy in the histogram of topic pairs (section 4.2, figure 4.8), when similarity is almost 1. A solution to avoid this, can be to use more powerful hardware. More heuristic solutions are to fragment the data or to aggregate topics. Fragmentation has the downside that it disturbs the coherence of topics, while aggregation simply destroys information. However, based on the assumption that topics with very high proximity are studied together by anyway, those connections will not contribute to the prediction model and can therefore be aggregated. This latter notion might be a useful addition in the preprocessing of data.

An important effort for the validity of the research was the separation of data per year. A yearly series of topics as history to compare with the actual developments was applied seven times. This is a small sample to generalise the trends that are discovered, but it makes the obtained results sufficiently robust as it enables detection of anomalies. A more difficult part of validation was encountered in the comparability of prediction models. The intention of our research was to compare a combination of one-dimensional prediction models with a more dynamic co-evolutionary model. Since the first few attempts gave poor results in terms of predictive value very clearly, no obvious reference was provided for the final model.

Although we were able to show significant effects with the support of χ^2-tests, assessment of predictive value needs to be linked to results from other models or from practical application, judged by policy makers. This assessment should also involve whether results are specific and sensitive. For our prediction models we used respectively the precision and recall as indicators for that purpose, which can be linked to reality by the innovation theory about absorptive capacity (section 2.3), and will reflect the policy implications of our research.

Prediction of city adjacencies can be adjusted to the preference of increasing either precision or recall, with a decrease of the other indicator as a result. For policy decisions, one of those might have greater impact than the other. When they are responsible for specialisation and diversification of knowledge, a more precise forecast of some interesting new topics will have more value than the entire scale of topics for the next year. Policy makers responsible for distribution of funding on the other hand are more eager to have a broad picture to support the realisation of potential knowledge production in general. In the second example the recall of a prediction model needs to be maximised. This separation leads to two designs, with either a preference for high precision or high recall. Both designs can provide a maximal score per region, which enables policy makers to compare the learning curve of their region to other regions. This is an example of relative absorptive capacity, a concept of Cockburn and Henderson (1998), which helps to determine whether the potential absorptive capacity is actually realised. There role of realisation is a theoretical contribution of Zahra and George (2002) to the original concept of absorptive capacity of Cohen and Levinthal (1990). Finding a practical measure to determine the difference would be a useful addition.

From a theoretical point of view, absorptive capacity has been helpful in this research for an optimal cognitive topic proximity. We applied this qualitative idea of Nooteboom et al. (2007), which led to a refinement of the path-dependent model. This was a co-evolutionary approach of path- and place-dependency in the sense that we experimented with similarity ranges in the topic network, while the objective was to obtain more decent predictions for city adjacencies, which involves the dual network. In an attempt to quantify the relation between this duality and absorptive capacity, we found some correlations. When we tried to find a balance between the precision and recall of the model, the shape of their plots showed some resemblance with the suspected underlying patterns of absorptive capacity and novelty value. Repeated testing of those values, also in other fields than transportation, will clarify whether these concepts are indeed correlated with respectively recall and precision of prediction models. However, we did not succeed to generate an exact cognitive proximity to improve predictions. Therefore, it might be of more value to use the relation between absorptive capacity and optimal cognitive proximity in a reverse direction. Our results already showed that it is possible to experimentally optimise cognitive proximity (section 4.5). Then, based on desirable values of precision and recall, the presence of absorptive capacity can be assessed by this optimised value of proximity.

For the place-dependent model, based on indicator A, we have found no theoretical basis for improvement in the research. Perhaps a comparable inference can be made as for the path-dependent model, but there seems to be a larger difference between the design of both models. In our design of a place-dependent model, we have not included change of cities over time. This has resulted in a static model that predicts the same adjacencies every year. The model was intended that way and χ^2-tests showed a significant effect with city adjacencies, a lot of geographic dynamics are neglected due to this approach. Based on the characteristics of the path-dependent model, we will touch upon two possible extensions of a place-dependent model. First, we could involve geographical movement of scientists, obtaining a scientist portfolio per city that can change over time. A second option would be to distinguish more regions. Then, to reduce complexity of analysis, subsets of topics can be analysed in which a threshold should be set for cities, depending on their activity in knowledge production in that particular scientific sub field.

As a more general implication for further research, we will end with ideas for the improvement of the design of our prediction models. To obtain better predictive results for knowledge production in a scientific field, qualitative concepts need to guide the relative importance of precision and recall of a prediction. The score of precision and recall behave similar to respectively a topic's novelty value and a city's absorptive capacity (section 4.5, figure 4.9 and 4.10 versus section 4.4.1, figure 2.1). It is unclear whether these are related to each other, but both pairs require a certain balance for an optimal learning curve. When the relative importance of precision and recall is established, a suitable range of topic similarities can be chosen. Then, a path-dependent model, based on this optimal cognitive topic proximity range, can be combined with the one-dimensional model of place-dependency, based on geographic city proximity.

Chapter 6

Conclusion

By researching path- and place-dependency we have designed prediction models for knowledge diffusion. We also have refined both designs by exploring interdependencies between those concepts of dependency. These attempts together enable us to answer our main research question: "How can a model of path- and place-dependency explain evolutionary patterns of knowledge production in the field of transportation?"

In the conceptual model (section 2.4, figure 2.2) we explained the composition of the research. The lower half of the conceptual model resembles the empirical study. We have replicated this part and pointed out which indicators lead to the intended prediction models (figure 6.1). Indicators A and D/E have provided the basis for one-dimensional prediction models, which will provide answers to the first two sub questions of the research. The other sub questions are concerned with the results of the integrated prediction models. The first integrated model aimed to combine the one-dimensional models, while the second integrated model involved an approach of co-evolution of cities and topics, obtained by indicators C and F. The last design is very similar to the fourth, with the difference that the required choice for topic proximity was made experimentally. For all models, both one-dimensional and integrated, we calculated the significance level and predictive value. The significance level shows whether there is a correlation at all, while the precision and recall determine the presence of path- and place-dependent patterns in our case study.

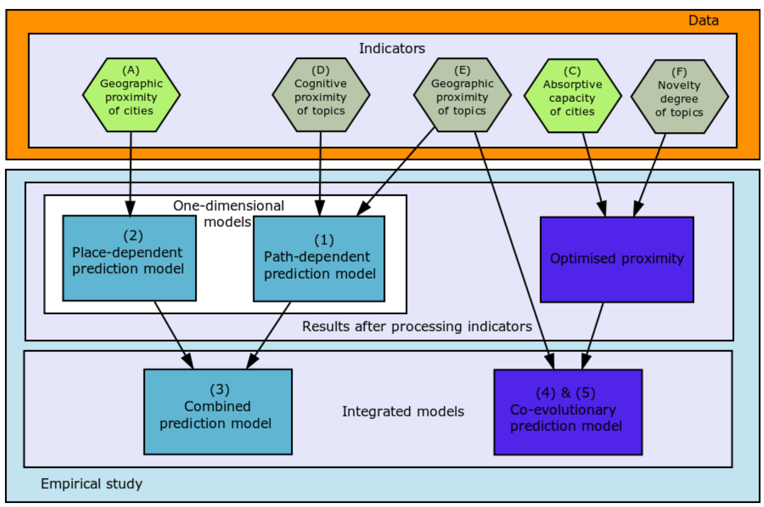

Figure 6.1: Prediction Models

The first objective of the research was to design a prediction model that resembles path-dependency. Based on analysis of cognitive topic proximity (indicator D), we aimed to set a topic similarity threshold to predict future adjacencies. Since we decided to measure evolution in the network by the behaviour of cities, the results of this indicator should be adjusted to the city level representation as well. Combining cognitive topic proximity with aggregation at city level is similar to the general idea of geographic topic proximity as described by the geographic dimension of the same topic network (indicator E). The first exploration of a path-dependent prediction model uses indicator E together with the assumption that topic similarity is positively correlated with the tendency to gain adjacencies between cities. We found a significant effect (for $\alpha = 0.05$) to support this conjecture (section 4.3, table 4.2 and 4.3). However, since both precision and recall of this indicator score low, the method does not fulfil the aim of explaining evolutionary patterns. So we can conclude there is still need for improvement of this path-dependent prediction model.

The second objective was to design a prediction model that emphasises place-dependency. Based on the analyses of geographic city proximity (indicator A), we designed a prediction model that assumes a positive correlation between geographic similarity of cities and the tendency for future adjacency. The effect of geographic similarity has been shown to be significantly present (for $\alpha = 0.05$). This place-dependent prediction model has, compared to its path-dependent counterpart, an acceptable F-score for its precision and recall (section 4.2, table 4.1). We will compare these values more thoroughly when we discuss the integrated model of co-evolution.

The third part of the research intended to focus on optimisation of the ratio of path- and place-dependent prediction models. However, due to the lack of a satisfying one-dimensional model of path-dependency, there neither is a suitable method for a combination. Therefore, we are not able to answer the third sub question of our research. Instead of combining path- and place-dependent models, we have to combined cognitive and geographic perspectives to refine the current path-dependent model.

For the fourth part of the research, we improved our original path-dependent prediction model, designed by cognitive and geographic topic proximity (indicators D and E). For this purpose, we used the intersection of absorptive capacity (indicator C) and novelty degree (indicator F), which brings the co-evolving behaviour of cities and topics together. We attempted to deduce an exact value of proximity from this intersection to determine an optimal prediction. The qualitative conjecture was that an intersection at the cognitive topic proximity value of 0.5 should be optimal. Although this value seemed very arbitrary, it provided an improved version in terms of its precision (section 4.4, table 4.4). However, the recall of the model is still very low, which is inconvenient.

The fifth part of the research involved a final effort to increase the recall of predictions, based on the same notion of co-evolution. We experimentally optimised the learning curve for the scientific field of transportation. Instead of holding on to the idea that this curve determines the similarity threshold for topics, we reversed the alleged causality. We made the assumption that these adjacencies are a resemblance of the optimal absorptive capacity of a city. When this assumption can be founded, practical meaning can be deduced from it. In this research, we only focused on the experimental part of this reversed idea. We chose various cognitive topic proximity values to acquire the proximity range that predicts city adjacencies as precise and specific as possible. Our refinement uses a wider range of cognitive topic proximity ($0.60 - 0.75$), which induces a decrease of precision, but an increase of recall (section 4.5, table 4.5).

To come to a final answer for our research question, we will compare the various designs. For a high precision, the co-evolutionary model, which uses absorptive capacity for a refinement, is most suitable (section 4.4, table 4.4); this is a improved design of the path-dependent model. For a higher recall, the place-dependent prediction model can be used (section 4.2, table 4.1). A slightly lower recall has been the outcome of another co-evolutionary refinement as well (section 4.5, table 4.5). This latter design experimentally optimises topic proximity and provides the most promising prediction of knowledge diffusion. We regard this final model most promising, as it has the best result in terms of its F-score, which represents the harmonic mean of a prediction's precision and recall.

Acknowledgements

I would like to thank all the people who have helped and supported me during the research and writing period of my master thesis. First of all, my gratitude goes out to my supervisor, dr. Gaston Heimeriks, for his ideas, patience and reflection on my work. Also, I would like to thank the second reader of my research proposal and final report, dr. Floortje Alkemade, for taking time to read my work and reflect on it. Finally, I would like to thank my friends and family for their unconditional support during this project.

Appendices

Appendix A

Extensive results of the analysis in section 4.5.

	2005	2006	2007	2008	2009	2010	2011
$0.00 \leq$ sim ≤ 0.05							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	1	3	0	0	0	1	1
TP	0	1	0	0	0	0	0
FP	1	2	0	0	0	1	1
FN	5686	2709	3301	2799	4915	4068	4733
TN	29558	32533	31944	32446	30330	31176	30511
Precision	0.00	0.33	0	0	0	0.00	0.00
Recall	0.00	0.00037	0.00	0.00	0.00	0.00	0.00
F-value	0	7e-04	0	0	0	0	0
$0.05 \leq$ sim ≤ 0.10							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	2	0	2	0	0	0	0
TP	0	0	0	0	0	0	0
FP	2	0	2	0	0	0	0
FN	5686	2710	3301	2799	4915	4068	4733
TN	29557	32535	31942	32446	30330	31177	30512
Precision	0.00	0	0.00	0	0	0	0
Recall	0.00	0.00	0.00	0.00	0.00	0.00	0.00
F-value	0	0	0	0	0	0	0
$0.10 \leq$ sim ≤ 0.15							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	2	2	2	3	4	6	12
TP	2	2	2	3	4	6	12
FP	0	0	0	0	0	0	0
FN	5684	2708	3299	2796	4911	4062	4721
TN	29559	32535	31944	32446	30330	31177	30512
Precision	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Recall	0.00035	0.00074	0.00061	0.0011	0.00081	0.0015	0.0025
F-value	7e-04	0.001	0.001	0.002	0.002	0.003	0.005
$0.15 \leq$ sim ≤ 0.20							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	7	7	15	24	24	24	27
TP	7	7	14	24	23	22	24
FP	0	0	1	0	1	2	3
FN	5679	2703	3287	2775	4892	4046	4709
TN	29559	32535	31943	32446	30329	31175	30509
Precision	1.00	1.00	0.93	1.00	0.96	0.92	0.89
Recall	0.0012	0.0026	0.0042	0.0086	0.0047	0.0054	0.0051
F-value	0.002	0.005	0.008	0.02	0.009	0.01	0.01

Table 6.1: City adjacency predictions: an evaluation of topic similarity values (1/5)

	2005	2006	2007	2008	2009	2010	2011
$0.20 \leq \text{sim} \leq 0.25$							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	19	36	37	54	49	45	62
TP	19	35	37	49	45	41	53
FP	0	1	0	5	4	4	9
FN	5667	2675	3264	2750	4870	4027	4680
TN	29559	32534	31944	32441	30326	31173	30503
Precision	1.00	0.97	1.00	0.91	0.92	0.91	0.85
Recall	0.0033	0.013	0.011	0.018	0.0092	0.01	0.011
F-value	0.007	0.03	0.02	0.03	0.02	0.02	0.02
$0.25 \leq \text{sim} \leq 0.30$							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	58	59	63	57	65	96	90
TP	53	52	55	53	61	86	81
FP	5	7	8	4	4	10	9
FN	5633	2658	3246	2746	4854	3982	4652
TN	29554	32528	31936	32442	30326	31167	30503
Precision	0.91	0.88	0.87	0.93	0.94	0.90	0.90
Recall	0.0093	0.019	0.017	0.019	0.012	0.021	0.017
F-value	0.02	0.04	0.03	0.04	0.02	0.04	0.03
$0.30 \leq \text{sim} \leq 0.35$							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	79	108	119	136	135	117	121
TP	77	94	112	119	116	102	107
FP	2	14	7	17	19	15	14
FN	5609	2616	3189	2680	4799	3966	4626
TN	29557	32521	31937	32429	30311	31162	30498
Precision	0.97	0.87	0.94	0.88	0.86	0.87	0.88
Recall	0.014	0.035	0.034	0.043	0.024	0.025	0.023
F-value	0.03	0.07	0.07	0.08	0.05	0.05	0.04
$0.35 \leq \text{sim} \leq 0.40$							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	139	182	199	212	189	197	213
TP	125	135	166	170	168	166	183
FP	14	47	33	42	21	31	30
FN	5561	2575	3135	2629	4747	3902	4550
TN	29545	32488	31911	32404	30309	31146	30482
Precision	0.90	0.74	0.83	0.80	0.89	0.84	0.86
Recall	0.022	0.05	0.05	0.061	0.034	0.041	0.039
F-value	0.04	0.09	0.09	0.11	0.07	0.08	0.07

Table 6.2: City adjacency predictions: an evaluation of topic similarity values (2/5)

	2005	2006	2007	2008	2009	2010	2011
$0.40 \leq \text{sim} \leq 0.45$							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	245	277	316	286	263	248	280
TP	218	164	221	199	224	210	243
FP	27	113	95	87	39	38	37
FN	5468	2546	3080	2600	4691	3858	4490
TN	29532	32422	31849	32359	30291	31139	30475
Precision	0.89	0.59	0.70	0.70	0.85	0.85	0.87
Recall	0.038	0.061	0.067	0.071	0.046	0.052	0.051
F-value	0.07	0.11	0.12	0.13	0.09	0.10	0.10
$0.45 \leq \text{sim} \leq 0.50$							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	311	350	426	367	397	436	377
TP	270	166	249	223	306	341	283
FP	41	184	177	144	91	95	94
FN	5416	2544	3052	2576	4609	3727	4450
TN	29518	32351	31767	32302	30239	31082	30418
Precision	0.87	0.47	0.58	0.61	0.77	0.78	0.75
Recall	0.047	0.061	0.075	0.08	0.062	0.084	0.06
F-value	0.09	0.11	0.13	0.14	0.12	0.15	0.11
$0.50 \leq \text{sim} \leq 0.55$							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	408	496	605	586	501	554	554
TP	338	172	299	254	311	374	371
FP	70	324	306	332	190	180	183
FN	5348	2538	3002	2545	4604	3694	4362
TN	29489	32211	31638	32114	30140	30997	30329
Precision	0.83	0.35	0.49	0.43	0.62	0.68	0.67
Recall	0.059	0.063	0.091	0.091	0.063	0.092	0.078
F-value	0.11	0.11	0.15	0.15	0.11	0.16	0.14
$0.55 \leq \text{sim} \leq 0.60$							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	551	673	796	791	782	841	753
TP	406	164	306	265	383	437	372
FP	145	509	490	526	399	404	381
FN	5280	2546	2995	2534	4532	3631	4361
TN	29414	32026	31454	31920	29931	30773	30131
Precision	0.74	0.24	0.38	0.34	0.49	0.52	0.49
Recall	0.071	0.061	0.093	0.095	0.078	0.11	0.079
F-value	0.13	0.10	0.15	0.15	0.13	0.18	0.14

Table 6.3: City adjacency predictions: an evaluation of topic similarity values (3/5)

	2005	2006	2007	2008	2009	2010	2011
$0.60 \leq \text{sim} \leq 0.65$							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	792	908	1021	1082	1044	1139	1121
TP	481	165	274	268	393	442	449
FP	311	743	747	814	651	697	672
FN	5205	2545	3027	2531	4522	3626	4284
TN	29248	31792	31197	31632	29679	30480	29840
Precision	0.61	0.18	0.27	0.25	0.38	0.39	0.40
Recall	0.085	0.061	0.083	0.096	0.08	0.11	0.095
F-value	0.15	0.09	0.13	0.14	0.13	0.17	0.15
$0.65 \leq \text{sim} \leq 0.70$							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	1076	1301	1505	1557	1460	1655	1648
TP	521	164	317	287	428	453	477
FP	555	1137	1188	1270	1032	1202	1171
FN	5165	2546	2984	2512	4487	3615	4256
TN	29004	31398	30756	31176	29298	29975	29341
Precision	0.48	0.13	0.21	0.18	0.29	0.27	0.29
Recall	0.092	0.061	0.096	0.10	0.087	0.11	0.10
F-value	0.15	0.08	0.13	0.13	0.13	0.16	0.15
$0.70 \leq \text{sim} \leq 0.75$							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	1474	1696	2068	2047	2050	2227	2272
TP	580	151	298	256	437	443	513
FP	894	1545	1770	1791	1613	1784	1759
FN	5106	2559	3003	2543	4478	3625	4220
TN	28665	30990	30174	30655	28717	29393	28753
Precision	0.39	0.089	0.14	0.13	0.21	0.20	0.23
Recall	0.10	0.056	0.09	0.091	0.089	0.11	0.11
F-value	0.16	0.07	0.11	0.11	0.13	0.14	0.15
$0.75 \leq \text{sim} \leq 0.80$							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	2085	2313	2754	2755	2655	3044	2899
TP	564	191	274	184	446	416	461
FP	1521	2122	2480	2571	2209	2628	2438
FN	5122	2519	3027	2615	4469	3652	4272
TN	28038	30413	29464	29875	28121	28549	28074
Precision	0.27	0.083	0.099	0.067	0.17	0.14	0.16
Recall	0.099	0.07	0.083	0.066	0.091	0.10	0.097
F-value	0.15	0.08	0.09	0.07	0.12	0.12	0.12

Table 6.4: City adjacency predictions: an evaluation of topic similarity values (4/5)

	2005	2006	2007	2008	2009	2010	2011
$0.80 \leq \text{sim} \leq 0.85$							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	2806	3061	3360	3407	3309	3566	3415
TP	553	180	214	149	444	280	343
FP	2253	2881	3146	3258	2865	3286	3072
FN	5133	2530	3087	2650	4471	3788	4390
TN	27306	29654	28798	29188	27465	27891	27440
Precision	0.20	0.059	0.064	0.044	0.13	0.079	0.10
Recall	0.097	0.066	0.065	0.053	0.09	0.069	0.072
F-value	0.13	0.06	0.06	0.05	0.11	0.07	0.08
$0.85 \leq \text{sim} \leq 0.90$							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	3285	3450	3858	3810	3707	3694	3665
TP	425	158	185	99	316	131	267
FP	2860	3292	3673	3711	3391	3563	3398
FN	5261	2552	3116	2700	4599	3937	4466
TN	26699	29243	28271	28735	26939	27614	27114
Precision	0.13	0.046	0.048	0.026	0.085	0.035	0.073
Recall	0.075	0.058	0.056	0.035	0.064	0.032	0.056
F-value	0.09	0.05	0.05	0.03	0.07	0.03	0.06
$0.90 \leq \text{sim} \leq 0.95$							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	3517	3501	3692	3643	3469	3581	3442
TP	332	165	134	61	228	49	131
FP	3185	3336	3558	3582	3241	3532	3311
FN	5354	2545	3167	2738	4687	4019	4602
TN	26374	29199	28386	28864	27089	27645	27201
Precision	0.094	0.047	0.036	0.017	0.066	0.014	0.038
Recall	0.058	0.061	0.041	0.022	0.046	0.012	0.028
F-value	0.07	0.05	0.04	0.02	0.05	0.01	0.03
$0.95 \leq \text{sim} \leq 1.00$							
Adjacencies	5686	2710	3301	2799	4915	4068	4733
Predicted adjacencies	2214	1986	1954	1794	1695	1540	1630
TP	179	74	61	18	119	7	38
FP	2035	1912	1893	1776	1576	1533	1592
FN	5507	2636	3240	2781	4796	4061	4695
TN	27524	30623	30051	30670	28754	29644	28920
Precision	0.081	0.037	0.031	0.01	0.07	0.0045	0.023
Recall	0.031	0.027	0.018	0.0064	0.024	0.0017	0.008
F-value	0.05	0.03	0.02	0.008	0.04	0.002	0.01

Table 6.5: City adjacency predictions: an evaluation of topic similarity values (5/5)

Appendix B

Separate plots per year of the analysis in section 4.5. Note that those plots are discrete. The line segments in between are only to visualise the trend.

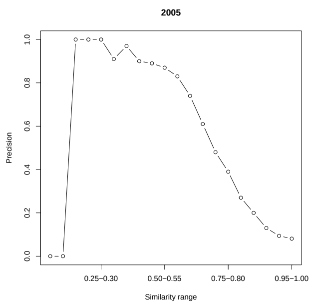

Figure 6.2: Precision 2005

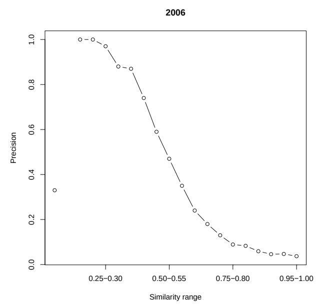

Figure 6.3: Precision 2006

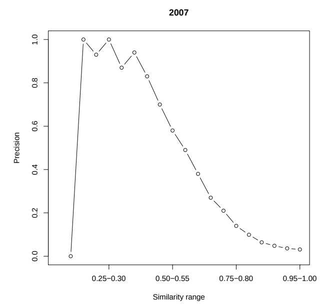

Figure 6.4: Precision 2007

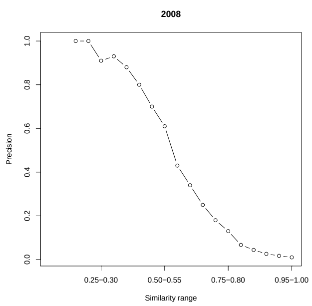

Figure 6.5: Precision 2008

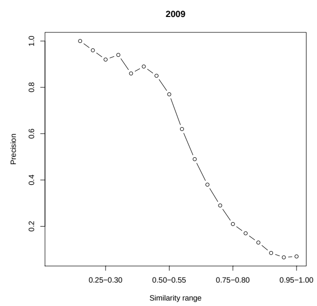

Figure 6.6: Precision 2009

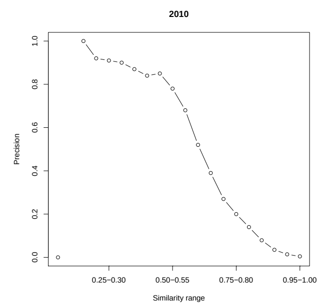

Figure 6.7: Precision 2010

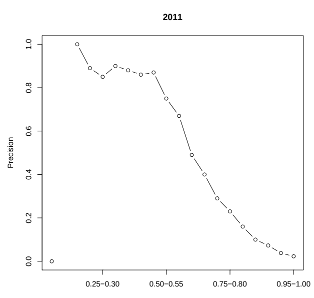

Figure 6.8: Precision 2011

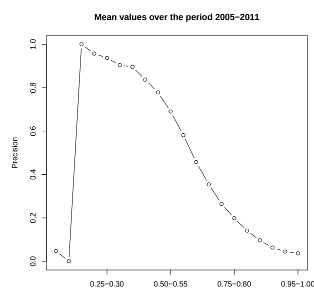

Figure 6.9: Precision: Mean

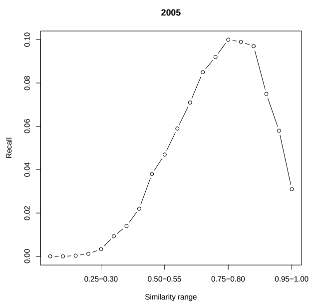

Figure 6.10: Recall 2005

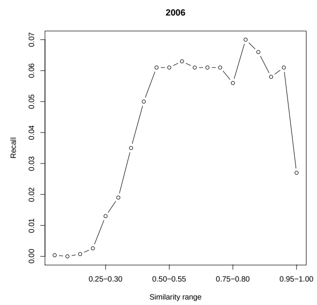

Figure 6.11: Recall 2006

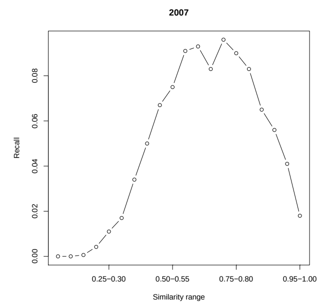

Figure 6.12: Recall 2007

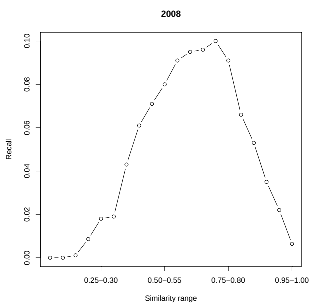

Figure 6.13: Recall 2008

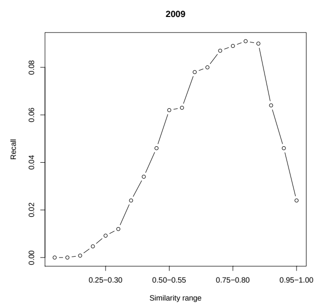

Figure 6.14: Recall 2009

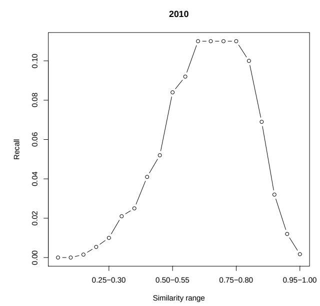

Figure 6.15: Recall 2010

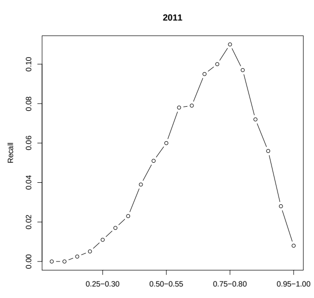

Figure 6.16: Recall 2011

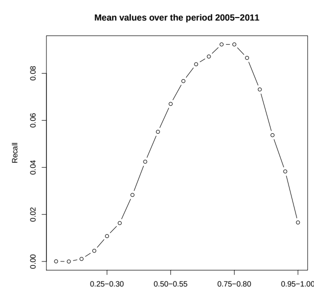

Figure 6.17: Recall: Mean

References

Agarwal, P. (2004). Contested Nature of Place: Knowledge Mapping for Resolving Ontological Distinctions between Geographical Concepts. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 3234*: 1-21.

Almeida, P., Kogut, B. (1999). Localization of Knowledge and the Mobility of Engineers in Regional Networks. *Management Science 45*(7): 905-917.

Amin, A., Cohendet, P. (2000). Organisational Learning and Governance through Embedded Practices. *Journal of Management and Governance, 4*(1-2): 93-116.

Amin, M., Mabe, M. A. (2003). Impact factors: Use and abuse. *Medicina 63*(4): 347-354.

Arthur, W. B., Polak, W. (2006). The Evolution of Technology within a Simple Computer Model. *Complexity, 11*(5): 23-31.

Arthur, W. B. (2009). The Nature of Technology: What it is and How it Evolves. New York: The Free Press.

Asheim, B. T., Coenen, L. (2006). Contextualising Regional Innovation Systems in a Globalising Learning Economy: On Knowledge Bases and Institutional Frameworks. *Journal of Technology Transfer 31*(1): 163-173.

Bailón-Moreno, R., Jurado-Alameda, E., Ruiz-Baños, R., Courtial, J. P., Jiménez-Contreras, E. (2007). The Pulsing Structure of Science: Ortega y Gasset, Saint Matthew, Fractality and Transfractality. *Scientometrics 71*(1): 3-24.

Balconi, M., Breschi, S., Lissoni, F. (2004). Networks of Inventors and the Role of Academia: An Exploration of Italian Patent Data. *Research Policy 33*(1): 127-145.

Balland, P. -A. (2009). Proximity and the Evolution of Collaboration Networks: Evidences from R&D Projects within the GNSS Industry. *Papers in Evolutionary Economic Geography 9.14*, Utrecht University, Utrecht.

Balland, P. -A., Vaan, M. de, Boschma, R. (2011). The Dynamics of Interfirm Networks along the Industry Life Cycle: The Case of the Global Video Games Industry 1987-2007. *Papers in Evolutionary Economic Geography 11.14*, Utrecht University, Utrecht.

Beal GM, Dissanayake W, Konoshima S. (1986). Some key issues. In: Beal GM, Dissanayake W, Konoshima S, editors. Knowledge Generation, Exchange and Utilization. *Boulder CO: Westview Press*: 463-467.

Becker, R. A., Wilks., A. R. (2012). mapdata: Extra Map Databases. Original S code, R version by Ray Brownrigg. R package version 2.2-1. `http://CRAN.R-project.org/package=mapdata`.

Besselaar, P. van den, Heimeriks, G. (2006). Mapping Research Topics Using Word-reference Co-occurrences: A Method and an Exploratory Case Study. *Scientometrics 68*(3): 377-393.

Boschma, R. A. (2004). Competitiveness of Regions from an Evolutionary Perspective. *Regional Studies 38*(9): 1001-1014.

Boschma, R. A. (2005). Proximity and Innovation: A Critical Assessment. *Regional Studies 39*(1): 61-74.

Boschma, R. A., Fornahl, D. (2011). Cluster Evolution and a Roadmap for Future Research. *Papers in Evolutionary Economic Geography 11.17*, Utrecht University, Utrecht.

Boschma, R. A., Martin, R. (2010). The Aims and Scope of Evolutionary Economic Geography. *Papers in Evolutionary Economic Geography 10.1*, Utrecht University, Utrecht.

Boschma, R. A., Heimeriks, G., Balland, P.-A. (2013). Scientific Knowledge Dynamics and Relatedness in Bio-Tech Cities. *Papers in Evolutionary Economic Geography 13.04*, Utrecht University, Utrecht.

Boschma, R. A. (2013). Constructing Regional Advantage and Smart Specialization: Comparison of Two European Policy Concepts. *Papers in Evolutionary Economic Geography 13.22*, Utrecht University, Utrecht.

Broström, A. (2010). Working with Distant Researchers - Distance and Content in University-industry Interaction. *Research Policy 39*(10): 1311-1320.

Carter, C. R. (2008). Knowledge Production and Knowledge Transfer: Closing the Research-practice Gap. *Journal of Supply Chain Management 44*(2), 78-82.

Cockburn, I.M., Henderson, R.M. (1998). Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery *Journal of Industrial Economics 46*(2): 157-182.

Cohen, W. M., Levinthal, D. A. (1990). Absorptive Capacity: A New Perspective on Learning and Innovation *Administrative Science Quarterly (Special Issue: Technology, Organizations, and Innovation) 35*(1): 128-152.

Cooke, P. (2007). To Construct Regional Advantage from Innovation Systems First Build Policy Platforms. *European Planning Studies 15*(2): 179-194.

Ernst, D., Kim, L. (2002). Global Production Networks, Knowledge Diffusion, and Local Capability Formation. *Research Policy 31*: 1417-1429.

European Union (2010). Intelligent Transport Systems in the field of road transport and interfaces with other transport modes *** II European Parliament legislative resolution of 6 July 2010 on the Council position at first reading with a view to the adoption of a directive of the European Parliament and of the Council on the framework for the deployment of Intelligent Transport Systems in the field of road transport and for interfaces with other modes of transport (06103/4/2010 C7-0119/2010 2008/0263(COD)).

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters 27*(8)861-874.

Feinerer, I., Hornik, K., Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software 25*(5): 1-54. `http://www.jstatsoft.org/v25/i05/`.

Feinerer, I., Hornik, K. (2013). tm: Text Mining Package. R package version 0.5-8.3. `http://CRAN.R-project.org/package=tm`.

Fienberg, S. E. (2011). The Analysis of Contingency Tables: From Chi-Squared Tests and Log-Linear Models to Models of Mixed Membership. *Statistics in Biopharmaceutical Research 3*(2): 173-184.

Furman, J. L., MacGarvie, M. J. (2007). Academic Science and the Birth of Industrial Research Laboratories in the U.S. Pharmaceutical Industry. *Journal of Economic Behavior and Organization 63*(4): 756-776.

Gao, J., Buldyrev, S. V., Havlin, S., Stanley, H. E. (2011). Robustness of a Network of Networks. *Physical Review Letters 107*(19): 195701.

Giannopoulos, G. A. (2009). Towards a European ITS for Freight Transport and Logistics: Results of Current EU Funded Research and Prospects for the Future. *European Transport Research Review, 1*(4): 147-161.

Giuliani, E., Bell, M. (2005). The Micro-determinants of the Meso-level Learning and Innovation: Evidence from a Chilean Wine Cluster. *Research Policy 35*: 47-68.

Glänzel, W. (2001). National Characteristics in International Scientific Co-authorship Relations. *Scientometrics, 51*(1): 69-115.

Glänzel, W., Schubert, A. (2001). Double Effort = Double Impact? A Critical View at International Co-authorship in Chemistry. *Scientometrics 50*(2): 199-214.

Gray, M., Schubert, L. (2012). Sustainable Social Work: Modelling Knowledge Production, Transfer, and Evidence-based Practice. *International Journal of Social Welfare 21*(2): 203-214.

Hausmann, R., Hidalgo, C.A., Bustos, S., Coscia, M., Chung, S., Jimenez, J., Simoes, A., Yildirim, M.A. (2011). The Atlas of Economic Complexity: Mapping Paths to Prosperity. New Hampshire: Puritan Press.

Heimeriks, G., Boschma, R. (2012). The Path- and Place-dependent Nature of Scientific Knowledge Production in Biotech 1986-2008. *Papers in Evolutionary Economic Geography 12.10*, Utrecht University, Utrecht.

Heimeriks, G., Leydesdorff, L. (2012). Emerging search regimes: measuring co-evolutions among research, science, and society. *Technology Analysis & Strategic Management 24*(1): 51-67.

Hekkert, M. P., Suurs, R. A. A., Negro, S. O., Kuhlmann, S., Smits, R. E. H. M. (2007). Functions of Innovation Systems: A New Approach for Analysing Technological Change. *Technological Forecasting and Social Change 74*(4): 413-432.

Hoekman, J., Frenken, K., van Oort, F. (2009). The Geography of Collaborative Knowledge Production in Europe. *Annals of Regional Science, 43*(3 SPEC. ISS.), 721-738.

Katz, J. S. (1994). Geographical Proximity and Scientific Collaboration. *Scientometrics 31*(1): 31-43.

Kauffman, S. A. (1995). At Home in the Universe: The Search for the Laws of Self-Organization and Complexity. New York: Oxford University Press.

Kostoff, R.N., Schaller, R.R. (2001). Science and technology roadmaps. *IEEE Transactions on Engineering Management 48*(2): 132-143.

Lane, P. J., Lubatkin, M. (1998). Relative Absorptive Capacity and Interorganizational Learning. *Strategic Management Journal 19*: 461477.

Larsen, K. (2008). Knowledge Network Hubs and Measures of Research Impact, Science Structure, and Publication Output in Nanostructured Solar Cell Research. *Scientometrics 74*(1): 123-142.

Lazaric, N., Longhi, C., Thomas, C. (2008). Gatekeepers of Knowledge Versus Platforms of Knowledge: From Potential to Realized Absorptive Capacity. *Regional Studies Volume 42*(6): 837-852.

Lei, G., Xin, G. (2011). Social Network Analysis on Knowledge Sharing of Scientific Groups. In proceeding of: LISS 2011 - Proceedings of the 1st International Conference on Logistics, Informatics and Service Science 3, Beijing, China, 8 - 11 June, 2011.

Leydesdorff, L. (1989). Words and Co-Words as Indicators of Intellectual Organization. *Research Policy 18*: 209-223.

Leydesdorff, L., Cozzens, S. E. (1993). The delineation of specialties in terms of journals using the dynamic journal set of the SCI. *Scientometrics 26*(1): 135-156.

Leydesdorff, L. (2004a). Clusters and Maps of Science Journals Based on Bi-connected Graphs in Journal Citation Reports. *Journal of Documentation 60*(4): 371-427.

Leydesdorff, L. (2004b). Top-down Decomposition of the Journal Citation Report of the Social Science Citation Index: Graph- and Factor-analytical Approaches. *Scientometrics 60*(2): 159-180.

Leydesdorff, L. (2005). Similarity Measures, Author Cocitation Analysis, and Information Theory. *Journal of the American Society for Information Science & Technology 56*(7): 769-772.

Leydesdorff, L., Vaughan, L. (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the web environment. *Journal of the American Society for Information Science and Technology, 57*(12): 1616-1628.

Leydesdorff, L., Rafols, I. (2011). Local emergence and Global Diffusion of Research Technologies: An Exploration of Patterns of Network Formation. *Journal of the American Society for Information Science and Technology 62*(5): 846-860.

Leydesdorff, L., Welbers, K. (2011). The semantic mapping of words and co-words in contexts. *Journal of Informetrics 5*: 469-475.

Lucio-Arias, D., Leydesdorff, L. (2008). Main-path Analysis and Path-dependent Transitions in HistCite^{TM}-based Historiograms. *Journal of the American Society for Information Science and Technology 59*(12): 1948-1962.

Lundvall, B. -Å. (2007). National innovation systems - analytical concept and development tool. Industry and Innovation, 14(1): 95-119.

Martin, B. R., Irvine, J. (1983). Assessing Basic Research: Some Partial Indicators of Scientific Progress in Radio Astronomy. *Research Policy 12*(2): 61-90.

Mattes, J. (2012). Dimensions of Proximity and Knowledge Bases: Innovation between Spatial and Non-spatial Factors. *Regional Studies 46*(8): 1085-1099.

McPherson, M., Smith-Lovin, L., Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology 27*: 415-444.

Musiolik, J., Markard, J., Hekkert, M. (2012). Networks and Network Resources in Technological Innovation Systems: Towards a Conceptual Framework for System Building. *Technological Forecasting and Social Change 79*6:1032-1048.

Neffke, F., Henning, M., Boschma, R., (2011). How Do Regions Diversify over Time? Industry Relatedness and the Development of New Growth Paths in Regions. *Economic Geography 87*(3): 237-265.

Nelson, R. R., Winter, S. G., (1977). In Search for a Useful Theory of Innovation. *Research Policy 6*: 36-76.

Nerkar, A., Paruchuri, S. (2005). Evolution of R&D Capabilities: The Role of Knowledge Networks within a Firm. *Management Science 51*(5): 771-785.

Nooteboom, B., van Haverbeke, W., Duysters, G., Gilsing, V., van den Oord, A. (2007). Optimal cognitive distance and absorptive capacity. *Research Policy 36*: 1016-1034.

Palla, G., Barabási, A., Vicsek, T. (2007). Quantifying Social Group Evolution. *Nature 446*(7136): 664-667.

Orsenigo, L., Pammolli, F., Riccaboni, M. (2001). Technological Change and Network Dynamics: Lessons from the Pharmaceutical Industry. *Research Policy, 30*(3): 485-508.

Pettersson, F. (2013). From Words to Action: Concepts, Framings of Problems and Knowledge Production Practices in Regional Transport Infrastructure Planning in Sweden. *Transport Policy 29*: 13-22.

Quddus, M. A., Ochieng, W. Y., Noland, R. B. (2007). Current Map-matching Algorithms for Transport Applications: State-of-the Art and Future Research Directions. *Transportation Research Part C: Emerging Technologies, 15*(5): 312-328.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. `http://www.R-project.org`.

Ronde, P. (2003). Delphi Analysis of National Specificities in Selected Innovative Areas in Germany and France. *Technological Forecasting and Social Change 70*(5): 419-448.

RSiena, various contributors (2012). RSiena: Siena - Simulation Investigation for Empirical Network Analysis. R package version 1.1-212. `http://CRAN.R-project.org/package=RSiena`.

Sai, S., Altintas, O., Kenney, J., Tanaka, H., Inoue, Y. (2013). Current and future ITS. *IEICE Transactions on Information and Systems E96-D*(2): 176-183.

Schartinger, D., Rammer, C., Fischer, M. M., Fröhlich, J. (2002). Knowledge Interactions between Universities and Industry in Austria: Sectoral Patterns and Determinants. *Research Policy 31*(3): 303-328.

Shapin, S. (1998). Placing the View from Nowhere: Historical and Sociological Problems in the Location of Science. *Transactions of the Institute of British Geographers 23*(1): 5-12.

Shepard, R. N. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science 237*: 1317-1323.

Shinn, T. (2002). The Triple Helix and New Production of Knowledge: Prepackaged Thinking on Science and Technology. *Social Studies of Science 32*(4): 599-614.

Slavtchev, V. (2013). Proximity and the Transfer of Academic Knowledge: Evidence from the Spatial Pattern of Industry Collaborations of East German Professors. *Regional Studies 47*(5): 686-702.

Snijders, T. A. B., van de Bunt, G. G., Steglich, C. (2010). Introduction to Actor-based Models for Network Dynamics. *Social Networks 32*: 44-60.

South, A. (2012). rworldmap: Mapping Global Data, Vector and Raster. With contributions from Scutt-Phillips, J., Rowlingson, B., Bivand, R., Foster, P.; R package version 1.02. `http://CRAN.R-project.org/package=rworldmap`.

Stegmann, J., Grohmann, G. (2003). Hypothesis Generation Guided by Co-word Clustering. *Scientometrics, 56*(1): 111-135.

Tan, P. -N., Steinbach, M., Kumar, V. (2006). Introduction to Data Mining. Addison Wesley, Boston.

Torre, A., Rallet, A. (2005). Proximity and localization. *Regional Studies, 39*(1): 47-59.

Trippl, M. (2009). Islands of Innovation and Internationally Networked Labor Markets: Magnetic Centers for Star Scientists? *SRE - Discussion Papers, 2009*/06. Institut für Regional- und Umweltwirtschaft, WU Vienna University of Economics and Business, Vienna.

Tuominen, A., Himanen, V. (2007). Assessing the Interaction between Transport Policy Targets and Policy Implementation - A Finnish Case Study. *Transport Policy 14*(5), 388-398.

Tuominen, A., Leonardi, J., Rizet, C. (2008). Assessing the Fitness-for-purpose of Strategic Transport Research in Support of European Transport Policy. *European Journal of Transport and Infrastructure Research 8*(3):183-200.

Tuominen, A., Ahlqvist, T. (2009). Is the Transport System Becoming Ubiquitous? Socio-technical Roadmapping as a Tool for Integrating the Development of Transport Policies and Intelligent Transport Systems and Services in Finland. *Technological Forecasting and Social Change 77*(1): 120-134.

Urbina, E., Wolshon, B. (2003). National Review of Hurricane Evacuation Plans and Policies: A Comparison and Contrast of State Practices. *Transportation Research Part A: Policy and Practice 37*(3): 257-275.

Valdaliso, J., Elola, A., Aranguren, M., Lopez, S. (2011). Social Capital, Internationalization and Absorptive Capacity: The Electronics and ICT Cluster of the Basque Country. *Entrepreneurship & Regional Development 23*(910): 707-733.

Wal, L.J. ter (2009). The Structure and Dynamics of Knowledge Networks: a Proximity Approach. Disseration at Utrecht University.

Wild, F. (2011). lsa: Latent Semantic Analysis, R package version 0.63-3. `http://CRAN.R-project.org/package=lsa`.

Wu, J. (2013). Geographical Knowledge Diffusion and Spatial Diversity Citation Rank. *Scientometrics 94*(1): 181-201.

Xing, Z., Stroulia, E. (2006). Understanding the Evolution and Co-evolution of Classes in Object-oriented Systems. *International Journal of Software Engineering and Knowledge Engineering 16*(1): 23-51.

Yahoo! Geocoder (2013). GPS Visualizer, with Yahoo! as source of coordinates. `http://www.gpsvisualizer.com/geocoding.html`, last revised on 5 February 2013.

Yan, X., Zhang, H., Wu, C. (2012). Research and Development of Intelligent Transportation Systems. Paper presented at the Proceedings - 11th International Symposium on Distributed Computing and Applications to Business, Engineering and Science, DCABES 2012: 321-327.

Yuan, B. J. C., Hsieh, C., Chang, C. (2010). National Technology Foresight Research: A Literature Review from 1984 to 2005. *International Journal of Foresight and Innovation Policy 6*(1-3): 5-35.

Zahra, S. A., George, G. (2002). Absorptive Capacity: A Review, Reconceptualization, and Extension. *The Academy of Management Review 27*(2) (Apr., 2002):185-203.

Zucker, L. G., Darby, M. R., Armstrong, J. S. (2002). Commercializing Knowledge: University Science, Knowledge Capture, and Firm Performance in Biotechnology, Conference Paper. *Management Science 48*(1): 138-153.

Zhang, J. (2008). China's Dynamic Industrial Sector: The Internet Industry. *Eurasian Geography and Economics, 49*(5): 549-568.

Zitt, M., Ramanana-Rahary, S., Bassecoulard, E., Laville, F. (2003). Potential Science-technology Spillovers in Regions: An Insight on Geographic Co-location of Knowledge Activities in the EU. *Scientometrics 57*(2): 295-320.