

Dancing on the shoulders of giants

Citation for published version (APA):

Persoon, P. G. J. (2021). *Dancing on the shoulders of giants: knowledge dynamics of renewable energy technologies*. [Phd Thesis 1 (Research TU/e / Graduation TU/e), Industrial Engineering and Innovation Sciences]. Technische Universiteit Eindhoven.

Document status and date:

Published: 01/10/2021

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Dancing on the shoulders of giants

*Knowledge dynamics of
renewable energy technologies*

Peter Persoon

Dancing on the shoulders of giants
Knowledge dynamics of renewable energy technologies

Copyright © 2021 by Peter Persoon. All Rights Reserved.

Dancing on the shoulders of giants / by Peter Persoon.
Eindhoven: Technische Universiteit Eindhoven, 2021. Proefschrift.

Cover design by Chérique Cuppen

A catalogue record is available from the Eindhoven University
of Technology Library

ISBN 978-94-6423-490-9

The work in this thesis has been sponsored by NWO grant
number 452-13-010

Printed by proefschriftmaken.nl

Dancing on the shoulders of giants

Knowledge dynamics of renewable energy technologies

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische
Universiteit Eindhoven, op gezag van de rector magnificus
prof.dr.ir. F.P.T. Baaijens, voor een commissie aangewezen
door het College voor Promoties, in het openbaar te
verdedigen op
vrijdag 1 oktober 2021 om 16:00 uur

door

Petrus Gerardus Jozefus Persoon

geboren te Naaldwijk

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr.ir. Y.A.W. de Kort
1e promotor:	prof.dr. F. Alkemade
2e promotor:	prof.dr.ir.ing. R.N.A. Bekkers
externe leden:	prof.dr. K. Frenken (Utrecht University) prof.dr. R. Tijssen (Leiden University)
lid TU/e:	dr.ir. B. Walrave (TU/e Eindhoven)

Het onderzoek dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Summary

Mitigating global climate change requires, amongst other measures, the replacement of Fossil Fuel based Energy Technologies (FFETs) by Renewable Energy Technologies (RETs). Policies aiming to strengthen in particular RET development can benefit from a deep understanding of the characteristics of the knowledge required to develop RETs, i.e. the knowledge base of RETs.

Knowledge bases consist of various dimensions. A first relevant dimension of the knowledge base of a technology is the extent to which it builds on scientific knowledge, which is referred to as the 'science-dependence'. Technologies however also build on earlier technological knowledge, their development is 'cumulative'. A second dimension of the knowledge base of a technology is therefore the extent to which it depends on its earlier development, also referred to as 'technological cumulativeness'. A third relevant dimension is the mobility of knowledge, the extent to which knowledge travels geographically. This third dimension is expected to depend on the first two dimensions. In this research, I developed methodologies to study the science-dependence and technological cumulativeness, and systematically compared these to the knowledge mobility of various RETs.

In the first part of this research, I performed a detailed descriptive analysis of the science base of both RETs and FFETs, allowing me to study characteristic differences between both types of energy technologies. I found that RETs generally have a more substantial science base and draw on a more diverse set of scientific disciplines. On average, the science on which RETs build is more recent, less applied, and is published in journals with a higher WOS Journal Impact Factor. However, for different RETs (e.g., photovoltaics, wind turbines, and non-fossil fuels), I observed much more variation across these dimensions than for different FFETs (e.g., combustion and gas turbines). Furthermore, the broad spectrum of sciences on which RETs build largely includes the smaller spectrum on which FFETs build.

In the second part of this research, I performed a theoretical and empirical analysis of technological cumulativeness. Despite the recognized importance of this concept, approaches in the academic literature to cumulativeness vary, and it often remains unclear what role cumulativeness plays in developing technology. I characterize the cumulativeness of a technology by the structure of its knowledge base (that is, how knowledge flow connects inventions), which is different from, but closely related to, the size of its knowledge base (that is, the number of inventions). Approaching the knowledge base structure as a relational network, where inventions (nodes) are connected by knowledge flow (links), we can define indicators to proxy cumulativeness. A simple conceptual model of researching engineers allows me to analytically derive equations describing a proportional relationship between the cumulativeness and the size of the knowledge base, where the rate of proportion may vary across technologies. Empirical tests of this model, using patent data on in-

ventions, confirm this relation and indicate that the rate varies considerably across technologies. At the same time I found that across technologies, this rate is inversely related to the rate of invention over time. This suggests that the cumulateness increases relatively slow in rapidly growing technologies.

In the third part of my research, I investigated in more detail how the notions of 'network paths' and 'path length' can be used to study cumulative knowledge structures, where again the network nodes represent elements of knowledge (such as inventions or scientific papers) and links represent the knowledge flow between these. Starting from the Price model of network growth, I derive an exact solution for the path length distribution of all unique paths from a given initial node to each node in the network. I study the relative importance of the average in-degree and cumulative advantage effect and implement a generalization where the in-degree depends on the number of nodes. The cumulative advantage effect is found to fundamentally slow down path length growth. As the collection of all unique paths may contain many redundancies, I additionally consider the subset of the longest paths to each node in the network. As this case is more complicated, I only approximate the longest path length distribution in a simple context. Where the number of all unique paths of a given length grows unbounded, the number of longest paths of a given length converges to a finite limit, which depends exponentially on the given path length. Fundamental network properties and dynamics therefore characteristically shape cumulative structures in those networks, and should therefore be taken into account when studying those structures.

In the fourth part of this research, I determined for an extensive group of RETs both their science dependence and their technological cumulateness. I systematically compared these to the mobility of their knowledge base. Knowledge mobility measures the extent to which developing technological knowledge travels geographically, and is positively related to the analyticity of the knowledge base (i.e. the extent to which it builds on analytic knowledge) and negatively related to the cumulateness of a knowledge base. I identified a substantial group of RETs (photovoltaics, fuel cells, energy storage) which have a highly analytic knowledge base and (indeed) a substantial knowledge mobility, there is also a substantial group of RETs (wind turbines, solar thermal, geothermal, and hydro energy) for which the knowledge base is less analytic and (indeed) less mobile. Likewise, the technological cumulateness tends to be lower for the former than for the latter group.

The previously mentioned characteristics of the knowledge base of RETs have several implications for science and technology policies that aim to strengthen the development of RETs. RETs overall build strongly on scientific knowledge, significantly more than FFETs do. For that reason I expect policies promoting scientific research in general (and basic, high impact science in particular) to lead to a strengthening of RETs. At the same time, there is substantial variation across different RETs in both science-dependence and cumulateness, which characteristically relates to variation in other dimensions such as the knowledge mobility and the rate of invention over time. This calls for the policies which aim to strengthen the development of specific RETs to be specific for those RETs, i.e. taking for a given RET into account both the type of knowledge it builds on as well as the local presence of this knowledge and the difficulty of catching up with a possible knowledge gap.

Contents

1	Introduction	1
1.1	Knowledge base dimensions	2
1.1.1	The science dependence	3
1.1.2	The cumulateness of knowledge	4
1.1.3	The mobility of technological knowledge	6
1.2	Research contribution	8
2	The science base of renewables	11
2.1	Introduction	12
2.2	Theoretical aspects of science bases	13
2.3	Data and methods	15
2.3.1	Knowledge base definitions	15
2.3.2	RETs and FFETs patent classes	16
2.3.3	Linking NPL data with WOS journal entries	20
2.3.4	Science base characteristics, distributions and indicators	21
2.4	Results	22
2.4.1	Overall knowledge base characteristics	22
2.4.2	Science base distributions and indicators	26
2.5	Discussion	32
2.6	Conclusions and policy recommendations	32
	Appendices	35
2.A	Linking NPL data with WOS journal entries	35
2.B	JIF and NEF statistics	37
3	How cumulative is technological knowledge?	43
3.1	Introduction	44
3.2	Theoretical perspectives on technological cumulateness	45
3.3	Measuring cumulateness	47
3.3.1	The transversal dimension: Internal dependence	48
3.3.2	The longitudinal dimension: Internal path length	48
3.4	Modeling the knowledge dynamics	49
3.4.1	Invention as search process	49
3.4.2	Internal Dependence Dynamics	50
3.4.3	Internal Path Length Dynamics	50
3.5	Empirical analysis	53
3.5.1	Data description	53
3.5.2	Id and ipl for the focus technologies	56
3.5.3	Distributions of backward links and path length	58

3.5.4	Cross technology variations	60
3.5.5	Cumulativeness across technological fields	62
3.6	Discussion	65
3.7	Conclusions	67
3.8	Acknowledgments	69
3.9	Data availability	69
Appendices		70
3.A	Statistics of the linear fits	70
3.B	Overview of selected technologies and corresponding CPC classifications	74
3.C	Evaluating the distribution fits	74
3.D	Examiner versus applicant citations	77
3.E	Mathematical appendix	79
3.F	Data scripts	81
3.F.1	Technology selection using CPC group/sub group level	82
3.F.2	Technology selection using CPC class level	83
3.F.3	Selecting applicant-added references	84
4	Cumulative structure and path length in knowledge networks	87
4.1	Introduction	88
4.2	All unique paths in the Price model	90
4.2.1	Excluding the cumulative advantage effect	91
4.2.2	Including the cumulative advantage effect	92
4.2.3	Generalization for increasing in-degree	94
4.3	Sub-selecting the longest paths	95
4.3.1	Zeroth order approximation	96
4.3.2	First order approximation	97
4.3.3	Expected path length	98
4.4	Conclusions	99
4.5	Discussion	101
4.6	Acknowledgements	102
Appendices		103
4.A	Derivations with all unique paths	103
4.A.1	Excluding the cumulative advantage effect	103
4.A.2	Including the cumulative advantage effect	104
4.A.3	Generalization for increasing in-degree	112
4.B	Derivations with longest paths	115
4.B.1	First order approximation	115
5	The knowledge mobility of renewable energy technology	119
5.1	Introduction	119
5.2	Theory	120
5.2.1	Analyticity of knowledge	121
5.2.2	Technological cumulativeness	121
5.2.3	Knowledge mobility	122
5.2.4	Knowledge dimensions of RETs	123
5.3	Methodology	124
5.3.1	Patents	125

5.3.2	Indicators	125
5.3.3	Technology selection and descriptive statistics	128
5.4	Results	132
5.4.1	General relations between knowledge dimensions	132
5.4.2	Knowledge relations of RETs	134
5.5	Discussion	139
5.6	Conclusions and policy implications	140
5.7	Acknowledgements	141
Appendices		142
5.A	Country ranking by number of patents	142
5.B	Reference distances of inventors in and outside Europe and US	142
6	Conclusions and research implications	145
6.1	Main research conclusions	146
6.2	Policy implications	148
6.3	Implications to further research	150
6.3.1	Open problems	151
Author contributions table		153
Nederlandse samenvatting		155
Acknowledgements		159
Curriculum vitae		161
Bibliography		163

Chapter 1

Introduction

Whether it is a device to harvest flowers, or a ship to fly to the moon, humans have come up with ingenious technological solutions to fulfill their daily needs or solve their societal issues. By studying nature and putting natural phenomena to work, we have effectively created our own, second nature. While this is a success story in many ways, the negative effects of some technological solutions are not always obvious at the outset. This makes that our second nature introduces some problems of its own. One of those, which is of particular urgency today, is the issue of global warming caused by excessive greenhouse gas emissions (IPCC, 2018). Although this problem could partially be solved by a set of behavioral changes, for the remaining part we again depend on the solutions provided by technological development.

This applies in particular to the gradual replacement of Fossil Fuel based Energy Technologies (FFETs) by Renewable Energy Technologies (RETs). In 2016 the energy sector accounted for about 60% of the greenhouse gas emissions worldwide (Ritchie & Roser, 2020). While it is, on the one hand, encouraging that many countries and organizations have policies in place to strengthen the development of RETs and the share of renewable energy sources in our total energy mix has been steadily increasing (IRENA, 2018), it is, on the other hand, alarming that the FFETs still largely dominate the energy mixes worldwide and emissions are peaking. Furthermore, introducing greater shares of renewable energy will bring about all kinds of serious further technological challenges such as the requirement of greater grid and storage capacities (IRENA, 2018). While on our way, therefore, we still have a long way to go.

There are various ways for policies to strengthen or accelerate the development of RETs. A first way for policies to do this is by implementing a regulatory framework that (partly) prohibits the use of FFETs, thus forcing people to choose for RETs. If effective at all, the measures of this type are usually avoided in free-market economies. In a second way, policies intervene in the diffusion process of RETs through tax reductions and subsidies, resulting in economies of scale due to optimization of production processes. While these demand stimulation measures work well to create a level-playing field for RETs that are (almost) market-ready, their impact on the development of novel RETs is unclear and only indirect at best. Stimulating the development of the latter is however equally relevant, especially from a long-term perspective. Therefore, in a third way of strengthening RET development, policies intervene at a more fundamental level of knowledge development, specifically aiming to encourage the knowledge development of RETs and not

of FFETs. Coming up with policies that can effectively aim research and technology development towards a certain mission is, however, not straightforward (Mazzucato, 2016). These policies should make well-considered choices about stimulating the development of which technology where. As I discuss in more detail later, the difficulty to enter the development of a technology and the extent to which its development is location-bound are expected to partially depend on fundamental properties of the knowledge underlying the technology. Organizing such mission-oriented research therefore in the first place requires a deep understanding of the body of knowledge on which technologies build, that is the *knowledge base* of these technologies, and, in particular, the knowledge base dimensions that relate to the difficulty of entry and the extent to which the development is location-bound. While a number of contributions study various different knowledge aspects of RETs (Barbieri et al., 2020; Dechezleprêtre et al., 2014; Ocampo-Corrales et al., 2020), the literature currently lacks a systematic overview of what the knowledge base of RETs exactly entails, and how it may vary for individual RETs. In this research, I aim to improve this understanding, by developing methodologies to systematically study various knowledge base dimensions and applying these specifically to RETs.

1.1 Knowledge base dimensions

What are the relevant dimensions in which the knowledge base of a technology can be studied? Before I dive into that question, let me first clarify the subject matter, 'technological knowledge'. Technological knowledge exists in many forms, varying from the intuition of engineers to detailed descriptions of the workings of machines. In this research, which is both theoretical and empirical, I focus on technological knowledge in the tangible form of *inventions*. I define invention rather generally as *a new practical application of some theoretical principle*. This 'theoretical principle' can be some discovered natural law, or geometric property, or even some knowledge of a social process: it is essentially a description (not scientific per se) of a particular relation between objects or processes¹. A 'new practical application' conveys that the theoretical principle is for the first time used to achieve some specified real-world purpose or fulfill some human need. I understand invention to differ from science in the sense that the latter focuses on developing new theoretical principles and is less (or not at all) focused on the application aspect. At this point it is also useful to distinguish between invention and innovation: an invention can *become* an innovation when it is implemented in society. Technological change, therefore, encompasses both invention and innovation. While acknowledging that the concept 'technology' includes multiple facets and can be approached from various angles (Mitcham, 1978), I will mainly focus on the knowledge aspects of technology in this work. In line with Arthur's perspective on technology (Arthur, 2009), we will therefore mainly approach 'technology' as the collection of all inventions, a subset of which (following some classification principle) can then be considered 'a technology'.

What does the development of technological knowledge require and under which circumstances does it flourish? Fundamentally, it requires motivated, well-trained

¹Note this does not require the phenomena involved with the theoretical principle to be 'fully understood': the understanding of a phenomenon may change with the development of later theories.

researchers and/or engineers, which in turn require stimulating environments, ambitious organizations, and intensive collaborations, which in turn require effective research and innovation (public) policies at regional, national, and international levels. All of these (and more) factors are relevant to knowledge development, and their effects are, both separately and collectively, studied as *Science, Technology and Innovation Systems* (Asheim et al., 2016; Binz & Truffer, 2017; Markard & Truffer, 2006). While acknowledging the relevance of these factors, I note that they mostly relate to, or are properties of, the producers of that knowledge (henceforth the 'inventors'). In this research, my aim is to focus more directly on the properties of knowledge itself (or 'knowledge intrinsic properties'). Approaching technology as a (growing) body of knowledge, I arrive at another fundamental requirement or source for the development of new technological knowledge: other (technological) knowledge. Philosophers of science and technology generally agree that new knowledge or content is created at least partly by recombining earlier knowledge (Arthur, 2009; Basalla, 1989; Freeman & Soete, 1997). To better understand the requirements for knowledge development to proceed, it therefore makes sense to study the knowledge linkages and dependencies between different bodies of knowledge. I identify two main different sources of knowledge on which technological knowledge may depend: scientific knowledge and other technological knowledge. In the following two sections I will discuss how the dependence on these two respective sources can be studied using the dimensions *science dependence* and *technological cumulativeness*.

While the inventors can theoretically be distinguished from the content they produce, in reality, they are two sides of the same coin. Properties of the inventors may therefore closely relate to knowledge intrinsic properties. In particular, the earlier mentioned science dependence and technological cumulativeness are expected to relate closely to knowledge mobility, a dimension that measures the extent to which knowledge travels geographically (discussed in more detail in Section 1.1.3). Where the first two dimensions are more content-related, the third is arguably less content-related and more inventor-related. The knowledge mobility, apart from being a useful indicator for the geography of innovation, can therefore be interpreted as a linking pin between on the one hand content-related dimensions and on the other hand inventor-related dimensions. As science and technology policies typically act in the domain of inventor dimensions, the knowledge mobility may be useful to translate findings from content-related dimensions into evidence-based science and technology policies.

To study the knowledge base of a technology, I therefore in the following three sections I discuss in more detail: (i) the science dependence, (ii) the cumulativeness and (iii) the mobility of technological knowledge.

1.1.1 The science dependence

The first knowledge dimension I focus on is the *science dependence*. I define the science dependence of a given technology as *the extent to which the development of this technology depends on scientific knowledge*. A substantial part of technology owes its working principles to the understanding of natural laws provided by science. In turn, science owes the possibility of many experiments and observations to the tools provided by technology. This symbiosis between science and technology (or to be precise physical science and industrialism) was described by A.J. Toynebee as "...

a pair of dancers, both of whom know their steps and have an ear for the rhythm of the music. If the partner who has been leading chooses to change parts and to follow instead, there is perhaps no reason to expect that he will dance less correctly." (Toynbee, 1963) Although this image is also criticized by authors who stress the development of the two is largely independent (Price, 1965a), the idea that there is some dynamic interaction appeals to many scholars, at least more than the 'linear model' does, in which all technology is applied science. Indeed there are in the history of science and technology plenty of examples where technology was crucial to science, as well as examples where the roles were reversed. The invention of the telescope essentially meant the birth of modern astronomy, yet without Newton's laws, we would still only be staring at the moon. The metaphor of a dance thus seems rather fitting: sometimes the partners are close together and touch, sometimes they are far apart and move rather differently to the same rhythm (be that more of a 21st-century dance than the elegant waltz Toynbee perhaps had in mind).

In the end, the suitability of Toynbee's metaphor probably depends on the way we demarcate science and technology, which may be challenging especially for disciplines on the borderline. Yet these grey areas can of course also be interpreted as instances of strong interaction (Narin et al., 1997), and it all boils down to the perception of technology-science dependence as a varying scale. In this work, I mainly focus on the dependence of technology on science, where I distinguish between the strength of the dependence, i.e. the earlier mentioned *science dependence*, and the body of scientific knowledge on which a technology builds, i.e. the *science base* of that technology. The type of knowledge (basic-applied) that forms the basis of inventions and the type of organizations (universities-companies) in which they are developed vary with the science dependence. Studying this dimension therefore not only reveals interesting differences between technologies, but may also provide policies with a lever to encourage the development of specific technologies.

Despite its relevance to science and technology policies, it is not always clear how the science dependence can systematically be measured (Meyer, 2000; Narin & Noma, 1985). A problem that arises when we consider various technologies, is how we can account for variations across these technologies in dimensions other than the science dependence, such as, for example, variation in age and the number of inventions associated with a technology. More generally, apart from a number of relevant examples (Leydesdorff & Zhou, 2007; McMillan et al., 2000), there is no clear-cut approach in the literature to what exactly a science base study should entail. Before I can therefore study the science base of RETs in all detail, I need to in the first place develop the methodology to identify, characterize and measure the science base of a technology.

1.1.2 The cumulateness of knowledge

The second knowledge dimension I focus on is the *cumulateness* of knowledge, which I define as *the continuous relevance of knowledge developed at any earlier stage to later knowledge development*. In other words, the idea that today's knowledge forms the basis for tomorrow's knowledge, which in its turn forms the basis for knowledge thereafter. I distinguish between technological and scientific cumulateness, where the former concerns the relevance of technological knowledge to technological knowledge, the latter concerns the relevance of scientific knowledge to

scientific knowledge. Cumulativeness (or 'cumulativity') comes about when people adapt their creations based on learning about previous creations or learning from other people and is often believed, at least in the context of science and technology, to be an important mechanism for progress (Dean et al., 2014; Richerson & Boyd, 2008; Tennie et al., 2009). Cumulativeness of knowledge enables people to reach accomplishments they would not have been able to reach by themselves, or, in the often-quoted words of Newton: *"If I have seen further it is by standing on the shoulders of Giants."* (Newton, 1675). For these reasons philosophers of science and technology regard cumulativeness as a fundamental property of both scientific and technological knowledge (Arthur, 2009; Basalla, 1989).

Next to its relevance in philosophy of science and technology, the cumulativeness concept plays a key role in the economics and the geography of innovation. Scholars have suggested that the technological cumulativeness varies characteristically across technologies in the extent to which they develop cumulatively (Malerba & Orsenigo, 1996; Nelson & Winter, 1977; Winter, 1984), and that this partly determines the ease (or difficulty) with which inventors or innovators may enter or diversify into a technology. Where an entry in higher cumulativeness technologies requires more effort and specialized knowledge, an entry in lower cumulativeness technologies is relatively easier. Furthermore, recent studies from the geography of innovations indicate that regions are more likely to diversify into technologies that are related to their existing knowledge base (Balland, 2016; Balland & Rigby, 2017; Boschma et al., 2015), and that greater cumulativeness associates positively with the geographical concentration of innovative activities (Breschi et al., 2000; Malerba, 2005). An understanding of the cumulative nature of technological development is thus pivotal for public policies looking to promote regional innovative activities, such as 'smart specialization' (Foray, 2014), where regions seek out attractive technologies for future economic development.

Despite its theoretical and practical relevance, it is not always clear in the scholarly literature what the cumulativeness concept exactly entails. Perspectives on cumulativeness vary from the incremental change in artifacts (Basalla, 1989; Butler, 2014; Gilfillan, 1935b; Ogburn, 1922), to the persistence of inventive activity (Cefis, 2003; Malerba & Orsenigo, 1993; Suárez, 2014), to the building of technological knowledge on earlier findings (Enquist et al., 2011; Merges & Nelson, 1994; Scotchmer, 1991; Trajtenberg et al., 1997). Furthermore, most of these descriptions or perspectives of cumulativeness are strictly conceptual, hence it often remains unclear how - if at all - cumulativeness can be measured systematically across technologies. The notable exceptions which do attempt to measure cumulativeness, mostly using the 'persistence perspective', often do not implement or check for the key characteristic of cumulative development that later knowledge builds on/depends on earlier knowledge. Research aiming to systematically compare cumulativeness across technologies would therefore benefit from a methodology that internalizes the property of knowledge building on earlier knowledge.

The elements of technological knowledge and the dependencies between them should be clearly represented in such a methodology, which leads naturally to a network approach to technological knowledge, where nodes represent inventions and links represent the dependencies between them. This 'knowledge network' approach has already been applied extensively in science and technology studies and, often following a basic network model introduced by Price (Price, 1976), has led to im-

portant insights about the mechanisms of scientific development (Garfield, 1979; Wang & Guan, 2011). In the field of innovation studies too the network approach is applied extensively, for example in the analysis of breakthrough innovation (Dahlin & Behrens, 2005; Fleming, 2001; Verhoeven et al., 2016), main paths (Hummon & Dereian, 1989; Verspagen, 2007), emerging technologies (Érdi et al., 2013; Shibata et al., 2009) and technological network evolution (Valverde et al., 2007). It appears however that this approach has not yet been applied for the analysis of technological cumulativeness. It is therefore not yet clear how exactly network structures might be used to measure technological cumulativeness.

Cumulative technological knowledge structures are characterized by inventions building on inventions, which themselves build on inventions, etc. An important element of these structures is therefore that there are sequences of inventions connected by knowledge flow, which, using a network approach, are conveniently represented by the well-studied notion of network paths and path length (Katzav et al., 2015; Newman, 2010; Watts & Strogatz, 1998). Yet there are various ways in which network paths can be applied methodologically. Where most contributions use metrics based on *the shortest paths*, that choice is not at all obvious for the study of cumulative structures, where it is important to take into account intermediate steps of development (which may not be included in the shortest paths). An alternative could be to use *the longest paths* instead, which necessarily includes the maximum number of intermediate steps. Another alternative might be to consider the various distinct paths leading to an invention, hence introducing metrics based on *all unique paths* in the network. This would indirectly account for the possibility that a combination of various ideas leads to a new invention, thus internalizing the recombinative nature of discovery and invention (Arthur, 2009; Kaplan & Vakili, 2015; Strumsky & Lobo, 2015). It is however not clear how the well-studied metrics based on the shortest paths (such as network distances and diameters) can be generalized to metrics based on the longest or all unique paths. In particular, it is unclear how, for typical knowledge network dynamics such as those in the earlier mentioned Price-model, the number of paths might be distributed over the various path lengths, i.e. what typical path lengths we might encounter in such networks. Understanding the path length distributions is therefore a starting point for a deeper understanding of cumulative knowledge structures and how these structures arise in different scientific disciplines or technologies.

1.1.3 The mobility of technological knowledge

The third knowledge dimension I focus on is the *knowledge mobility*, that is, the extent to which knowledge travels geographically. I define the knowledge mobility of a technology as *the extent to which the knowledge it builds on was created at a geographical distance from where it was utilized*. More mobile (or 'footloose') knowledge is developed widespread in ever-changing geographical locations, whereas less mobile (or 'sticky') knowledge, is confined to specific locations that change little over time. Unlike the science dependence and cumulativeness, which we can interpret as intrinsic knowledge properties more closely related to the particular content of that knowledge, the mobility of knowledge is a result of the actions of the users, intermediates, and producers of that knowledge, which are indirectly determined by intrinsic properties of that knowledge. In this work, I therefore mainly

investigate how the knowledge mobility is affected by the science dependence and cumulateness.

Various factors determine the location where technologies are developed, amongst others: market prospects, advantageous circumstances for production, or geographical aspects of their application (wind turbines where the wind blows, solar panels where the sun shines). While acknowledging that these factors likely affect the knowledge mobility, I focus in this work on the role that knowledge intrinsic properties may play in the geographic dynamics of technology development. Earlier contributions have connected the knowledge mobility to different modes of knowledge production, associating global, 'footloose' knowledge to a 'Science-Technology and Innovation mode' observed in science-based industries, and local, 'sticky' knowledge to a 'Doing, Using and Interacting mode' observed in engineering-based industries (Asheim & Coenen, 2005; Binz & Truffer, 2017; Jensen et al., 2007). This suggests that technological knowledge bases with a high science dependence are generally more geographically mobile. This is rather different for cumulateness, which, as I mentioned earlier, is associated positively with the geographical concentration of innovative activities. This suggests that highly cumulative knowledge bases are generally less mobile.

It is however unclear if these expectations also count for RETs. Some contributions indeed find that RETs are on average highly science-based and, in line with expectation, develop rather widespread (Ocampo-Corrales et al., 2020). However, other contributions indicate that RETs are a rather heterogeneous collection of various technologies (Barbieri et al., 2020), which themselves not only build on a heterogeneous set of other technologies (Noailly & Shestalova, 2013a), but are also built on by a heterogeneous set of other technologies (Nemet, 2012). What counts for RETs overall might therefore not count for the different individual RETs. To better understand if the expected relationship between the science dependence, cumulateness, and knowledge mobility is therefore applicable, the different RETs need to be considered individually, and the knowledge dimensions need to be determined and measured separately for each of them.

To recap, I list the most relevant research challenges in studying the knowledge base of RETs using the science dependence, cumulateness, and knowledge mobility.

1. It is unclear how the science base and science dependence can be systematically determined for various technologies. As such it is unclear what characterizes the science base of RETs and how that is different for FFETs.
2. Interpretations of technological cumulateness vary and it is unclear how cumulative knowledge structures can systematically be measured across various different technologies (not necessarily limited to RETs and FFETs).
3. In a network approach to technological knowledge, a convenient way to study cumulative knowledge structures is by using the notion of network paths and the metrics derived from this notion. It is however unclear how the commonly used metrics based on the shortest paths can be generalized for the longest or all unique paths, which are of special interest in the study of cumulative knowledge structures.
4. RETs are considered to be a heterogeneous collection of different technologies. It is unclear how the different RETs can be categorized with respect to the sci-

ence dependence and technological cumulateness, and how these dimensions relate to their knowledge mobility.

1.2 Research contribution

In this research I use Toynbee's science-technology dance and Newton's cumulative giants to characterize the knowledge base of Renewable Energy Technologies (RETs), providing input for evidence-based policies aiming to steer technology onto the renewable energy course. I thereby focus on the following research questions, which link one-to-one with the earlier mentioned research challenges:

1. What is the science base of RETs and how does it differ from FFETs?
2. How can we identify and measure technological cumulateness?
3. How can metrics based on network paths be generalized to study cumulative knowledge structures?
4. How do different RETs vary with respect to science dependence and cumulateness, and how does this relate to their knowledge mobility?

To address these questions, I develop a methodology to empirically characterize the science base of a technology and a methodology to measure technological cumulateness. I develop and apply these methodologies by explicitly approaching technological knowledge as a network structure of inventions and associated knowledge flows. This to some extent allows me to approach the above questions both theoretically and empirically, using data on patents and patent references.

In the following, I explain how my research is structured in more detail. Chapter 2 is dedicated mostly to the first of the above research questions. In this chapter, I develop a methodology that allows for a detailed analysis of the science base of technologies. I do this by identifying and classifying the references in patents (belonging to a certain technology) to scientific journals (belonging to certain scientific disciplines). I then apply the analysis both for RETs and FFETs, allowing me to study characteristic differences between both types of energy technologies. Next to the strength of the science dependence, I compare the science bases on a number of aspects, such as the diversity of scientific disciples, the degree to which they are 'basic' or 'applied', and the 'scientific impact'. The scientific impact of a scientific finding is a measure for its importance to other scientific findings and is in this research mainly approached using indicators that count how often a finding is referred to by other scientific findings (i.e. 'forward citations').

Chapter 3 is dedicated mostly to the second of the above research questions. In this chapter, I develop a methodology that allows for a detailed analysis of the cumulateness of technologies. Despite the recognized importance of cumulateness from various disciplinary angles, its exact meaning is not always clear in the literature. For that reason, I dedicate a substantial part of this work to clarifying the meaning of this concept in the context of technological knowledge. Subsequently, I apply a methodology in which I explicitly approach the knowledge base of technologies as a network of inventions connected through knowledge flow. The cumulateness of a technology is thus approached as property of the structure of its knowledge base,

which is different from, but closely related to, the size of its knowledge base. Using a simple model of the invention process I analytically derive equations describing the relation between the cumulateness and the size of the knowledge base. In addition, I empirically test these ideas for a number of selected technologies, using patent data.

Chapter 4 is dedicated mostly to the third research question, where I explore how in knowledge networks, the notion of network paths can be used to identify cumulative structures. Where the metrics based on the shortest paths are rather well developed, it is unclear how these can be generalized when we consider *the longest paths*, or *all unique paths* instead. In particular, I am interested in the question of how the paths are, for these cases, distributed over the various lengths. Knowing these distributions allows for a calculation of the expected path length, which represents the typical value of such path lengths in real-life networks. Starting from the Price model, I derive an exact solution for the distribution of all unique paths and investigate in detail how the different network properties affect this distribution. Deriving the distribution for the longest paths instead is analytically more challenging, which is why we approximate it instead, choosing the simplest possible network dynamics. Finally, I investigate how the expected path length depends on various network properties and the knowledge base size.

Chapter 5 is dedicated mostly to the fourth of the above research questions. In this chapter, the earlier developed methodologies are combined, where I determine for an extensive group of RETs both the science dependence and technological cumulateness, and systematically compare this to the knowledge mobility of their knowledge base. I estimate the degree of knowledge mobility using geographical data about patents, thereby bringing together various data sets. I recap my main findings in Chapter 6, including a discussion of implications for further research and policy.

Chapter 2

The science base of renewables

Peter Persoon, Rudi Bekkers and Floor Alkemade

This chapter is published as: Persoon, P.G.J., Bekkers, R.N.A., Alkemade, F. (2020). The science base of renewables. Technological Forecasting & Social Change 158, DOI: 10.1016/j.techfore.2020.120121.

Abstract

Initiatives to foster the development of Renewable Energy Technologies (RETs) can benefit from a deep understanding of the science base that underlies such technologies, and especially how that science base differs from that of Fossil Fuel based Energy Technologies (FFETs). This paper investigates both science bases using citations in patents to scientific journals. We find that RETs generally build stronger on science and draw on a more diverse set of scientific disciplines. On average, the science on which RETs build is more recent, less applied and is published in journals with a higher WOS Journal Impact Factor. However, for different RETs (e.g., photovoltaics, wind turbines and non-fossil fuels), we observe much more variation across these dimensions than for different FFETs (e.g., combustion and gas turbines). Furthermore, the broad spectrum of sciences on which RETs build largely includes the smaller spectrum on which FFETs build. Based on these findings, we offer several policy recommendations to better stimulate the development of RETs.

2.1 Introduction

Reducing carbon emissions is high on the policy agenda of many countries. Often such policies seek to stimulate the development Renewable Energy Technologies (RETs) through subsidies and tax reductions. While these demand stimulation measures work well to create a level-playing field for RETs that are (almost) market ready, the influence on the development of novel and immature technologies is less clear. Moving from general stimulation of science and R&D to more targeted policies that stimulate knowledge development specifically for RETs, while ceasing support to Fossil Fuel based Energy Technologies (FFETs), is not trivial. Incentives for such mission-oriented research (Mazzucato, 2016) require a deep understanding of the technologies, their underlying scientific knowledge base, and the interaction between science and technology.

The interaction between science and technology is understood to be pivotal for technological development (Freeman & Soete, 1997; Rosenberg, 1976), yet the nature of this interaction may be very different for each field of technology (Mansfield, 1995; Verbeek et al., n.d.). Possible benefits of science, such as accelerating or improving the efficiency of technology development, are found to depend on the extent to which a technology consists of highly coupled components (Fleming & Sorenson, 2004) or highly novel combinations of components (Arts & Fleming, 2018). Recent findings indeed suggest that green technologies combine a higher number of technological components and are based on more novel combinations than their non-green counterparts (Barbieri et al., 2020), and that energy technology generally builds on a large and diverse set of other technologies (Nemet, 2012; Noailly & Shestalova, 2013b; OECD, 2010). This suggests that RETs may benefit greatly from developments in science, and more so than FFETs. In this study we therefore evaluate how strongly RETs build on science, how the RETs science base can be characterized and how this differs for FFETs.

In an earlier analysis of the science base of environmental technology (including RETs) for the period 2000-2007, the OECD found a broad dependence on scientific disciplines (OECD, 2010). With the OECD study as a starting point, this paper seeks to carve out the specifics of the RET knowledge base by answering the question: What is the scientific knowledge base of renewable energy technology, and how does that differ from non-renewable energy technology? To do so, we present an in-depth investigation, looking at various dimensions of the science bases, such as the relative importance of basic versus applied research, and the scientific impact of that research. We use recent data, and also investigate time trends, as several studies have suggested that policy for RETs should also take into account the development stage of the technology (Abernathy & Utterback, 1978; Anderson & Tushman, 1990; Huenteler, Schmidt, et al., 2016). Our analysis of these different aspects of the science base of RETs can help policymakers to make informed and targeted decisions.

The paper is structured as follows. In Section 2.2 we discuss the theoretical background of determining the science base of a technology and what may be expected for RETs and FFETs. Next, Section 2.3 explains how we identify RETs and FFETs and the methodology used for measuring technology-science links. We do so by constructing these science bases using patent data and references from patents to the non-patent literature. Section 2.4 describes the relevant science bases and science base quantities. We will explicitly address the distinction between RETs,

FFETs and overlapping technologies.

2.2 Theoretical aspects of science bases

In line with Arthur's perspective on technology (Arthur, 2009), we define technology as the *body of knowledge* which applies *science and/or engineering* to fulfill a human purpose. In evolutionary economics, technological change is often understood in terms of technological paradigms and trajectories (Dosi, 1982). A technological paradigm is an outlook on technological progress, kept by a certain engineering community and based on a particular collection of scientific and/or technological findings. A technological trajectory is the continuous development of a particular product or technological design within such paradigm. Important scientific discoveries largely set the playground for changing paradigms. Any attempt to better understand a given technology, therefore, should start with a deep understanding of its science base. Even though the respective knowledge bodies of science and technology may overlap, their differences and mutual interaction are relevant topics in innovation sciences. Some scholars thereby emphasize the importance of science to the development of technology (Freeman & Soete, 1997) while others emphasize the reverse relation (Rosenberg, 1976). Yet, most authors seem to agree that the influences in both directions are essential, resulting in a complicated, non-linear pattern of interaction. However, the extent to which technological innovation builds on scientific knowledge varies greatly across economic sectors (Pavitt, 1984) and technologies (Mansfield, 1995; Verbeek et al., n.d.). Fleming et al. approach inventing as a combinatorial search process, in which science may lead inventors more directly to useful combinations (Fleming & Sorenson, 2004). Their findings suggest that especially R&D in technologies that combine a large number of interrelated 'coupled' components may greatly benefit from the guiding role of scientific knowledge. Furthermore, the usage of scientific literature may help inventors overcome the uncertainties of exploring new fields and trying novel combinations of knowledge (Arts & Fleming, 2018). A recent study by Barbieri and co-authors focuses on the *technological* knowledge base differences between green and non-green technologies and indeed suggests that green technologies combine more technological components and are based on more unique combinations of knowledge than their non-green counterparts (Barbieri et al., 2020). We therefore expect that RETs may benefit more from scientific knowledge than FFETs, and hence to a larger extent build on scientific knowledge.

Energy technology consists of a diverse set of technologies, each extensively building on a larger collection of other technologies (Barbieri et al., 2020; Nemet, 2012; Noailly & Shestalova, 2013b). In general, we therefore expect a large disciplinary diversity for the science base of energy technology. This broad dependence may, however, be different for RETs and FFETs. The 2010 OECD report mentioned above indicates that green technology depends on a very broad spectrum of scientific fields. While technological knowledge is not the same as scientific knowledge, the earlier mentioned study by Barbieri suggests that the green technologies rely on more diverse technological knowledge than their green counterparts. Next to the technologies built on, this difference is also found for the technologies building on RETs and FFETs: RETs were found to have more spillovers, and to a greater vari-

ety of technological fields than FFETs (Barbieri et al., 2020; Dechezleprêtre et al., 2014). Dechezleprêtre et al. found that the spillover rates of RETs are comparable to upcoming tech-science fields such as biotechnology, nanotechnology, robotics, and 3D-printing. Finally, yet perhaps most importantly, RETs and FFETs differ greatly in terms of exploited phenomena. Where FFETs revolve around exothermic chemical processes (such as combustion), RETs reside to a multitude of fundamentally different phenomena, such as wind and sunlight but also, in the case of bio-fuels, on traditional exothermic chemical processes. Taking all these aspects into account, we expect a greater diversity in the science base of RETs than in the science base of FFETs.

In the industrialized countries, energy generation in the 20th century was largely fossil fuel based, leading to a strong development of FFET (Wilson, 2012). Even though early developments of RETs started over a hundred years ago, most RETs are generally considered to be in an earlier stage of development than FFETs. In terms of market readiness, RETs vary considerably (McKinsey, 2013), for instance wind is in a more mature phase than photovoltaics (Popp, 2017). Technologies in an early phase of development may also have a more diverse underlying science base. Where for technologies in a more mature phase a dominant design may be established, building on a (small) number of relevant scientific fields, for technologies in more early phase a number of designs may still be competing, each exploring ideas from several fields (Anderson & Tushman, 1990; Murmann & Frenken, 2005). The differences in distance to market may thus correspond to differences in the extent and diversity of the underlying science base.

To further understand science base differences between RETs and FFETs, it is useful to distinguish between radical innovation (understood to initiate new paradigms) and incremental innovation (understood to happen within technological paradigms). Innovation in FFETs is largely understood to happen within current energy technology paradigms, and can therefore mostly but not exclusively be associated with incremental innovation (Markard et al., 2012; Markard & Truffer, 2006). Likewise, RETs can mostly but not exclusively be associated with radical innovation. In the context of energy technology, incremental innovation is linked to large technological systems, characterized by slow innovation and path-dependent development. Radical innovation is linked to fast developments in niches (Markard & Truffer, 2006). Radical innovations are understood to be based on novel knowledge (combinations) of both technological and scientific nature (Verhoeven et al., 2016). Even though paradigm change may not be synchronized for science and technology, breakthroughs in science may lead to breakthroughs in technology (and vice versa). Given that scientific breakthroughs are characterized by high-impact factors, we expect radical innovation to build on high-impact science more than incremental innovation. Technology may incrementally co-develop with a certain field of science, and specialize along their trajectories, in tandem. Such specializing science may only be relevant to a limited number of neighboring fields of science and may therefore have a lower impact. That could be another reason why incremental innovation may be associated with lower impact science. Given this radical-incremental distinction, we therefore have three expectations: (1) for RETs we expect a larger science base than for FFETs, (2) the impact of the science RETs build on is expected to be higher than that for FFETs and (3) we expect a smaller science technology time lag for RETs than for FFETs.

2.3 Data and methods

2.3.1 Knowledge base definitions

We define the *knowledge base* of a technology (or technological field) as the body of knowledge on which it builds. In section 2.2, technology itself was identified as a body of knowledge as well. In this analysis, we assume that we can meaningfully attribute a *size* to both types of knowledge bodies. That allows us to define the *knowledge dependence* of a technology as the size of the knowledge base relative to the size of the technology. This measures to what extent a technology builds on earlier knowledge. The knowledge base may be of a technological, scientific and/or other type of knowledge¹. We define the *science base* of a given technology (or technological field) as the scientific part of the knowledge base. Accordingly, the *science dependence* is the size of the science base relative to the size of the technology. The *technology base* of a technology (or technological field) is the technological part of the knowledge base. Within the technology base, a part of the knowledge may refer to the technology itself (e.g. to earlier versions or ideas which are part of the same technology). The size of this part of the technology base relative to the size of the technology we define as the *intra-technology dependence*. It indicates to what extent a technology builds on itself rather than on other technologies. A high intra-technology dependence of a technology may indicate a more mature phase of a technological development, as in the beginning a technology mainly builds on other technologies (and/or science). The size of the other part of the technology base, we define as the *inter-technology dependence*.

Using a classification of scientific disciplines, the science base can be studied in more detail. The distribution of the science base over these disciplines also allows for *diversity* comparisons between technologies. Another perspective on science distinguishes between *basic and applied* science. Both are about acquiring new knowledge, yet where the first is directed primarily towards the underlying foundation of phenomena and observable facts, the second is primarily directed towards a specific, practical aim or objective (OECD, 2015). Views on what is basic and applied research vary however (Calvert & Martin, 2001). Still, most researchers would agree, as our definition of technology implies, that applied research is closer to practical application and technology than basic research. Arguably, a "technology" or "engineering" classification of a research field therefore signals a more applied character of research. To effectively stimulate the development of a technology, it is important to understand to what extent it builds on basic or applied science.

Finally, we consider the time scales of the science-technology interaction. Knowledge of typical time scales is vital to accelerate technological development, and the planning of policies. We define the *science-technology time lag* as the average time lag between the publication of the scientific knowledge and the usage of this knowledge by technology. This lag signifies how fast knowledge flows, and indicates the temporal proximity of the science and technology interaction.

We use patents to study the interactions between science and technology. A technology is represented by a body of patents applications (henceforth shortly 'patents') within a defined technological class. Patent data is arguably the most extensive and detailed source of technological developments. Yet, using patent data to represent

¹In this analysis we will only consider scientific or technological knowledge.

a technology also has its limitations. Not all technology that is developed, is or can be patented, and not all patents lead to successful technology. In this research, we consider patents filed at the European Patent Office (EPO) between 1977 and 2016, and use data from the Patstat Spring 2017 version.² Patents usually contain references, which are added by the patent examiner to identify the relevant body of prior art, in order to decide whether the application meets the patentability criteria. Such references are also known as backward citations. For EPO patents, references are exclusively determined by the examiner, even though the inventor may suggest relevant prior art. For innovation scholars, these references are a useful tool to retrieve the possible building blocks of a technology. Most references are to other (existing) patents, but they can also refer to other knowledge sources, generally known as Non-Patent Literature (NPL). Most NPL references are to scientific literature, mostly journal articles (Callaert et al., 2006; van Vianen et al., 1990).

The total number of NPL references in patents for a given technology is a useful indicator for the extent to which that technology builds on science: For example, using NPL references, Narin found a general trend of increasing dependence on science at the end of the 20th century (Narin et al., 1997). Additionally, NPL references can also provide insight into the scientific content exploited by the technology, for instance by looking at the scientific discipline of these references. NPL data, however, should be used with caution. For instance, increasing numbers of NPL references could also be the result of improved search mechanisms at the patent office (instead of reflecting an increased reliance on science). Moreover, NPL data may be incomplete (not exhaustive) or may not always be relevant (Meyer, 2000). The larger the dataset, the more such imperfections can be expected to level out.

Below, we will first discuss how we identified the relevant patent data sets for RETs and FFETs, and second, how we processed and cleaned the NPL data in those data sets. Third, we explain how we determined science base quantities from that data.

2.3.2 RETs and FFETs patent classes

To identify Renewable Energy Technologies (RETs), we use the Y02 classification scheme introduced in the Cooperative Patent Classification (CPC). The Y02 tag signals patents that enable or stimulate climate change mitigation (Veefkind et al., 2012). The Y02 classification further distinguishes a number of subclasses. For our study, we focus on patents that have at least one classification in the Y02E subclass, which contains technologies related to energy generation, transmission, or distribution. When we refer in this work to 'all RETs' or 'RETs aggregated' we refer to unique patents in this subclass. Y02E is further divided into a number of 'groups', a selection of which we will focus on in this research. We base this selection on the main categories of RETs as proposed by the International Renewable Energy Agency (IRENA) (IRENA, 2018): hydropower, bioenergy, solar energy, wind energy, geothermal energy and tide- wave- and ocean energy (also called 'energy from sea'). In addition to that, we include a number of relevant enabling technologies, which can complement RETs to become feasible alternatives to FFETs: energy storage, hydrogen energy, fuel cells, and smart grids. Table 2.1 provides an overview of the

²We chose EPO patents because of the relatively high quality NPL data for such patents in Patstat

Short term	CPC code	Nr of EPO patents	NPL of journal-type	Average earliest filing year
<i>RETs</i>	Y02E	69,904	40,421	2006
<i>Geothermal energy</i>	Y02E 10/1	544	20	2005
<i>Hydro energy</i>	Y02E 10/2	2,040	80	2006
<i>Energy from sea</i>	Y02E 10/3	1,308	109	2006
<i>Solar thermal energy</i>	Y02E 10/4	6,219	503	2005
<i>Photovoltaic energy</i>	Y02E 10/5	16,589	14,951	2007
<i>Wind energy</i>	Y02E 10/7	10,882	1,286	2008
<i>Enabling technology</i>	Y02E 60	18,110	6,397	2006
<i>Energy storage</i>	Y02E 60/1	8,396	2,548	2006
<i>Hydrogen technology</i>	Y02E 60/3	4,167	1,797	2005
<i>Fuel cells</i>	Y02E 60/5	3,882	1,508	2005
<i>Smart grids</i>	Y02E 60/7	1,522	496	2007
<i>Non-fossil fuels</i>	Y02E 50	7,156	16,430	2006
<i>Clean combustion</i>	Y02E 20	5,330	1,451	2004

Table 2.1: *RETs and CPC descriptions (CPC, 2018)*

RET technologies we consider individually in our study, as well as their CPC codes (Column 2), and the total number of patents for those technologies (Column 3). Column 4 shows the number of NPL references to (academic) journals³. We will use these NPL references - and specifically their academic disciplines - to investigate the science base of these technologies. (Note that three technologies - energy from sea, geothermal and hydro energy - have too few NPL observations to allow detailed analysis.) The last column shows the average (earliest) filing year, and here we can see that all these technologies are relatively young (between 2005-2008).

To identify Fossil Fuel based Energy Technologies (FFETs), we will largely follow the list of CPC subclasses constructed by Dechezleprêtre (Dechezleprêtre et al., 2014). These patent classes are presented in Table 2.2, along with relevant characteristics. When we refer in this work to 'all FFETs' or 'FFETs aggregated', we consider each unique patent with (at least one) classification mentioned in of Table 2.2. Note that most FFETs have an average (earliest) filing year between 2001-2002, which indicates that FFETs are, on average, older technologies than RETs.

Remarkably, many of the patents in the FFET (sub)classes also have a Y02 tag, for example, the inventions that aim to reduce carbon gas emission of fossil fuel based technologies. These inventions both have a 'clean' and 'dirty' element to them, and we can call them 'hybrid' technologies. In order to carefully deal with such hybrids in our analysis, we perform our analysis for both the total set of FFETs, as well as for a set of FFETs excluding these hybrids. Dechezleprêtre copes with this challenge by separately considering "clean", "grey" and "dirty" technologies on the group level of classification. Our exclusion, however, is on the more precise level of individual patents.

Patents also often list multiple CPC codes, and as a result, many patents are both present in multiple RET and/or multiple FFET technologies. In fact, the 140,874 unique patents in our data set include a total of 374,731 CPC classifications. Figure 2.2 illustrates which technologies often co-occur in these classifications. The connections between RETs show a lower density than those between FFETs. We

³Patstat offers a specific classification of NPL types, and references to (academic) journals are recognized as the 's'-type in Patstat)

Short term (<i>italics</i>) and extended description	CPC code	Nr of EPO patents	NPL of journal-type	Average earliest filing year
<i>Cracking</i> : cracking hydrocarbon oils; production of liquid hydrocarbon mixtures, recovery of hydrocarbon oils from oil-shale, oil-sand, or gases; refining mixtures mainly consisting of hydrocarbons; reforming of naphtha; mineral waxes	C10G	12,712	3,703	2001
<i>Gasification</i> : production of producer gas, water-gas, synthesis gas from solid carbonaceous material, carbureting air or other gases	C10J	2,661	294	2002
<i>Fuels</i> : fuels not otherwise provided for; natural gas; liquefied petroleum gas; adding materials to fuels or fires to reduce smoke or undesirable deposits or to facilitate soot removal; firelighters	C10L	8,165	3,305	2002
<i>Steam engines</i> : steam engine plants, steam accumulators; engine plants not otherwise provided for, engines using special working fluids or cycles	F01K	4,432	414	2006
<i>Gas turbines</i> : Gas-turbine plants, air intakes for jet-propulsion plants, controlling fuel supply in air-breathing jet propulsion plants	F02C	10,305	476	2007
<i>Steam generation</i>	F22	3,779	272	2001
<i>Combustion</i> apparatus, combustion processes	F23	20,632	853	2001
<i>Furnaces</i> , kilns, ovens, retorts	F27	8,154	722	2001
<i>Heat exchange</i> in general	F28	20,547	1,099	2002

Table 2.2: *FFETs and CPC descriptions (from CPC descriptions 2018 (CPC, 2018))*

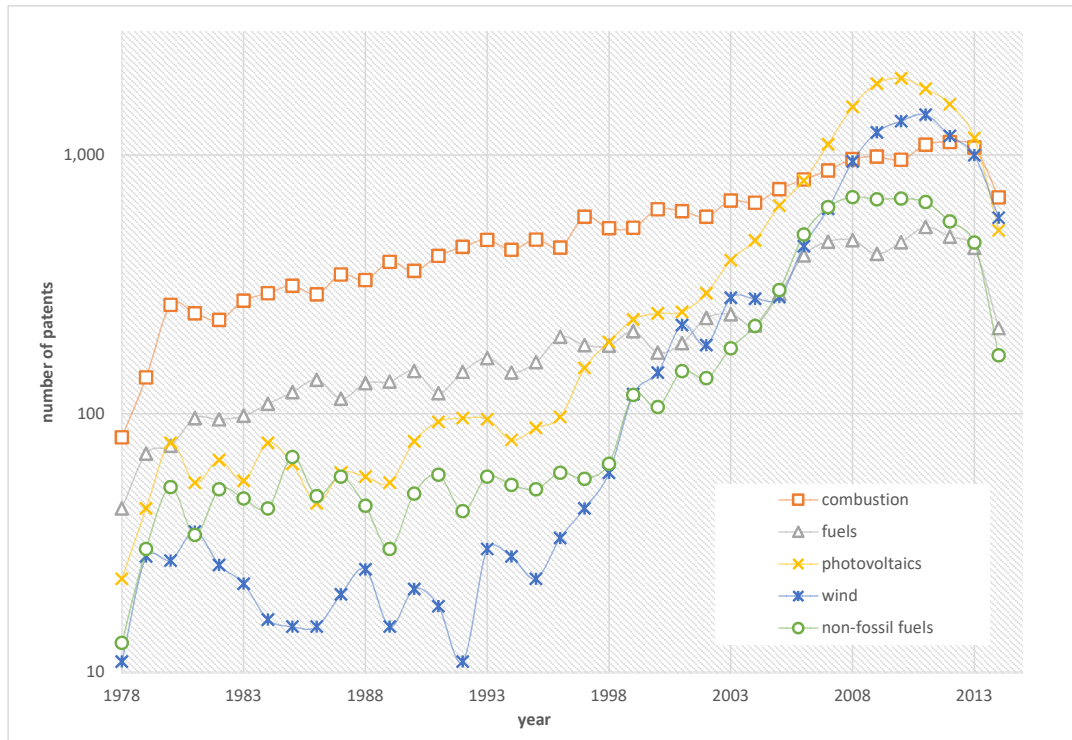


Figure 2.1: *Number of European patent applications by earliest filing year for selected RETs and FFETs* The number of patents increase exponentially for each depicted technology. The decrease towards 2015 is likely an truncation effect (see main text).

also observe that a number of RETs have more co-classifications with FFETs than with other RETs. Focusing on patents that are only classified in RETs and patents only classified in FFETs (and not considering patents classified in both), we see that these RET patents have an average of 1.3 CPC classifications, whereas the average for these FFETs is 3.3. This suggests that specific RETs (e.g., hydro and photovoltaics) are relatively unconnected, while FFETs are more connected.

Finally, we focus on the time dimension of the considered EPO patents. As explained above, our data set starts with patents from 1977 (one year after EPO was established). Figure 2.1 shows the number of applications by year, for the largest technologies considered in this research. A drop in patents can be observed in the last few years of our dataset: this is most likely not an actual decline, but due to the fact that it takes up to 18 months before a filed patent is published (which is necessary to be included in Patstat), and the update cycle of Patstat. The (few) patents in our data set applied for in 2016 are considered in our overall analyses, but not shown separately.

Figure 2.1 shows that in the early time frame there are fewer patents for RETs than for FFETs. We see that, after approximately 2001, the number of patents for RETs starts to grow very fast (note the logarithmic scale). Both observations are consistent with the earlier finding of RETs being a younger technology (Table 2.1 vs. Table 2.2).

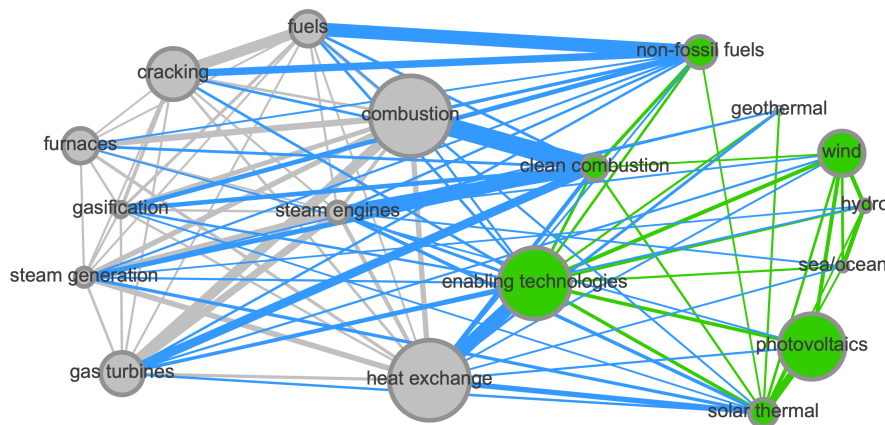


Figure 2.2: *Co-classification between RETs and FFETs* Node size corresponds to total number of patents. Line thickness corresponds to the number of patents co-classified in these technologies (green = RET-RET, grey = FFET-FFET, blue=FFET-RET)

2.3.3 Linking NPL data with WOS journal entries

In order to calculate the indicators discussed in section 2.3, we link the NPL journal references with journal entries in the Web Of Science (WOS) database. As a significant share of NPL data in Patstat is far from harmonized (e.g., a single journal may be referred to in many different ways), this also requires extensive cleaning of NPL data. While this section will only briefly explain the main steps we took, Appendix A provides a detailed explanation.

Our data set contained a total of 68,042 NPL references. We first considered the references that contained an ISSN (about 13,000). We used these to create a *matching list* that contained the full names of those journals as used in the WOS as well as other common abbreviations and alternative names used for that journal. To do so, we also used the online database of the ISSN International Centre (“ISSN international Centre”, n.d.). The matching list was further complemented with terms from the Science and Engineering Journal Abbreviations list from the University of British Columbia (Library, n.d.).

We then used this matching list to find matches for the NPL references that did not contain an ISSN code (approximately 55,000). To that end, we applied two different string matching algorithms (Van der Loo, 2014). Where the two algorithms provided the same best match (which was in 72 percent of all cases), the match was accepted and, if not, it was discarded.

Finally, for each technology, we determined the frequency distribution of cited journals, and downloaded the WOS Journal Impact Factor, Eigenfactor and WOS Category for the most frequently cited journals (where available). For all technologies together, this amounted to 1577 journals, representing 82 percent of all matches.

2.3.4 Science base characteristics, distributions and indicators

We calculate the *knowledge dependence* of a technology as the number of references per patent (to other patents and to NPL). Similarly, the *science dependence* of a given technology is the number of NPL references per patent and the *intra-technology dependence* of the technology is the number of references to patents in the same class per patent. The *science dependency by year* are calculated for patents applied for in a given year. The *science-technology time lags* are determined, for each technology, as the average time difference between the publication date of the NPL and the earliest filing date of the patent that cites it.

To determine the science base distribution over different fields of science, and calculate the values of related indicators we use the classification system for scientific journals by the Web Of Science (WOS). This classification distinguishes 252 'smaller categories' (ClarivateAnalytics, n.d.-b). We determine the science base distribution over these categories by counting the total number of NPL references to journals in these categories (note that our data set refers to journals in 140 of these categories). This method is similar to the approaches used by McMillan (McMillan et al., 2000) and by Leydesdorff (Leydesdorff & Zhou, 2007), who used it to study biotechnology and nanotechnology, respectively. Where a journal was classified in multiple categories, it was counted fractionally. To compare the diversities of different science bases, we calculate the Shannon index of the relative distributions.

The WOS also offers a more aggregated level of classification called 'broader categories', containing 'Life Sciences-Biomedicine', 'Physical Sciences', 'Social Sciences', 'Arts-Humanities' and 'Technology'. We use the 'Technology' broader category to distinguish between a *basic* and an *applied* science base. For every technology (RETs or FFETs), we calculate the *A-fraction* as the fraction of NPL references to journals in the 'Technology' broad category. We note that patents in our data set do not - or very rarely - refer to the broader categories 'Arts-Humanities' and 'Social Sciences'.⁴

Finally, WOS also offers data that allows us to calculate the average *Journal Impact Factor (JIF)* associated with the science base of a given technology. The JIF is a measure of the frequency with which the "average article" in a journal has been cited in a particular year (ClarivateAnalytics, n.d.-a), and is considered to reflect the scientific relevance of contributions published in that journal. This view however is also criticized, see Waltman, 2016 and references therein. Amongst other limitations, the JIF may not account for variation in citation densities across scientific fields. For robustness, we therefore (i) check for the JIF variations we may expect for citing journals in a certain scientific field and (ii) repeat the analysis for a second measure of scientific relevance which WOS offers, namely the normalized eigenfactor (West et al., 2010). While the normalized eigenfactor is also based on forward citations, the NEF also takes into account the impact of the source of the citations and excludes the effect of self-citations. As the NEF is based on proportions of citations rather than absolute numbers of citations, it is relatively insensitive to

⁴The WOS broad categories 'Physical Sciences' and 'Life Sciences-Biomedicine' may include 'smaller categories' which can reasonably be characterized as applied research as well. Therefore, we also constructed an alternative indicator, for which we separately considered each 'smaller category' that is part of these two broader categories, and characterized them manually as 'applied' or 'basic'. The results for this alternative indicator were mostly identical to that of the A-fraction, so we present the results for the latter and simpler indicator only.

variations in citation densities across scientific fields (West et al., 2010).

2.4 Results

This section consists of two parts. First, we discuss the overall knowledge base characteristics of RETs and FFETs. In the second part, we investigate the actual nature/content of these references to scientific sources, and determine their academic impact, diversity, their most relevant science categories and the extent to which they are basic or applied.

2.4.1 Overall knowledge base characteristics

Figures 2.3 and 2.4 present, for RETs and FFETs respectively, the *knowledge dependence*. For each technology, we break down the knowledge dependence (the total number of references per patent) into (1) the *science dependence* (references to NPL), (2) the *intra-technology dependence* (references to patents in the same technology) and (3) the *inter-technology dependence* (references to patents in another technology). For a substantial part of the RETs, the relatively high knowledge intensities are mainly the result of high science dependencies. We observe that the knowledge dependence varies more across different RETs, than it does across different FFETs. If we only focus on science dependence, we make the same observation (see Figure 2.5). Overall, RETs have higher science dependencies, especially photovoltaics and non-fossil fuels: for RETs aggregated, the science dependence represents about 22 percent of the total knowledge dependence. For FFETs aggregated, this is only about 8 percent, and even decreases to 7 percent when the Y02 patents are taken out. The intra-technology dependence also varies a lot more across different RETs than across the FFETs. Even though some RETs have high intra-technology dependence, the FFETs generally show higher intra-technology dependencies. Insofar individual FFETs build on other technologies, these are usually other FFET technologies. In summary, (i) RETs vary more than FFETs in the extent to which they build on previous knowledge, (ii) RETs build stronger on science than FFETs, (iii) RETs build less on themselves than FFETs. Figure 2.5 combines the different results indicating the ranges of the different science dependencies. The range is far wider for the RETs, mainly due to non-fossil fuels and photovoltaics. Taking out the Y02 patents from the FFETs, the FFETs range becomes even smaller.

Next, we consider the RETs in Figure 2.3 in more detail. The knowledge dependence varies considerable across different RETs with a standard deviation of 1.0 ref/pat. The knowledge dependence of non-fossil fuels (6.9 ref/pat) is the absolute maximum and that of energy from sea (2.9 ref/pat) the minimum, with other technologies mostly on the lower side of this spectrum. The science dependencies appear to be proportional to the knowledge intensities (standard deviation 0.8 ref/pat). Geothermal and solar thermal are exceptions in this respect: they show a relatively low science dependence of 0.21 ref/pat, while their knowledge intensities are relatively high. This appears to be largely due to relatively high intra-technology dependencies. The fact that these technologies build more on themselves than on external (scientific) knowledge may indicate a more mature phase of development for these technologies. The science dependence therefore indicates key differences

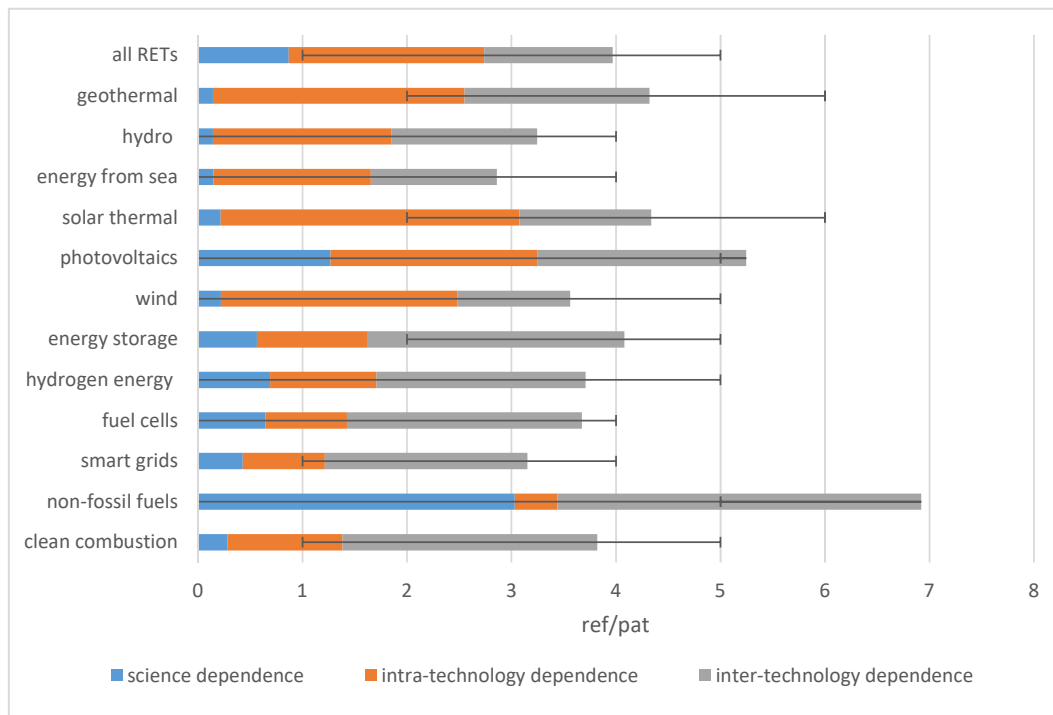


Figure 2.3: **RETs knowledge dependence** Break down of the knowledge dependence of RETs in science dependence, intra- and inter-technology dependence. The shown error bounds are defined by the 35th percentile and 65th percentile of the knowledge dependence. Note that for photovoltaics and non-fossil fuels the mean can be observed to exceed the 65 percentile.

between the RETs. We have on the one hand strong science dependencies for photovoltaics, non-fossil fuels and most enabling technologies. For wind, geothermal, solar thermal, hydro and energy from the sea, science appears less important and the intra-technology dependencies are relatively high. Photovoltaics however is exceptional in this demarcation, which despite a high science dependence also has a relatively high intra-technology dependence. This seems to indicate that although the technology is maturing, science is still important to its development.

Subsequently, we turn to the FFETs, shown in Figure 2.4. We first discuss the FFETs in general, which includes patents with a Y02 tag, and then the FFETs without Y02 patents. There is little variation in the knowledge dependence across different FFETs (standard deviation of 0.4 ref/patent). Only fuels seem to have a relatively high knowledge dependence, which appears to be largely due to a relatively high science dependence and inter-technology dependence.

Overall, the science dependence and intra-technology dependence are remarkably uniform for different FFETs, with respective standard deviations of 0.1 ref/pat and 0.6 ref/pat. Interesting to note is that the knowledge dependence of all FFETs together consists for almost 90 percent of intra-technology dependence. The inter-technology dependencies for the individual FFETs can therefore be understood to almost exclusively refer to other FFET technologies.

Further, we consider the FFETs in Figure 2.4 without the Y02 patents. Most knowledge intensities are similar to their versions including Y02 patents. The science dependencies however appear to slightly decrease for cracking and fuels. The intra-technology dependencies of these technologies instead appear to increase, which also

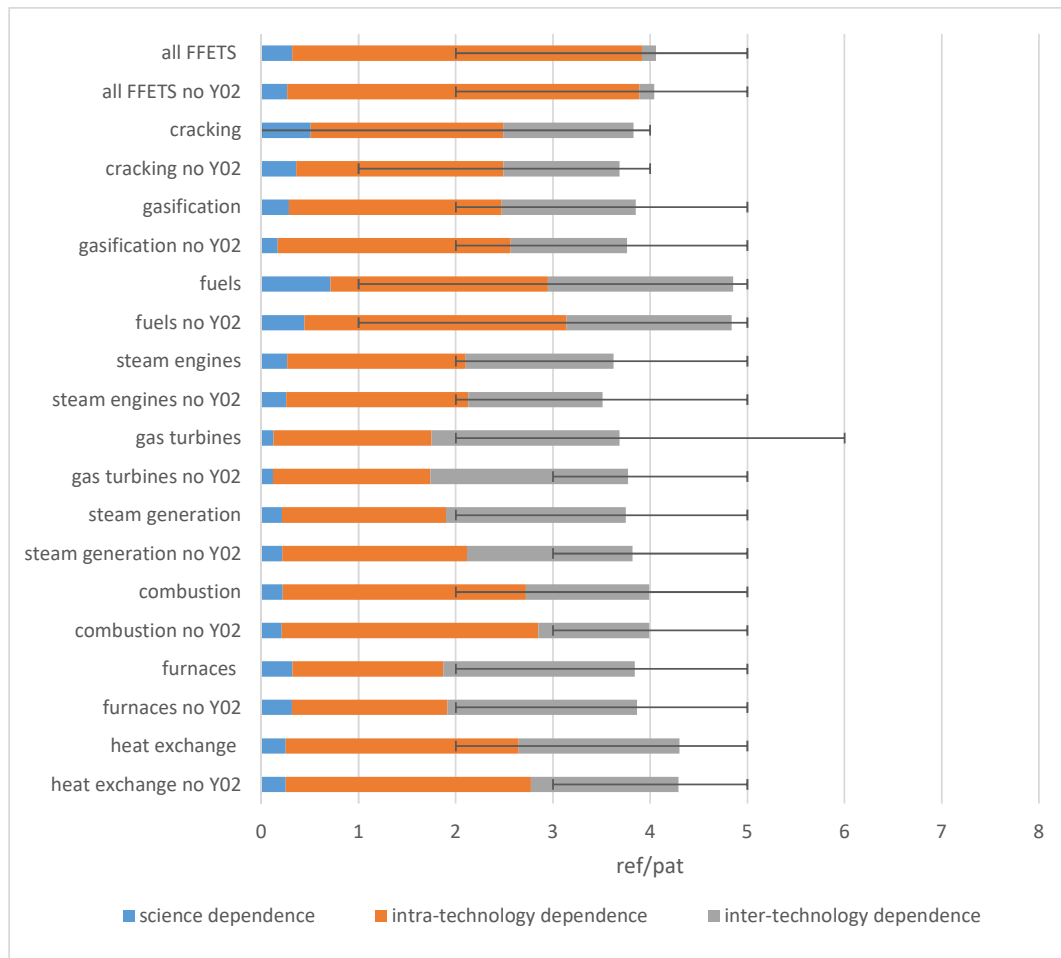


Figure 2.4: *FFETs knowledge dependence* Similar to caption Figure 2.4, but then for FFETs. We also distinguish between FFETs and FFETs without Y02

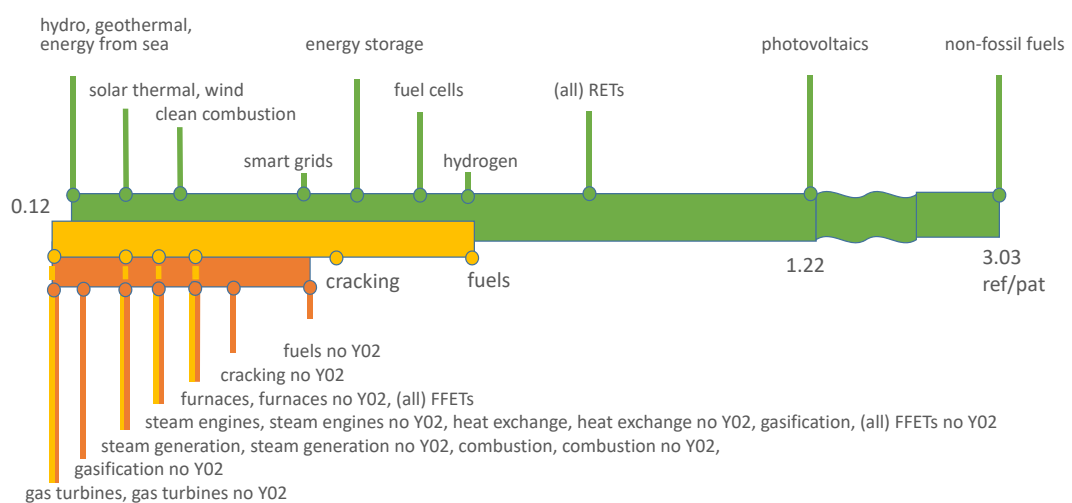


Figure 2.5: *Schematic representation of science dependence range.* Here, we show the range in science dependence for the RETs (green), FFETs (orange) and FFETs without the Y02 patents (red).

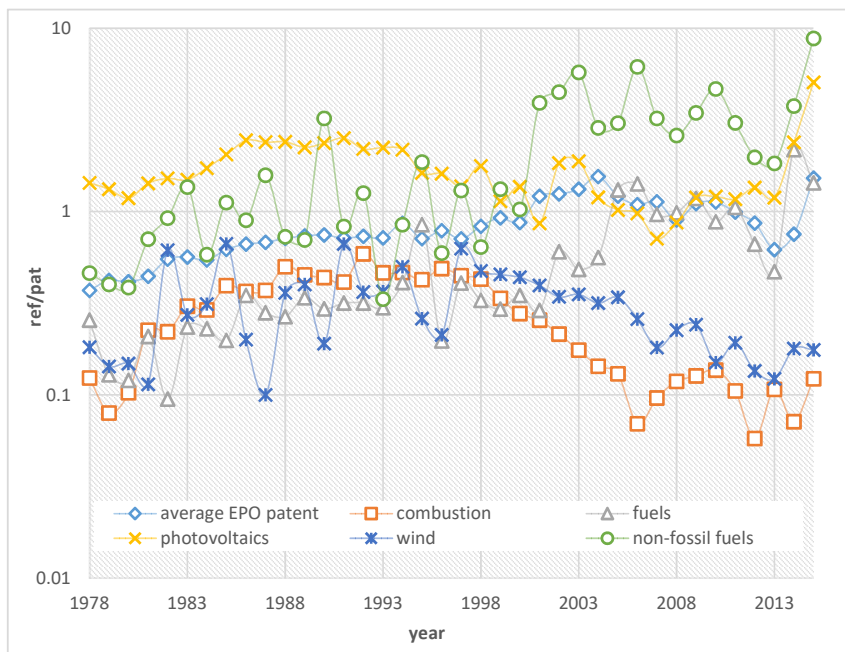


Figure 2.6: *science dependence by year*. Yearly science dependencies for a number of RETs, FFETs and average EPO patents.

holds for a number of other FFETs, such as combustion and heat exchange.

Finally, we discuss the time dimension of the overall science base characteristics of the technologies under investigation. From tables 2.1 and 2.2, we have seen that the average earliest filing years are somewhat older for FFETs than RETs. We have also seen that all RETs and FFETs from the early beginning show continuous growth (e.g., have a non-zero and increasing number of patents over the period 1978-2015). The propensity to cite science, however, does not follow a similar regularity. In Figure 2.6, which depicts the science dependence by year, we observe considerable variation across different technologies. Where the science dependence of non-fossil fuels and fuels seem to continuously increase, that of wind and combustion seem to first increase and then decrease. The science dependence of photovoltaics follows a pattern which is somewhat in between those patterns. If we study the science dependence by year for the other technologies, we find similar variation both across RETs as FFETs. The importance of science may therefore not just be technology-specific, but also be specific to the phase of development of that technology.

We conclude this subsection by considering the science-technology time lags. In Figure 2.7 we present the time lags of both the FFETs and RETs along with the number of NPL references. For the FFETs, we plot the time lag both on the CPC classification group level as on the CPC (sub)class level. Though we are mainly interested in the (sub)class level, we include the values of the subgroup level to demonstrate a general relation between the time lags and the number of NPL references. It appears that the higher the number of NPL references, the less variation in time lags, resulting in a cone shape with a horizontal axis at the height of about 10 years. The FFETs on (sub)class level neatly fall within the cone, their time lags vary only little between 8.8 and 12.5 years. These lags do not significantly change once the Y02 patents are removed from the FFETs. A number of RETs,

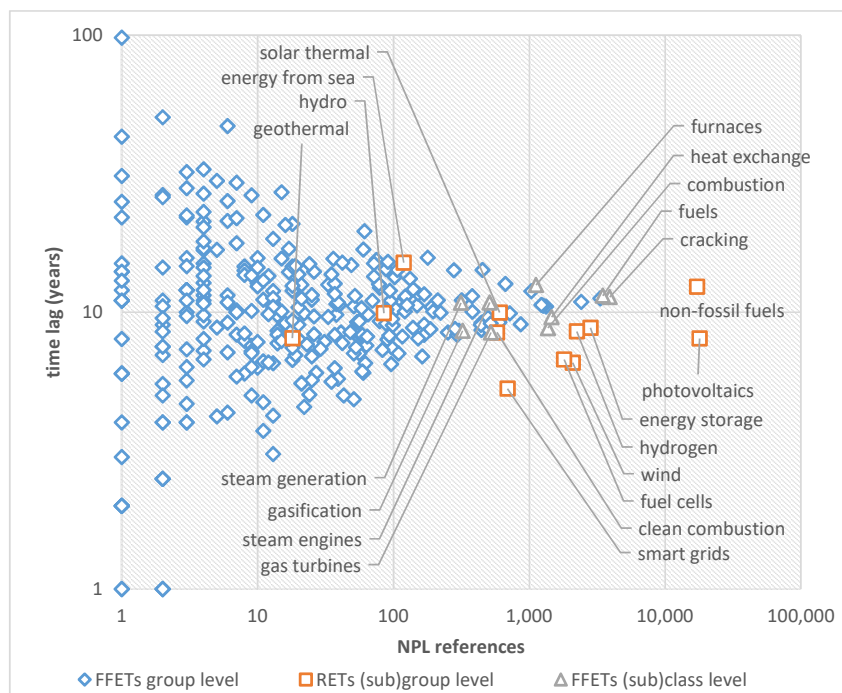


Figure 2.7: **Science Technology Time Lags.** Time lags (in years) for each RET and FFET ordered by number of NPL references.

however, appear to fall outside the cone to the lower side, their time lags ranging between 5.3 and 12.4 years, despite rather high numbers of NPL references. For fuel cells, wind and smart grids, the time lags are particularly small, indicating a rapid interaction between science and technology in these fields. Overall, given that the RETs generally have more NPL references than FFETs, the RETs time lags seem relatively shorter than those of FFETs. Again, however, we observe larger heterogeneity for RETs than for FFETs.

2.4.2 Science base distributions and indicators

The following section discusses the actual nature/content of the references to scientific sources for the various technologies⁵. We determine the academic impact of these sources, their diversity, and their most relevant science categories and level of applied versus basic science.

To determine the academic impact of these sources, we determine the average WOS Journal Impact Factor (JIF) of references in each technology and the technologies aggregated, (see Table 2.B.1 in appendix 2.B). For RETs aggregated the average JIF is 7.57 (standard deviation 10.3), for FFETs aggregated it is 5.36 (standard deviation 7.9), for FFETs without Y02 patents this drops to 4.64 (standard deviation 6.9). A simple Welch t-test points out that these averages are significantly different. The majority of RETs we consider individually also have higher average JIFs than the FFET maximum (6.30 for fuels). To account for the JIF bias towards

⁵Note that the relatively small number of NPL references of the RETs hydro, energy from the sea and geothermal energy, (see Table 2.1) did not allow for these more detailed analyses and will therefore not be considered individually in this section

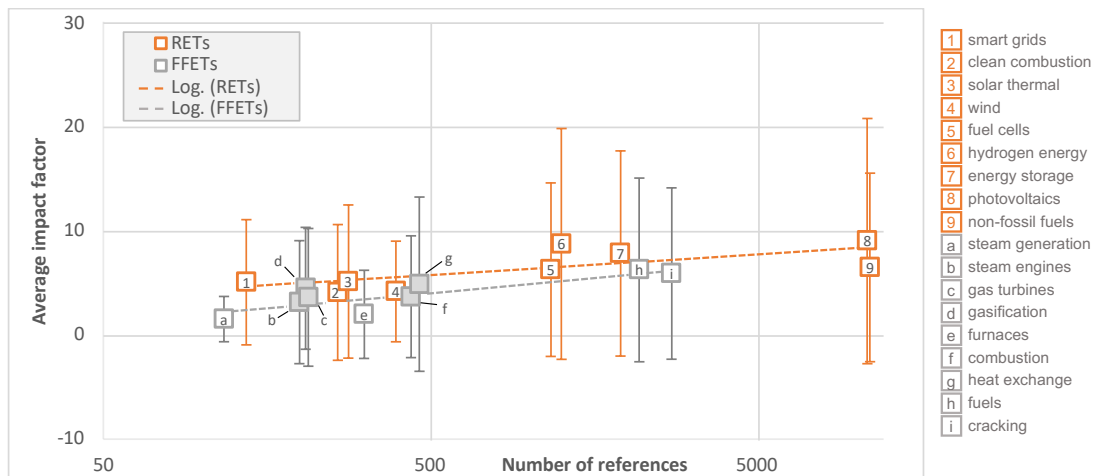


Figure 2.8: **Average JIF RETs and FFETs with Y02** We plot for all RETs and FFETs with Y02 patents the average JIF for the number of references, where the grey bars indicate the standard deviations and the lines indicate a logarithmic fit. Note that the positive relation between the JIF and log number of references equally counts for RETs and FFETs and the coefficients hardly differ.

certain scientific categories, we calculate the expected JIF based on the distribution of references over the scientific categories for each technology and the technologies aggregated (see appendix 2.B). These expected values indicate that the average JIF is indeed expected to be slightly larger for RETs than for FFETs. However, when we consider the proportion between these expected values with the (measured) average JIF per technology, we find these proportions are substantially larger for RETs than for FFETs. This suggests that, within a scientific category, RETs tend to build on sources with a scientific impact which is relatively high for that category, and more so than FFETs.

These conclusions however are more nuanced when we consider how the average JIF relates to the number of references, which we depict for the different RETs and FFETs in Figure 2.8 and similarly for FFETs without Y02 in Figure 2.9. We find that both for the RETs and FFETs, there is a significant, positive relation between the logarithm of the number references and the average JIF, the coefficient of which does not significantly differ for the RETs and FFETs (for the statistical details see appendix 2.B). We find this relation considering the average JIF on the level of technologies and to some extent also on the level of individual patents. While the coefficients are significant, most constants (i.e. the value of the JIF when the number of references =1) are not. This therefore does not allow us to conclude that the RETs have higher values than FFETs for a similar number of references, even though Figures 2.8,2.9 suggest this. The fact that the aggregated RETs refer to sources with a higher average scientific impact than the aggregated FFETs therefore appears to be closely related to the greater tendency of RETs to refer to scientific literature. Finally, as a robustness check, we repeat a large part of this analysis for the normalized eigenfactor, the results of which are presented in appendix 2.B. These results are very much in line with the earlier conclusions based on the JIF.

Considering now the diversity of the science base, as measured with the Shannon entropy (Figure 2.10), we come to a rather different conclusion. Note that for the FFETs, we plot the entropy both on the CPC classification group level as on the CPC (sub)class level. Though we are mainly interested in the (sub)class level, we

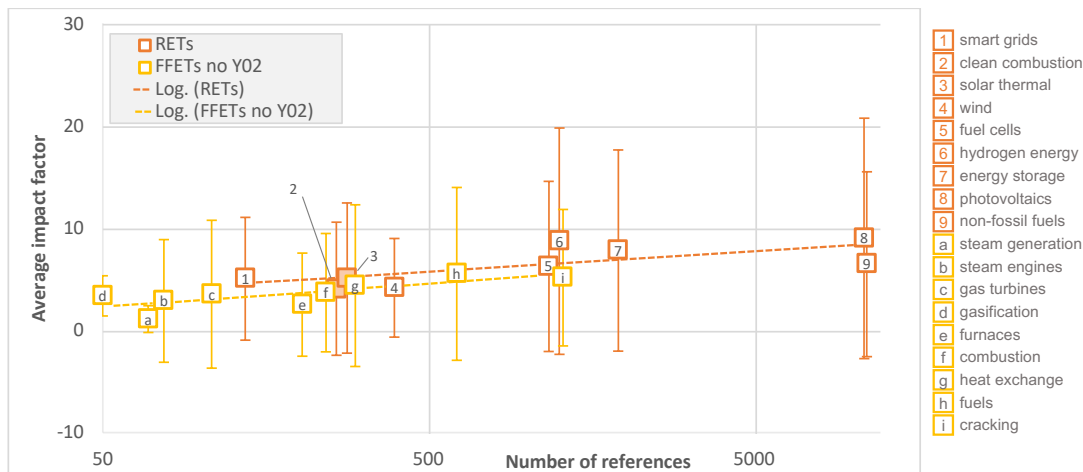


Figure 2.9: *Average JIF RETs and FFETs without Y02* Same as Figure 2.8 except for the FFETs without Y02 patents. Both the JIF and number of references are lower when the Y02 patents are removed.

include the values of the subgroup level as they better demonstrate a positive relation between the number of NPL references and the Shannon entropy. While we again observe a positive relation between the number of NPL references and the Shannon entropy, RETs are rather on the lower end of the spectrum, compared to FFETs with similar numbers of NPL references. This conclusion does not change if the Y02 patents are taken out from the FFETs (sub)classes in Figure 2.11. The RETs aggregated have a Shannon entropy of 3.3, for the FFETs, this is 3.4 and when the Y02 patents are removed this reduces to 3.3. If we instead take the Herfindahl index as a measure of diversity, it leads to similar conclusions. These measures of diversity, though attractive for their simplicity, do not account for variations in mutual similarity or dissimilarity relations between research categories. To account for these relations, we will therefore also take a more qualitative perspective by considering the most important research categories in the science bases.

The most relevant science categories in the science base distributions for both the RETs and FFETs are illustrated schematically in Figures 2.12 and 2.13. The pies represent the most important science categories, where the size corresponds to the total number of references to these categories. In Figure 2.12 the green part of the pies represents the number of references from RETs, the red part those from FFETs. The vertical coordinate of each pie indicates the number of different RETs which build on it for 5% or more, the horizontal coordinate indicates the number of different FFETs which build on the category for 5% or more.

From Figure 2.12 it is immediately clear how much stronger the RETs build on science: almost all pies have a larger degree of green. The only considerably large category showing the contrary is 'chemical engineering'. The categories 'physical chemistry', 'multidisciplinary chemistry' and 'energy and fuels' are important for FFETs, yet are positioned high and hence important to multiple RETs. In other words, there appear to be few scientific categories only relevant to FFETs, and those that are, are only sparsely referred to. RETs, however, have a number of categories on which they uniquely build, such as 'polymer science', 'electrochemistry' and 'organic chemistry'. These findings become even stronger once the Y02 patents are

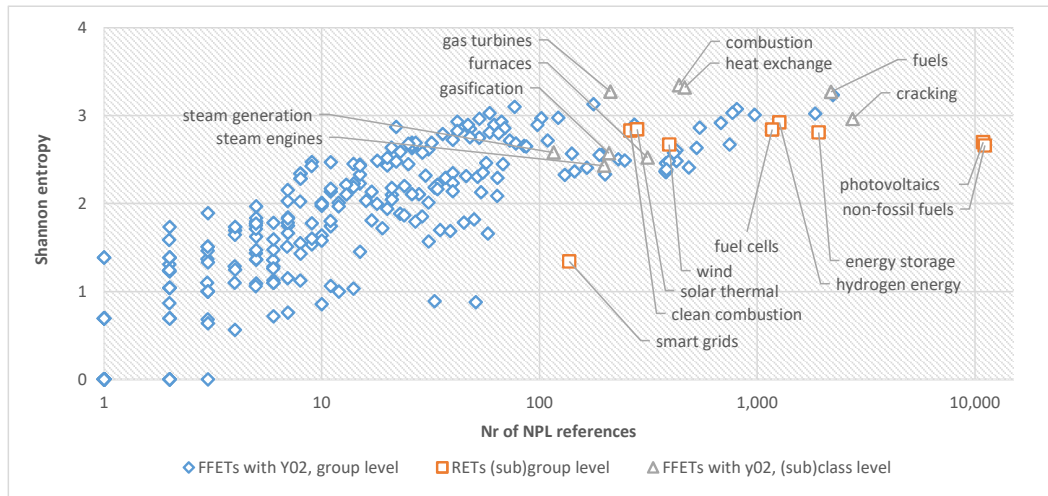


Figure 2.10: *Shannon entropy of the science base of each technology* The Shannon entropy of the science base of the RETs (orange), FFETs on the group level (blue) including Y02 patents and FFETs on subclass level including Y02 patents (grey).

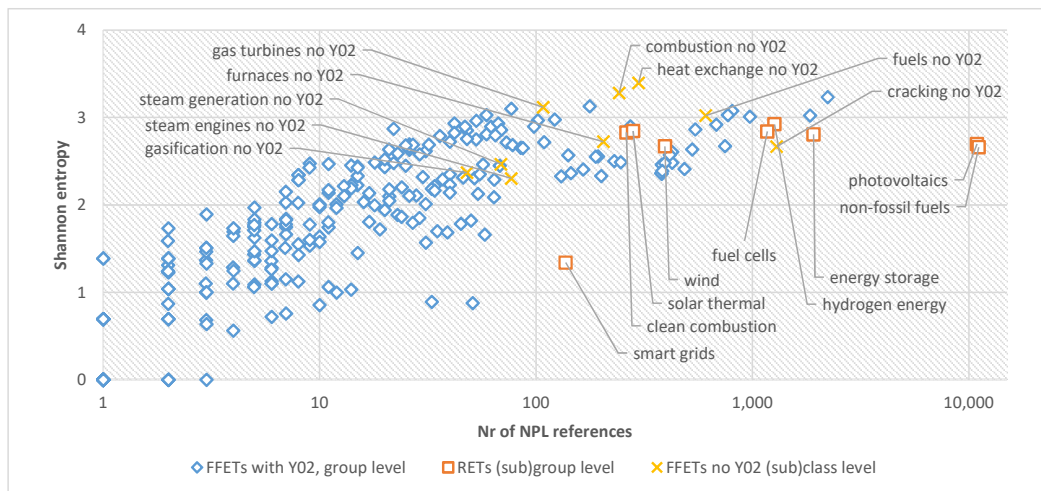


Figure 2.11: *Shannon entropy of the science base of each technology, FFETs on subclass level without Y02*

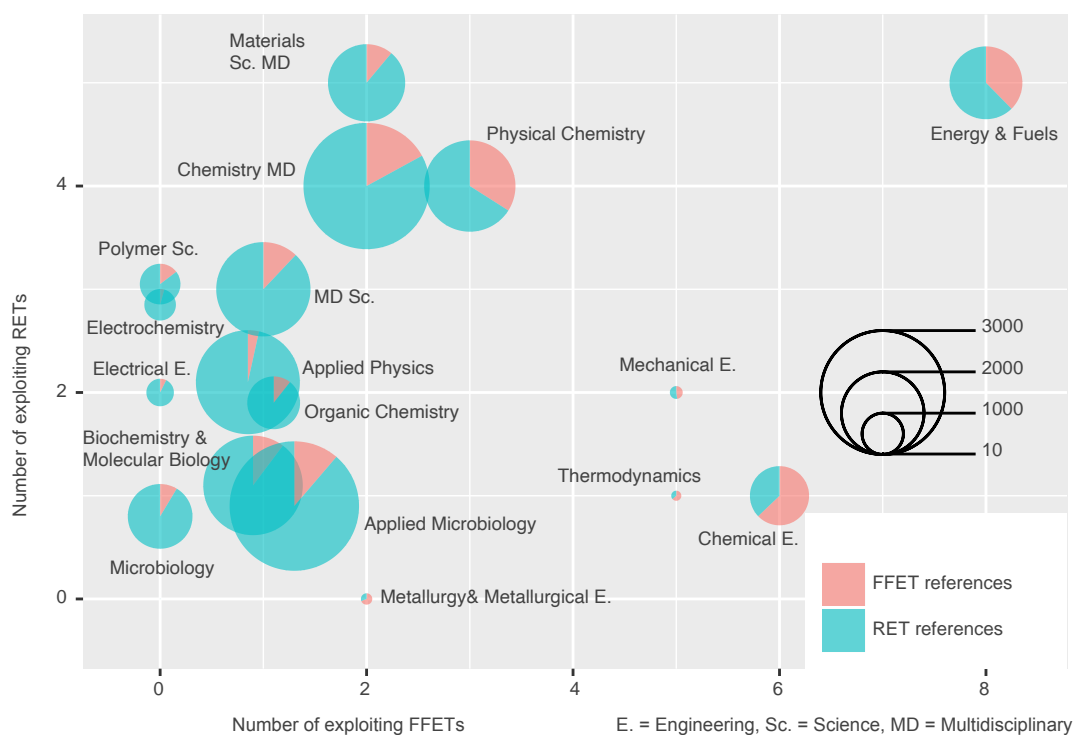


Figure 2.12: **Most important WOS Research Categories.** Each bubble represents a WOS research category which makes up at least 5% of the science base of at least one of the RETs (without hydro, geothermal, and sea energy) and/or FFETs. The size of the bubble indicates the total number of references from all technologies to journals in these categories, where the proportion of references coming from RETs is depicted in green and from FFETs is depicted in red. The vertical coordinate of a bubble indicates the number of different RETs which build on it for 5% or more, the horizontal coordinate indicates the number of different FFETs which build on the category for 5% or more. This figure was created using the R package *scatterpie* (Yu, 2018).

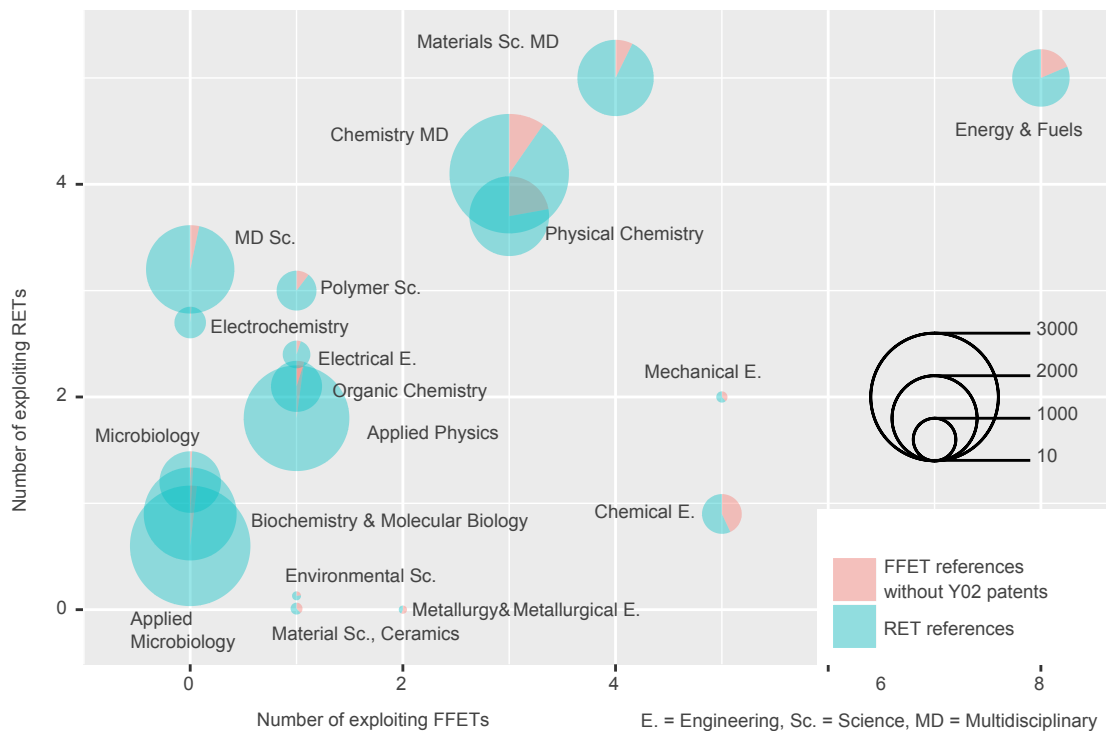


Figure 2.13: *Most important WOS Research Categories (Y02 removed)*. Similar to Figure 2.12, but now the FFETs are without Y02 patents.

taken out of the FFETs (Figure 2.13): almost all pies are now on the higher left. The considerable importance to FFETs of some large categories, such as 'biotechnology' and 'microbiology', visible in Figure 2.12, disappears in Figure 2.13. This is likely the effect of in- or excluding bio-fuels in FFETs. Figures 2.12 and 2.13 offer a rather different perspective on the diversity of the science base than the Shannon entropy. RETs strongly refer to various fields of physics, chemistry, and biology as well as a number of interdisciplinary and engineering fields. References in FFETs are mostly limited to fields of chemistry, such as chemical engineering, multidisciplinary- and physical chemistry. Therefore, from a disciplinary perspective, the RET science base is a lot more diverse than the FFET science base.

In conclusion, we observe an asymmetry in the extent to which RETs and FFETs build on scientific categories. RETs build on most categories more strongly and have various large categories uniquely useful to them. Furthermore, the relevant categories RETs build on, cover a large spectrum of scientific disciplines. FFETs, in turn, weakly build on most categories. The few categories they strongly build on are mostly related to fields of chemistry only.

Finally, we discuss the A-fraction, which indicates the extent to which a science base consists of applied science. Here, we find no relation at all between the number of NPL references in the science base and the A-fraction. Yet, as Table 2.3 shows, this indicator does signal important differences between the science bases of RETs and FFETs. In general, the A-fractions of the RETs are lower than those of FFETs, which suggests that RETs build less on applied science than FFETs. However, the A-fractions vary more across different RETs, with very high values for wind and smart grids (0.92) and very low values for non-fossil fuels (0.30) and photovoltaics (0.22). For most FFETs, the A-fraction increases when the Y02 patents are removed.

RETs	A-fraction	FFETs	A-fraction	A-fraction no Y02
all RETs	0.20	all FFETs	0.40	0.46
solar thermal	0.54	cracking	0.32	0.35
photovoltaics	0.22	gasification	0.59	0.67
wind	0.84	fuels	0.29	0.38
energy storage	0.27	steam engines	0.71	0.75
hydrogen	0.30	gas turbines	0.65	0.69
fuel cells	0.29	steam generation	0.73	0.82
smart grids	0.92	combustion	0.63	0.66
non-fossil fuels	0.07	furnaces	0.79	0.74
clean combustion	0.63	heat exchange	0.56	0.55

Table 2.3: *A-fraction for RETs, FFETs and FFETs without the Y02 patents. A high A-fraction (between 0-1) signals a high level of applied science.*

2.5 Discussion

There is inherent uncertainty about understanding future technological developments on the basis of a snapshot of those technologies today. Identified characteristics of those technologies, such as the extent to which it builds on science, may or may not change throughout the further development of such technologies. The indicators, while useful, therefore have to be used with caution and in context. Further, we discuss a number of limitations on a more technical level. Firstly, we base our study on non-patent references in patents. While this allows us to observe the influence of science on technology, it does not allow us to see the reverse influence. Secondly, we rely on several classifications for our analyses; the CPC Y02 coding by the EPO and USPTO to identify RETs, and the coding by Dechezleprêtre to identify FFETs. We analyzed RETs at the class or subclass level and FFETs at the group or subgroup level. In addition, we build on existing classifications of scientific areas into disciplines, and into basic and applied science. While these classifications have been validated, their use to study emerging technologies is not trivial. Finally, we focused in this research on EPO patents. Repeating this research for USPTO patents would be interesting, as citation behavior may be very different in those patents (Criscuolo & Verspagen, 2008). The greater tendency of US patent applicants to cite is likely to intensify the science dependences and broaden the spectra of scientific disciplines in the science bases found in this work.

2.6 Conclusions and policy recommendations

This paper compares the science bases of Renewable Energy Technologies (RETs) and Fossil Fuel based Energy Technologies (FFETs), driven by the assertion that such an understanding is helpful in formulating policies to stimulate the development of RETs. Our empirical analysis shows that:

1. RETs build on science stronger than FFETs. The aggregated RETs on average refer 3 times more to scientific literature than the aggregated FFETs.
2. RETs build on a larger and more diverse set of relevant scientific disciplines than FFETs. Where RETs build on very different WOS research categories

such as applied physics, chemistry, materials science, and various fields of biology, FFETs build on more closely related categories, such as chemical engineering and other fields of chemistry. Furthermore, the broad spectrum of sciences on which RETs build includes the smaller spectrum on which FFETs build.

3. For RETs, the time lags between scientific publication and usage in a technological invention are generally shorter than for FFETs.
4. RETs generally build on higher impact science than FFETs. The average Journal Impact Factor of journals referred to is 1.5 times greater for the aggregated RETs than for the aggregated FFETs.
5. RETs generally build more on basic research, whereas FFETs rely more on applied research. The fraction of applied research fields in the science base of RETs is on average 2 times less than for FFETs.

In our data, we distinguish RETs and FFETs on the basis of patent technology classification (CPC) codes. A complicating factor here is that some technologies can be classified both as RETs and FFETs. For the above conclusions (1-5), these technologies were included with the FFETs. If we exclude these 'hybrid' technologies from our definition of FFETs, all five conclusions become more pronounced.

Yet, our empirical findings also point out a number of subtleties. Across the dimensions in all five conclusions, we find much more variety between different RETs than we find for different FFETs, which are more homogeneous. RETs such as photovoltaics and non-fossil fuels intensively refer to scientific literature, and rely on very basic, high-impact science. Other RETs such as wind- and geothermal energy refer less often to scientific literature, and, in the case of wind energy, more to applied science. In fact, the backward citation rates of these last two technologies are more comparable to FFETs. As we will discuss below, this heterogeneity of RETs does have consequences for policy. Furthermore, we note that the average scientific impact of the sources referred to by a technology is related logarithmically to the number of references of such a technology. The greater reliance on high-impact science of RETs as opposed to FFETs is therefore closely related to the greater tendency of RETs to refer to scientific literature.

On the basis of our findings, we formulate a number of policy recommendations for stimulating research and development for RETs. This study has shown that RETs rely much more on science than FFETs (and especially on basic, high-impact science). As such, a policy that promotes scientific research in general (and basic, high-impact science in particular) is expected to lead to a strengthening of RETs. Further, this study has shown that RETs rely on a broad spectrum of scientific disciplines, a spectrum that encompasses the smaller spectrum FFETs rely on. While this study has identified a number of scientific disciplines particularly relevant to a number of FFETs, the relevance of these disciplines to RETs is at least comparable. It appears therefore that reducing support for particular scientific disciplines is not likely to be a successful policy for fostering RETs rather than FFETs, or an accelerated phasing out of FFETs. At the same time, our study has shown considerable heterogeneity across different RETs: photovoltaics and non-fossil fuels are rather different in their science base from wind- and geothermal energy, for instance. For policies to be as effective as possible, policymakers are advised to take

these differences into account, and develop technology-specific policies that focus on strengthening fields in science that are known to promote specific renewable technologies. Such technology-specific policies should consider to what degree specific RETs depend on science (some do more so than others, as we have shown), and the spectrum of scientific disciplines that would qualify for strengthening, given the choice for a specific RET. Such policies could also consider the time lag between science and application (which is much shorter for some RETs than for others) and the degree to which RETs build on applied science (instead of basic science). Taking the above considerations into account, smart grids and wind energy are among the most suitable candidates for technology-specific policies with short-time goals (investments in related applied science), whereas non-fossil fuels and photovoltaics are the main candidates for technology-specific policies with a longer time window (investment in the related basic science).

Appendix

2.A Linking NPL data with WOS journal entries

Classifying a large number of NPL references can be a non-trivial task.⁶ The Patstat Spring 2017 version is advantageous in this respect, as it is the first to have a more harmonized recording of NPL references.⁷ The classes in Table 2.2 plus the Y02E patents amount to a total of 68,042 NPL references, where we only consider `npl_type='s'`, which indicates journal citations. For about 13,000 references the ISSN of the journal is recorded. As a preliminary step, we did a check on patents with large number of NPL references. We recorded examples with more than 300 NPL references, and some of these patents appeared to occur multiple times (with sequential application numbers and almost identical references). Applying for a patent several times, each time with only minor variations, is strategy used by some patentees, yet could bias our research if the number of NPL references is high. We therefore performed a quick scan of all patents with numbers of NPL with 30 or higher to erase 'duplicates' of these patents, which were 94 in total. Then we were ready to further clean and sort the data in a number of systematic steps.

- First, these ISSN references were selected and their journal names were looked up in WOS and if not available in an online ISSN database ("ISSN international Centre", [n.d.](#)). The result was the *ISSN journal list*.
- Second, to classify the remaining about 55,000 NPL references, which translated to 37,849 unique references, the journal name data under the Patstat header '`npl_title2`' was used. As this is still a substantial number, the matching with journal names was automatized using two string comparison algorithms. The first, "Optimal String Alignment" (OSA), minimizes the Levenshtein distance and at the same time allows for transposition of adjacent characters. The second, "Jaro distance" (JARO), is a more heuristic distance measure which matches characters between two strings that are not a given number of positions apart (Van der Loo, 2014).
- Third, the string matching algorithms need a list with journal names to compare the raw data to. We created this list combining *ISSN journal list* with the Science and Engineering Journal Abbreviations list from the University of British Columbia (Library, [n.d.](#)), which includes approximately 13100 journal names and their abbreviations. After a number of trials, the list was further

⁶For the linkage of NPL references in patents to scientific articles see also the database by Marx and Fuegi, 2019 (which was not yet available at the time of this research)

⁷This recording was in May 2017 still work in progress, yet was surprisingly complete for the EP patents.

improved by including new terms on the basis of frequently occurring mismatches. This resulted in a reference list of 6042 search terms, where search terms resulting in no matching were discarded.

- Fourth, in the final automated comparison of the `npl_title2` data and the match list, we selected only the item with the best matching score for both algorithms. On the basis of this we created two lists: the first with matched journal names on which both the OSA and JARO algorithms agreed, which happened for the majority of the cases (72%), and a second where they disagreed (28%).
- Fifth, we did a manual test on a random 500 piece sample of the first list, which revealed that for 7 percent of the references the matching was inaccurate. A manual test on a random 500 piece sample of the second list revealed that for 11 percent of the references the OSA method was correct, in 13 percent of the references the JARO method was correct. This test gave further confidence to use only the matched terms on the first list and discard the other. A second quick inspection of the discarded list revealed that for a substantial part (estimated 14 percent) of these references it was hard to identify a journal in the first place.
- Sixth, given the large number of items on the match list, we were required to make a selection of search terms for which further journal data would be acquired. To this end the following steps were taken. (a) We ordered the matched journal names/abbreviations by frequency of occurrence and selected the top 500. (b) We made a separate frequency ordering for the smaller RETs and FFETs and from these the top 350 terms were selected. (c) We complemented this selection with the journal names from the *ISSN journal list* and removed any doubles, which completed the shortlist of a total of 2041 journal names and/or abbreviations.
- Seventh, we searched for all terms on the shortlist in Web Of Science (WOS). For 464 of the journal names there was no data available in WOS, reducing the shortlist to 1577 search terms. Fortunately, the journal names not available in WOS were not frequently occurring in the references. The shortlist was sufficient to cover 82% of all identified NPL references and attained coverage percentages well over 55% for all individual RETs and FFETs separately. The only exceptions to this were wind energy and smart grids with respective coverage of 40% and 32%. However, a more careful analysis of the matched search terms of wind energy, including the ISSN matching, showed that 35% of the wind energy references was in the group of 464 journal names about which WOS has no data. Similarly, for smart grids this was a 26%. These percentages are relatively high: for photovoltaics it was only 6% for example. The low coverage percentages of wind and smart grids were therefore mainly the result of frequently referring to journals which were/are not represented in WOS at the time of reference. For the 1577 journal names/abbreviations we downloaded the WOS categories, Journal impact factor, normalized eigenfactor and how these change over all available years. For the large majority of journals recent data was available.

RET	average JIF (SD)	J_k^E	f	with Y02				without Y02			
				FFET	average JIF(SD)	J_k^E	f	average JIF (SD)	J_k^E	f	
all RETs	7.57(10.3)	3.61	2.10	all FFETs	5.36(7.9)	3.56	1.50	4.64(6.9)	3.58	1.29	
solar thermal	5.17(7.3)	2.88	1.80	cracking	5.95(8.2)	3.88	1.54	5.21(6.6)	3.97	1.31	
photovoltaics	9.08(11.7)	3.61	2.52	gasification	4.52(5.8)	3.26	1.39	3.41(1.9)	3.10	1.10	
wind	4.21(4.8)	3.94	1.07	fuels	6.30(8.8)	3.50	1.80	5.59(8.4)	3.58	1.56	
energy storage	7.88(9.8)	3.77	2.09	steam engines	3.19(5.8)	2.77	1.15	2.93(5.9)	2.85	1.03	
hydrogen energy	8.80(11.0)	3.83	2.30	gas turbines	3.65(6.5)	3.03	1.21	3.59(7.2)	2.92	1.23	
fuel cells	6.32(8.3)	3.63	1.74	steam generation	1.55(2.1)	2.78	0.56	1.13(1.3)	2.75	0.41	
smart grids	5.10(6.0)	5.03	1.01	combustion	3.71(5.8)	3.05	1.22	3.74(5.7)	3.02	1.24	
non-fossil fuels	6.54(9.0)	3.51	1.86	furnaces	2.01(4.2)	3.15	0.64	2.56(5.0)	3.15	0.81	
clean combustion	4.12(6.5)	3.03	1.36	heat exchange	4.92(8.3)	3.03	1.63	4.43(7.9)	3.06	1.45	

Table 2.B.1: *Comparison average JIF and expected JIF* For each technology we determine the average JIF, the expected JIF J_k^E based on the reference distribution over the WOS scientific categories, and the factor of proportionality f between these quantities.

2.B JIF and NEF statistics

In this appendix, we provide as statistical background of the analysis done for the Journal Impact Factor (JIF) and Normalized EigenFactor (NEF).

The JIF is known to be biased towards certain WOS scientific categories (Waltman, 2016). We therefore retrieved from WOS for each scientific category i the aggregated JIF (J_i). From the distribution of references of a technology k to scientific categories specific, we calculate the expected JIF (J_k^E) based on the aggregated JIF of these scientific categories, i.e.,

$$J_k^E = \sum_i \frac{J_i r_{i,k}}{r_k^T} \quad (2.1)$$

where $r_{i,k}$ are the number of NPL references in technology k to scientific category i and r_k^T is the total number of NPL references of technology k (where we only count the NPL references for which a JIF factor could be retrieved). In Table 2.B.1 we compare the J_k^E with the measured JIF averages for each technology and for the aggregated RETs⁸ and aggregated FFETs. We observe small variations in J_k^E across technologies, where the values of the RETs indeed appear slightly larger (for RETs aggregated 3.61 versus for FFETs aggregated 3.56). However, as we indicate in the same table, the average JIF are generally a factor $f > 1$ larger than J_k^E , where the values of f shown in the same table. Where f is for RETs aggregated 2.10 and for the different RETs on average 1.75, it is for different FFETs aggregated 1.50 and for different FFETs on average 1.23. Within a certain scientific category, the RETs therefore appear to cite journals with impact factors which are relatively high for that category, and more so than FFETs. When we remove the Y02 patents from the FFETs, this difference becomes even more pronounced. This suggests that, while there are small differences between RETs and FETs due to the bias of JIF towards certain scientific categories, the large differences found between RETs and FFETs are mostly due to a tendency of RETs to build on higher impact science within those categories.

The average JIF in table 2.B.1 is for the aggregated RETs and FFETs determined using respectively 27509 and 5881 references. While the distribution of the JIF itself

⁸Note that the aggregated RETs include more than the 9 other technologies in Table 2.B.1, see table 2.1

is highly skewed, we trust that our sample size is large enough that the distribution of the average JIF can be approximated to be normal. We can thus do a Welch t-test comparing the average JIF of aggregated RETs and FFETs, which points out that the averages are significantly different ($t = 18.504$, $df = 10650$, $p < 2.2e - 16$). The average JIF of the FFETs and FFETs without Y02 is similarly found to be significantly different ($t = -4.3096$, $df = 6103.3$, $p = 1.661e - 05$).

We repeat the impact analysis for a second measure of scientific relevance offered by WOS: the normalized eigenfactor. The eigenfactor is an 'influence measure' based on applying pagerank (or eigenvector centrality) to citation networks of scientific journals (West et al., 2010). Like the JIF, the eigenfactor of a source is mainly based on the number of forward citations this source receives. Unlike the JIF, the eigenfactor of a journal takes into account the eigenfactor of the source of citations, thereby rewarding being cited by a source which is itself highly cited. Further, where the JIF varies across disciplines, the eigenfactor is relatively insensitive to these differences, because it focuses on the proportion of citations going to a given source rather than on the absolute number going to that source (West et al., 2010). Finally, self-citations do not play a role in the calculation of the eigenfactor. The *normalized* eigenfactor is determined "by rescaling the total number of journals in the JCR [Journal Citation Report] each year, so that the average journal has a score of 1" ("Web of Science Core Collection Help", n.d.). Analogous to the procedure with the JIF, we then (1) retrieve when available the normalized eigenfactor (NEF) from WOS for each cited journal cited by the technologies, (2) determine the distribution of references to each journal for each technology and (3) calculate the average NEF per technology.

For the average NEF the aggregated RETs and FFETs have respective values 22.88 and 14.07, with standard deviations 39.57 and 31.81 and number of references 25844 and 5229. Similarly applying a Welch t-test points out that the NEF average is significantly different for RETs than for FFETs ($t = 17.477$, $df = 8838.8$, $p < 2.2e - 16$). The NEF average of the aggregated FFETs without Y02, valued 11.78 with standard deviation 26.98, is likewise found to be significantly different from that of aggregated FFETs ($t = -3.2625$, $df = 5521$, $p = 0.0011$).

Next, we have a closer look at the relation between the log of the number of references and the JIF in Figures 2.8,2.9, and likewise for the NEF in Figures 2.B.1, 2.B.2 (found in this appendix). The observed positive relation is tested and fitted with an OLS estimation for the JIF in Table 2.B.2 and for the NEF in Table 2.B.3. We draw three conclusions based on these regressions

1. The coefficients are significant to a 0.05 level for the RETs, FFETs and FFETs without Y02, both for the JIF and NEF.
2. The coefficients are rather similar for RETs, FFETs and FFETs without Y02. As a matter of fact the coefficient difference between the RETs and FFETs (and FFETs without Y02) turns out to be insignificant for both indicators. We conclude this using the test for small sample sizes described in (COHEN, 2016), the details of which can be found in Table 2.B.4.
3. While the coefficients are significant, most constants (i.e. the value of the JIF or NEF when the number of references =1) are not. This therefore does not allow us to conclude that the RETs have higher values than FFETs for

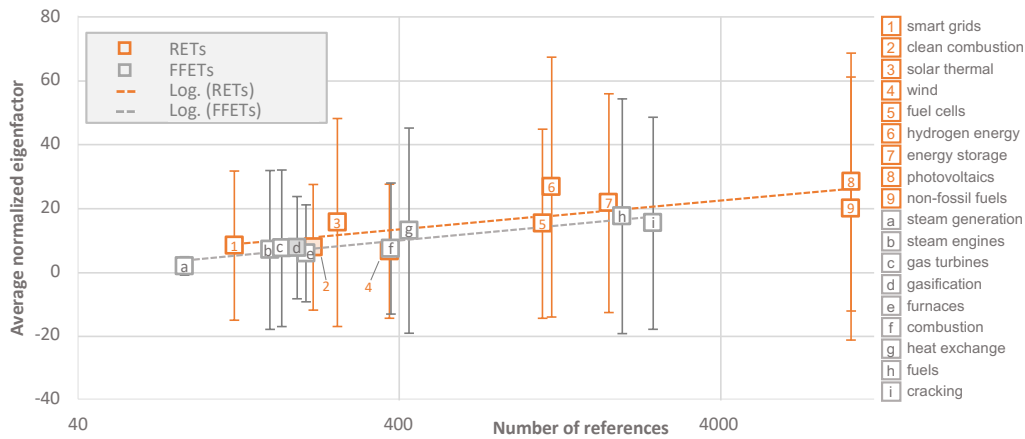


Figure 2.B.1: **Normalized eigenfactor RETs and FFETs with Y02** We plot for all RETs and FFETs with Y02 patents the average normalized eigenfactor for the number of references, where the grey bars indicate the standard deviations and the lines indicate a logarithmic fit. Note that the positive relation between the JIF and log number of references equally counts for RETs and FFETs and the coefficients hardly differ.

a similar number of references, even though Figures 2.8,2.9 and 2.B.1 suggest this.

Together, these three conclusions point out that the higher average impact of the sources built on by RETs (as opposed to FFETs) is closely related to the greater tendency of RETs to refer to scientific literature.

Table 2.B.2: **Normalized Journal Impact Factor for the number of references** This table presents three regressions, where the dependent variable is the average JIF for the RETs, FFETs and FFETs without Y02 patents and the independent variable is the log of the number of references. All coefficients are significant on a 0.05 level, yet the constants are not.

	Dependent variable:		
	JIF RETS	JIF FFETs	JIF FFETs no Y02
	(1)	(2)	(3)
log references	0.864** (0.305)	1.235*** (0.314)	0.984** (0.303)
Constant	0.363 (2.167)	-3.474 (1.925)	-1.542 (1.623)
Observations	9	9	9
R ²	0.534	0.688	0.600
Adjusted R ²	0.467	0.643	0.543
Residual Std. Error (df = 7)	1.373	0.969	0.923
F Statistic (df = 1; 7)	8.018**	15.419***	10.512**

Note:

*p<0.1; **p<0.05; ***p<0.01

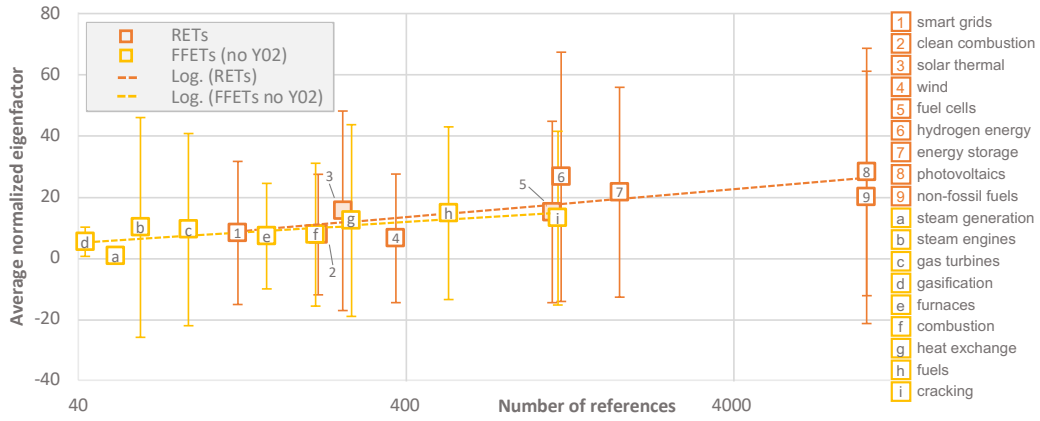


Figure 2.B.2: *Normalized eigenfactor RETs and FFETs without Y02* Same as Figure 2.B.1 except for the FFETs without Y02 patents. Both the NEF and number of references are lower when the Y02 patents are removed.

Table 2.B.3: *Normalized eigenfactor for the number of references* This table present three regressions, where the dependent variable is the average NEF for the RETs, FFETs and FFETs without Y02 patents and the independent variable is the log of the number of references. All coefficients are significant to a 0.05 level, yet not all constants are.

	Dependent variable:		
	NEF RETs	NEF FFETs	NEF FFETs no Y02
	(1)	(2)	(3)
<i>log</i> references	3.945** (1.182)	4.062*** (0.585)	2.914** (0.933)
Constant	-10.433 (8.314)	-14.576*** (3.481)	-5.794 (4.832)
Observations	9	9	9
R ²	0.614	0.873	0.582
Adjusted R ²	0.559	0.855	0.523
Residual Std. Error (df = 7)	5.397	1.909	2.973
F Statistic (df = 1; 7)	11.134**	48.283***	9.766**

Note:

*p<0.1; **p<0.05; ***p<0.01

coefficient pair	JIF		NEF	
	t statistic	p	t statistic	p
RET-FFET	-0.793	0.22	-0.077	0.47
RET-FFET no Y02	-0.258	0.40	0.619	0.27

Table 2.B.4: *Coefficient differences* We statistically evaluate for the JIF and NEF the coefficient difference between RETs & FFETs and the RETs & FFETs no Y02 using the test described in (COHEN, 2016), which applies to the difference of coefficients based on pairs of independent, small samples. As we see all p values are larger than 0.05 and the differences are therefore not significant.

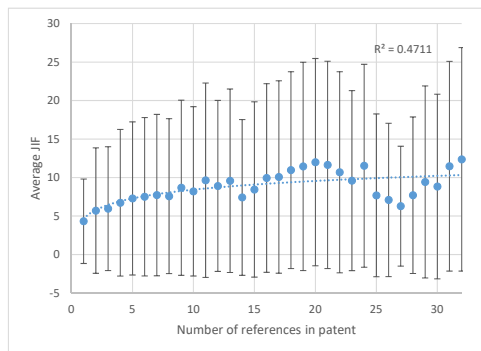


Figure 2.B.3: **Average JIF for the number of references** We sort each reference (for which a JIF could be identified) by the number of references of the patent in which it appears, and we accordingly determine the average JIF and standard deviation (see error bars). The average JIF is reasonably fitted by the logarithm of the number of references, which is confirmed by the relatively high correlation coefficient.

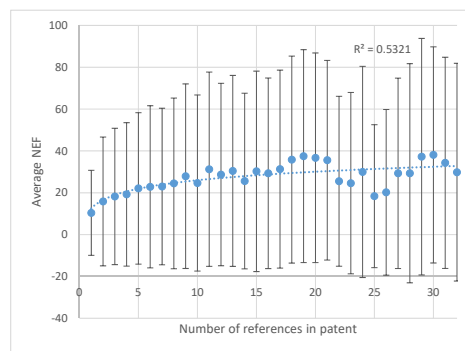


Figure 2.B.4: **Average NEF for the number of references** We sort each reference (for which a NEF could be identified) by the number of references of the patent in which it appears, and we accordingly determine the average NEF and standard deviation (see error bars). The average NEF is reasonably fitted by the logarithm of the number of references, which is confirmed by the relatively high correlation coefficient.

Finally, we note that the positive relation is found between the (log) number of references and the *average* JIF (and NEF) for different technologies. This does not necessarily imply the relation also counts on the level of individual patents. Comparing the JIF averages of technologies in Figure 2.8 to the science dependences of technologies in Figures 2.3 and 2.4 however appears to indicate a close relation between these as well, suggesting that the average JIF is not only related to the number of references, but also to the number of references per patent. To test the relation with the number of references per patent more directly we plot it for the JIF and NEF respectively in Figures 2.B.3 and 2.B.4 (for all references in this study). Indeed for these, we also observe a logarithmic relation between the number of references per patent and the average JIF and NEF (both are highly significant). The standard deviations are large however, suggesting there is substantial variation across patents. While there appears to be a clear relation, it is not exactly clear what mechanism is behind this relation. Finding this mechanism is however beyond the scope of this research.

Chapter 3

How cumulative is technological knowledge?

Peter Persoon, Rudi Bekkers and Floor Alkemade

This chapter is published as: Persoon, P.G.J., Bekkers, R.N.A., Alkemade, F. (2021). How cumulative is technological knowledge? Quantitative Science Studies, 1-27. DOI: 10.1162/qss_a_00140

Abstract

Technological cumulateness is considered one of the main mechanisms for technological progress, yet its exact meaning and dynamics often remain unclear. To develop a better understanding of this mechanism we approach a technology as a body of knowledge consisting of interlinked inventions. Technological cumulateness can then be understood as the extent to which inventions build on other inventions within that same body of knowledge. The cumulateness of a technology is therefore characterized by the *structure* of its knowledge base, which is different from, but closely related to, the *size* of its knowledge base. We analytically derive equations describing the relation between the cumulateness and the size of the knowledge base. In addition, we empirically test our ideas for a number of selected technologies, using patent data. Our results suggest that cumulateness increases proportionally with the size of the knowledge base, at a rate that varies considerably across technologies. Furthermore, this rate is inversely related to the rate of invention over time. This suggests that the cumulateness increases relatively slow in rapidly growing technologies. In sum, the presented approach allows for an in-depth, systematic analysis of cumulateness variations across technologies and the knowledge dynamics underlying technology development.

3.1 Introduction

Technology progresses when engineers adapt their designs based on learning about previous designs. Consequently, a key element of theories of technological change is the cumulative nature of knowledge and invention: the idea that new results build on - or recombine - previous results (Basalla, 1989; Freeman & Soete, 1997; Nelson & Winter, 1982; Trajtenberg et al., 1997). Indeed, many of today's technologies have rich histories of development, some going back all the way to antiquity. While the size of the knowledge base of these technologies is substantial, this does not necessarily imply the underlying knowledge structure is cumulative: a pile of stones is different from a stone wall, and some walls are higher than others.

Cumulativeness (or sometimes 'cumulativity') may therefore vary per technology and over time. A better understanding of the underlying mechanisms of technological cumulativeness is important for a number of reasons. From an economics perspective, the extent to which a technology develops in a cumulative manner has implications for how easy it is to enter or diversify into that technology. Entry is considered more difficult in complex technologies that require extensive and in-depth knowledge about the underlying principles (Breschi, 2000; Breschi et al., 2000; Winter, 1984). Recent contributions from the geography of innovation describe how regions are more likely to diversify into technologies that are related to their existing knowledge base (Balland, 2016; Balland & Rigby, 2017; Boschma et al., 2015). An understanding of the cumulative nature of technological development is thus pivotal for ongoing efforts of smart specialization (Foray, 2014), where regions seek out attractive technologies for future specialization. From a philosophical perspective, a better understanding of cumulativeness and its role in the evolution of technological knowledge (Arthur, 2009) may help to clarify the relation between knowledge accumulation and the complexity of that knowledge, which is an ongoing discussion in the 'cumulative culture' literature (Dean et al., 2014; Tennie et al., 2009; Vaesen & Houkes, 2017). Developing this understanding starts from a clear definition and measure of cumulativity.

Surprisingly, despite the recognized importance of cumulativity, the exact meaning of the concept often remains unclear. Characterizations vary from the incremental change in artifacts (Basalla, 1989; Butler, 2014; Gilfillan, 1935b; Ogburn, 1922), to the persistence of innovative activity (Cefis, 2003; Malerba & Orsenigo, 1993; Suárez, 2014), to the building of technological knowledge on earlier findings (Enquist et al., 2011; Merges & Nelson, 1994; Scotchmer, 1991; Trajtenberg et al., 1997).

In this contribution we aim to develop a better understanding of technological cumulativeness by taking the following steps: In Section 3.2 we present a comprehensive review of the various perspectives on cumulativeness and identify their common grounds. In Section 3.3 we use this analysis to formulate two indicators which measure cumulativeness: the *internal dependence* and *internal path length*. In Section 3.4 we then discuss how the values of these indicators are expected to change as a technology develops. In Section 3.5 we test these expectations empirically for a number of technologies, using patent data as a proxy for inventions. Finally, we discuss some deeper implications of our contribution to the understanding of technological cumulativeness in Section 3.6 and summarize our main conclusions in Section 3.7.

3.2 Theoretical perspectives on technological cumulativeness

Where in most texts 'cumulative' simply means 'summed up', in the innovation literature the term has come to represent a type of technological development. Perspectives on cumulative technological development however vary across contributions.

The earliest ideas about technological cumulativeness arise in studies of the gradual change in pre-20th century artifacts (Butler, 2014; Gilfillan, 1935a; Pitt-Rivers, 2018), which are reminiscent of fossil records of gradually evolving species. Inspired by evolutionary theory, these theories understand technological change as a process in which antecedent artifacts are *replicated with incremental modifications*, thereby creating descendant artifacts (Gilfillan, 1935b; Ogburn, 1922). In this first perspective, artifacts are literally the sum of many incremental modifications, justifying the term 'cumulative'.

While the cumulative aspect of technology arises naturally in this perspective, it is unclear when a development is not cumulative: as in genetic lineage, *each* descendent is supposed to have an antecedent. Some authors have argued that in reality, technological developments occasionally 'jump'; when a radical finding breaks fundamentally with past engineering practices and ideas (Schoenmakers & Duysters, 2010; Verhoeven et al., 2016) it may initiate a new model of solutions to selected technological problems, i.e., a new *technological paradigm* (Dosi, 1982). In this second perspective, cumulative development is *the opposite of radical development*, and interpreted as the incremental change happening within a technological paradigm.

Yet, to base cumulative change solely on the notion of incremental change raises two difficulties. First, there is a certain arbitrariness to when a change is incremental or not. Depending on the context and their knowledge of the subject, different people may characterize incrementality differently. Second, even if the change from an antecedent to descendant is radical, the antecedent may still be of crucial importance to the formation of the descendant (Basalla, 1989).

These difficulties are sidestepped in a third perspective, where a development is cumulative if a later result *depends* or *builds on* an earlier result (Breschi et al., 2000; Enquist et al., 2011; Merges & Nelson, 1994; Trajtenberg et al., 1997). 'Dependence' or 'dependency' is here interpreted in the context of technology as a body of knowledge, where new technological ideas or inventions (the 'results') draw on earlier insights, and are themselves used in later ideas and inventions. Note in this perspective, cumulativeness is a property of the *development* (not of one of the results). If we are interested in the cumulativeness of a technology, we therefore consider all developments within that technology, i.e. all dependencies between results that are part of that technology. Alternatively, authors have studied the cumulativeness of the union of multiple (or all) technologies (Acemoglu et al., 2016; Clancy, 2018; Napolitano et al., 2018), thereby focusing on inter-technology developments or dependencies. Both approaches are relevant to better understand the advancement of technology and knowledge production. In this work, we however focus on the former approach, as we are mainly interested in the question to what extent cumulativeness is an intrinsic property of a technology, and how this property varies for different technologies.

The relevance of cumulativeness as an intrinsic property of a technology is re-

flected by its role as defining element of a *technological regime* (Nelson & Winter, 1982), which defines the relevant circumstances under which innovating firms or organizations compete, thrive or fail. Within a technological regime, higher cumulateness is associated with greater appropriability of innovation and greater (geographical) concentration of innovative activity (Breschi et al., 2000; Malerba & Orsenigo, 1996; Winter, 1984). The framework of technological regimes gave rise to a number of contributions which use yet another perspective of cumulateness, where the emphasis is not so much on the dependence of later generations of a technology on earlier ones, but more on the continuation of those generations (Apa et al., 2018; Breschi, 2000; Cefis, 2003; Frenz & Prevezer, 2012; Hölzl & Janger, 2014; Malerba et al., 1997). Cumulateness is then characterized by the *persistence* of inventive and innovative activity in a technology: the longer a development continues (without significant interruption), the greater the cumulateness. Where previous perspectives focus more on cumulateness as an intrinsic property of technology, this fourth perspective also attributes a role to the creators of the technology (and their persistence to continue along a given path).

In summary, we recite from these four different perspectives the key notions of technological cumulateness: (1) as replication with incremental modifications, (2) as within-paradigm (opposite to radical) development, (3) as dependence or building on earlier technology and (4) as persistence of inventive or innovative activity. The first two perspectives approach cumulateness as 'incremental change', the latter two perspectives approach cumulateness as 'continuous dependence' of technology on earlier generations of technology. Though apparently very different, there are similarities between incremental change and continuous dependence. Incremental change supposes a series of modifications to what is, in some sense, a single object (often pictured as an artifact). Similarly, continuous dependence supposes a series of dependencies between objects which are, in some sense, different (often pictured as a set of inventions). Essentially therefore, the discrepancy is about the object(s) to which a series of changes is applied, yet both advocate the relevance of *a series of developmental steps*. Further, both for incremental change and continuous dependence, cumulateness appears in two dimensions: (i) the size of each developmental step: if the modification is small (dependence is great), the cumulativity is large and (ii) the number of steps in the process: if there are many small modifications (a long chain of dependency links) the cumulativity is large. While (i) and (ii) both relate to cumulativity, they are theoretically very different, and we shall henceforth refer to them as the *transversal-* and *longitudinal dimension* of cumulativity respectively. Although both dimensions can be meaningfully interpreted in all four cumulateness perspectives, it appears the first two perspectives focus more on the transversal dimension and the latter two perspectives more on the longitudinal dimension. In the next section, we will propose a separate indicator for each dimension. We emphasize that both are measured *within* a certain technological field or technology. Although the interaction between multiple fields or technologies is interesting and worth studying, the focus of this work is on understanding these cumulateness dimensions within a single technology.

Finally, we discuss the relation between technological cumulateness and complexity. In this contribution we will not enter the discussion about the exact meaning of technological complexity (for a good overview see Vaesen and Houkes, 2017), but instead work with the general description of a complex system consisting of many,

non-trivially interacting subsystems (Simon, 1962). One way to interpret this in the context of technology, is to consider an invention to be a system consisting of subsystems, which are (parts of) other inventions or borrowed ideas. The complex character of an invention is therefore in an abstract sense captured by the transversal dimension of cumulateness, which focuses on these direct dependencies. Intuitively, the more subsystems and dependencies, the greater the complexity (although this strongly depends the chosen measure for complexity). However, this is not the entire story. A relevant criterion for increasing complexity in the context of evolutionary systems is that a representative sample of lineages of descent increases in complexity (McShea, 1991; Vaesen & Houkes, 2017). Not only therefore should 'more complex' systems appear in time, but these should also fit into the lines connecting antecedents and descendants. In the context of technological knowledge, the lines of descent appear rather literally in the mentioned first perspective of cumulateness, and correspond to the longitudinal dimension of cumulateness. Especially the joint consideration of the transversal and longitudinal dimensions of cumulateness therefore allows us to study the dynamics of technological complexity.

3.3 Measuring cumulateness

In most contributions mentioning cumulative technological development, cumulateness remains an abstract property without explicit measure. There are a number of exceptions however, in particular the contributions adhering to the earlier mentioned 'persistence perspective' of cumulateness. These contributions base their measures of cumulateness on a variety of sources: survey data (Breschi et al., 2000; Frenz & Prevezer, 2012; Hölzl & Janger, 2014), licensing data (Lee et al., 2017) and statistical properties of patent count time series (Breschi, 2000; Cefis, 2003; Malerba et al., 1997). While all of these highlight interesting aspects of cumulative processes, none of them seem to directly proxy the key property of knowledge building on knowledge. Survey data may offer detailed information on the usage of particular knowledge, yet it is challenging to quantify and generalize this information in order to compare different technologies. Approaches based on counting backward citations (Apa et al., 2018) arguably do measure the extent to which knowledge builds on earlier knowledge, yet without specifying *which* technologies are cited, only partially capture the underlying knowledge structure of technologies. However, as was argued in the previous section, to understand technological cumulateness along both the transversal and longitudinal dimension, studying the underlying knowledge structure is pivotal. In this contribution, our starting point is to interpret this structure as a network of interconnected elements of knowledge. Each node then represents a single invention, and each link represents a knowledge flow. A link thus naturally corresponds to a dependence, or knowledge building on other knowledge. This approach has been successfully applied to the analysis of breakthrough innovation (Dahlin & Behrens, 2005; Fleming, 2001; Verhoeven et al., 2016), main paths (Hummon & Dereian, 1989; Verspagen, 2007), emerging technologies (Érdi et al., 2013; Shibata et al., 2009) and technological network evolution (Valverde et al., 2007). We denote the knowledge flows *to* an invention (i.e., the links which indicate on which knowledge the invention builds) as 'backward links' and the knowledge flows *from* an invention as 'forward links'.

Further, we assume that there is a technology classification that allows us to

assign each invention to at least one class, hence allowing us to distinguish between *internal links* (link to an invention in the same class) and *external links* (link to an invention of another class)¹. In the previous section we introduced the transversal and longitudinal dimensions of cumulativeness. Exploiting useful network structures, we will in the next two subsections introduce two indicators measuring the cumulativeness along these dimensions. For the transversal dimension we introduce the internal dependence and for the longitudinal dimension we introduce the internal path length.

3.3.1 The transversal dimension: Internal dependence

The transversal dimension of cumulativeness reflects the extent to which findings in a given technology *depend* on other findings within that technology. In a network of inventions, each directed link can rather literally be interpreted as a relation of dependence. Ideally, we would go into the content of each knowledge link to distinguish a degree of dependence. Yet this approach would be difficult to automate when the number of links and inventions becomes large (which is the case for most technologies). Most network approaches to technology therefore count each knowledge link equal, so the number of internal links becomes a measure for the dependence. Each invention that is added to the technology introduces a number of backward internal links, see Figure 3.3.1 (left panel) for a network illustration. The more internal backward links it introduces, the more the technology builds on itself. As a measure for the transversal dimension, we therefore define the *internal dependence (id)* of a technology as the *average number of backward internal links per invention*. A high id signals high cumulativity in the transversal dimension.

3.3.2 The longitudinal dimension: Internal path length

The longitudinal dimension of cumulativeness reflects the number of steps in a series of technological developments. Approaching technology as a network of inventions, we can translate this rather literally to a chain of internal inventions connected by links, which translates to the notion of a 'path' in the terminology of network analysis, see Figure 3.3.1 (right panel) for a network illustration. The longer the internal paths, the longer a series of developments within a technology is continued. As multiple knowledge aspects of a technology may develop in parallel, we generally deal with several, intertwined paths. As a measure for the longitudinal dimension, we therefore define the *internal path length (ipl)* of a technology as the *average length of all paths within*

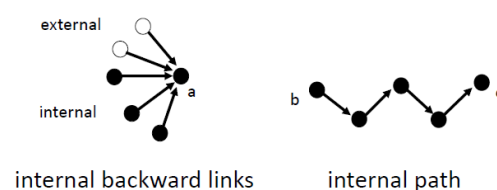


Figure 3.3.1: **Useful network structures** Left: The number of internal backward links of node *a* is 3. Right: the length of the internal path between node *b* and *c* is 4. For a precise definition of path and path length we refer to section 3.4.3.

¹Inevitably, there is some room for interpretation here as there can be various grounds on which technologies are classified. In the Section 3.6 we discuss a number of alternative approaches to making the external-internal distinction.

that technology. A high ipl signals high cumulativity in the longitudinal dimension.²

3.4 Modeling the knowledge dynamics

In this section we discuss how the values of the internal dependence (id) and internal path length (ipl) are expected to change as a technology develops, i.e. when the size of its knowledge base increases. More specifically, we analyze (a) how the id and ipl change as the number of inventions increase and (b) how the id and ipl are interrelated. We thereby describe both general and technology-specific elements.

3.4.1 Invention as search process

In this section, we sketch a highly simplified model of the invention process in a certain technology, which consists of an inventor performing a series of searches. Essentially, the inventor searches until he or she succeeds in completing an invention, where a knowledge flow (equivalent to a backward internal link) is picked up along with each search. The relevant quantity in this process is the probability ρ of completing an invention before performing another search, which may depend on the size of the knowledge base of a technology, as measured by the number of inventions n . For each n , the probability of inventing is therefore $\rho(n)$, the probability for performing a search is $1 - \rho(n)$. We have two main assumptions in this model:

1. The probability $\rho(n)$ decreases proportionally with the number of inventions n . This reflects the intuition that it becomes harder, as a technology develops, to produce an invention without using any prior knowledge developed in that technology. In other words, the inventor needs to consider some knowledge in a certain field before delivering a contribution to that field, and the larger the field, the more the inventor needs to consider.
2. The probability of success is independent of the number of searches: in the invention process, there is no guarantee that a certain amount of effort leads to success.

Given these assumptions we may write down for the probability $\rho(n) = \frac{1}{qn+m_1}$, introducing the technology specific constants $q > 0$ and $m_1 > 1$. Here, the parameter m_1 describes the need to have knowledge of the technology in order to invent at the initial stage of this technology, and q describes how fast this need increases as the technology develops. As a consequence of the two assumptions, the probability for a node to have m backward internal links (i.e. the probability that m searches take place before invention) is given by $P_n(m) = (1 - \rho(n))^m \rho(n)$, i.e. the number of backward internal links per node is distributed geometrically.³ This distribution is characterized by a highly skewed shape towards lower values of m , yet as n increases, it slowly becomes less skewed.

²Similar ideas are presented in (Frenken et al., 2012), where innovations attain 'higher quality' with longer path lengths. The study of Frenken et al., which is based on numerical simulations, thereby focuses on the (re)combination principle in relation to diffusion.

³We then assume that the number of backward links per node stays well below n , which appears to be reasonable if we consider technologies with large n .

The rate q is related to the type of technological knowledge and we therefore assume it is a technology-specific quantity. Yet we hypothesize it is also related to the rate of invention over time. Our reasoning is as follows. If the rate of invention over time is high, this means that more people work on the same technology at the same time. If multiple researchers work on the same technology, they tend to specialize, focusing only on a particular sub-field or sub-part of the technology. As an effect, multiple aspects of the technology develop in parallel, perhaps more so than if a smaller group of people had worked on it. As a result, the development of the technology is more fragmented into sub-fields, which causes inventors active in these sub-fields to focus on the relevant findings within their sub-field. We may therefore suppose that there is structurally less need for these inventors to master the entire knowledge base, which leads to lower values of q . In reverse, it is possible that a low need for prior knowledge of a technology accelerates innovative activities in a technology, as it may then be more easily accessible, thus inviting more people to contribute. Deriving a more precise form and causal direction of the inverse relation between q and rate of inventing over time however is beyond the scope of this work. For a more elaborate discussion of the causality we refer to Section 3.6.

3.4.2 Internal Dependence Dynamics

Using the distribution of the number of backward internal links, we can calculate $\langle m \rangle = \sum_{m=0}^n m P_n(m)$ the expected value of the number of backward internal links per invention, i.e. the internal dependence (id). Assuming that n is large, we can approximate this sum by choosing infinity for the upper limit and using the expression $P_n(m) = (1 - \rho(n))^m \rho(n)$, obtaining

$$\langle m \rangle = \frac{1}{\rho(n)} - 1 = qn + m_1 - 1 = qn + m_0, \quad (3.1)$$

introducing $m_0 = m_1 - 1$ for convenience. We therefore conclude that the id is expected to increase proportionally with the number of inventions (i.e. with the size of the knowledge base), where the rate can be approximated by q for a large number of inventions. This technology-specific coefficient q describes how fast the need to have specialized knowledge increases in order to produce an invention in that technology.

3.4.3 Internal Path Length Dynamics

Next we will discuss how we expect the internal path length (ipl) to depend on the number of inventions. Although these results can be generalized by including external links, we focus in this contribution for simplicity on the role of internal links. A new invention creates at least one new path with each of its internal backward links. The internal dependence, besides measuring a complementary dimension of cumulativity, therefore also plays a key role in the ipl dynamics. Let us again consider a technology with n inventions, where the n th invention has *on average* $\langle m \rangle$ internal backward links. Some inventions however will have no backward links, which we will refer to as *initial inventions*. As a first assumption, we take that the number of initial inventions n_0 is a fixed fraction r of n , i.e., $n_0 = rn$.⁴ We use the

⁴As we explain in more detail in Appendix 3.E this assumption is compatible with the found backward link distribution if q is small compared to m_0 , we then have that $r \approx 1/(m_0 + 1)$.

initial inventions to define a path and path length:

- A *path* is a sequence of inventions $\mathbf{i}_0, \mathbf{i}_1, \dots, \mathbf{i}_k$ in which for any $k \geq 0$ and $x > 0$, \mathbf{i}_x has a backward link to invention \mathbf{i}_{x-1} and \mathbf{i}_0 is an initial invention.
- The *path length* of path $\mathbf{i}_0, \mathbf{i}_1, \dots, \mathbf{i}_k$ is k .

We denote the number of paths of length k by $f_k(n)$. From the first assumption⁵, we have that, $f_0(n) = rn$. As a second assumption, each invention is equally likely to be used as prior knowledge with probability $\frac{1}{n}$. Let us consider what happens to $f_k(n)$ for $k > 0$ when we introduce the $n + 1$ th invention. If that invention builds on a prior invention \mathbf{i} that has $l_{i,k-1}$ paths of length $k - 1$, each of these paths will increase by 1, hence $f_k(n)$ increases by $l_{i,k-1}$. This holds for all inventions, which in total have $\sum_i l_{i,k-1} = f_{k-1}(n)$ paths of length $k - 1$. For $\Delta_n f_k(n)$, i.e. the expected increase in $f_k(n)$ from n to $n + 1$, we therefore have $\Delta_n f_k(n) \propto f_{k-1}(n)$, and for $k > 1$ we have

$$\Delta_n f_k(n) = \langle m \rangle \frac{f_{k-1}(n)}{n}. \quad (3.2)$$

In the previous section we established that $\langle m \rangle \approx qn + m_0$. When n gets large, $\langle m \rangle/n \rightarrow q$, further reducing Equation 3.2 to

$$\Delta_n f_k(n) = q f_{k-1}(n). \quad (3.3)$$

As there are no paths for $n = 0$, we take that $f_k(0) = 0$ for all k . Using this initial condition and the expression for $f_0(n)$, the solution to Equation 3.3 is derived to be

$$f_k(n) = r q^k \binom{n}{k+1}, \quad (3.4)$$

where $\binom{x}{y}$ is the binomial coefficient. The steps leading to this solution and later ones are explained in more mathematical detail in Appendix 3.E. Summing over all k we obtain the total number of paths $\sum_{k=0}^n f_k(n) = r(1+q)^n/q - r/q$. The total number of paths is therefore expected to increase exponentially in n . For the normalized path length distribution $\tilde{f}_k(n)$, describing the probability to have a path of length k , we subsequently obtain

$$\tilde{f}_k(n) = \binom{n}{k+1} \frac{q^{k+1}}{(1+q)^n - 1}, \quad (3.5)$$

which is a distribution closely related to the binomial distribution. This indicates that as n increases, the path length distribution will shift from a skewed shape towards more symmetric, parabolic shape (on a log scale) and its maximum, the most frequent path length, will continuously shift to higher values. Subsequently, we can calculate the expected path length $\langle k \rangle = \sum_{k=0}^n k \tilde{f}_k(n)$, i.e. the ipl, which reduces for large n to

$$\langle k \rangle \approx \frac{q}{q+1} n + k_0, \quad (3.6)$$

⁵If we also consider external inventions, we can choose a more general definition, where a path can also start at an external invention. Note that, ignoring the links to external inventions, the inventions which only link to external inventions become initial inventions

where k_0 is some constant value. As we focus on large n behavior, we are less interested in this constant. What is more important is the expectation that the ipl increases proportionally with the number of inventions, by a rate $p = q/(q + 1)$. This rate p is a number between 0 and 1: for large q it is close to 1 and for small q , it is close to q . We end this section by mentioning two extensions of the model which improve its explanatory power.

- In this derivation, we assumed that $\langle m \rangle/n \approx q$, even though we know it in fact only *approaches* q for large n . This approximation can be significantly improved by instead calculating the average $\langle m \rangle/n$ for n inventions. We can determine this quantity in two ways, (a) by directly using the data of the number of backward links for each invention, i.e. by calculating $q'_a = 1/n \sum_i m_i/i$ where m_i is the number of backward links of invention i and (b) by using estimates for parameters in the relation $\langle m \rangle = m_0 + nq$, i.e. calculating $q'_b = 1/n \sum_i q + m_0/i = q + m_0 H(n)/n$, where $H(n)$ is the n th harmonic number. Analogous to Equation 3.6, we then have $q'_a/(1 + q'_a) = p'_a$ and similar for p'_b . p'_a is likely to be more accurate as it is more directly based on the backward link data, yet p'_b is less sensitive to outliers in this data. Both predictions should however be close to one another. Note that this correction depends proportionally on m_0 .
- Equation 3.2 implies that as we add the n^{th} invention to the system, the number of paths of length n increase from 0 to some positive value. In fact, this equation therefore establishes a 'maximum speed' v of 1 path length per invention, faster than which the path lengths cannot increase. This maximum speed is rather lenient: technologies with paths increasing with 1 length per invention (i.e. forming perfect chains) would be highly unrealistic. While Equation 3.2 is accurate for the more frequent path lengths (i.e. the lengths close to the mean), it may therefore be less accurate for the less frequent path lengths (i.e. the shortest and longest lengths). A more realistic estimate of the maximum speed v may therefore help establish a better description of the overall distribution of path lengths. Let us suppose that we at once add δn inventions to the system which do not connect amongst themselves, and of which the total added number of backward links is $M(n)$. Equation 3.2 then becomes

$$f_{k+1}(n + \delta n) - f_{k+1}(n) = M(n) \frac{f_k(n)}{n} \quad (3.7)$$

If we choose δn such that $M(n) \approx n$, then each of the n inventions in the system approximately obtains 1 forward link. This implies that all paths in the system increase on average by 1, including the longest path(s). δn Therefore defines a typical interval for the longest path to increase by 1, and $1/\delta n$ therefore presents a more reasonable estimate for the maximum speed v . We will use this idea to derive a new expression for the path length distribution. Note that Equation 3.7 then becomes

$$f_{k+1}(n + \delta n) - f_{k+1}(n) = f_k(n). \quad (3.8)$$

If we introduce the variable $n' = n/\delta n$ and the function $f'_k(n') = f_k(n)$, we may write this relation as $f'_{k+1}(n' + 1) - f'_{k+1}(n') = f'_k(n')$, which is solved by

$f'_k(n') = r \binom{n'}{k+1}$ (this time using the condition that $f'_k(n') = 0$ for $k < n' = nv$). This leads to the normalized distribution

$$\tilde{f}'_k(n') = \frac{1}{2^{n'} - 1} \binom{n'}{k+1} \quad (3.9)$$

and expected path length (i.e. the ipl)

$$\langle k \rangle' \approx \frac{n'}{2} + k'_0, \quad (3.10)$$

where k'_0 is again a constant we are less interested in. Rewriting this expression in terms of n gives the coefficient $\frac{1}{2\delta n}$ or $\frac{v}{2}$, describing how fast the ipl increases with n . Assuming the earlier analysis with a greater maximum speed is accurate for the mean path length values, this should coincide with the earlier established coefficient p . We can therefore approximate the maximum speed as $v \approx 2p$.⁶ This implies that the paths with maximum length grow about twice as fast as paths with mean length, i.e. the distribution becomes more symmetric as n increases. Noting that $n' = nv$, we identify n' as the maximum path length after n inventions, which can be used to evaluate Equation 3.9. Alternatively, we use the expression for $v = 2p$ to rewrite this expression in terms of n and p ,

$$\tilde{f}_k(n) = \frac{1}{4^{pn} - 1} \binom{2pn}{k+1}. \quad (3.11)$$

3.5 Empirical analysis

In this section, we empirically test the models developed in Section 3.4 using patent and patent citation data. We start with a discussion of our type of data and a number of limitations of these data. Subsequently, we perform the analysis on three different levels: first, we consider the development and distributions of both cumulateness indicators for four focus technologies into detail. Second, we consider the relation between the two indicators and the consistency of the indicators, using a larger set of technologies. Third, we choose a more aggregated level of technology classification to obtain a more general overview of the cumulateness variation across different technological fields, which also allows us to compare our findings to earlier results from the literature and to some extent validate the indicators.

3.5.1 Data description

In order to study the knowledge dynamics empirically, we need some codification of that knowledge. Patents are an important codification of technological knowledge, as each patent is a detailed description of a new, non-trivial technological development. Furthermore, patent systems have two elements that allow us to study technological content without necessarily having to consider the detailed meaning of

⁶This is consistent with the earlier assertion that $M(n) \approx n$. To see this, note that the total number of links is $n \langle m \rangle$ (as $\langle m \rangle$ is an average), hence between n and $n + \delta n$ we add $\delta n(m_0 + q(\delta n + 2n))$ links. For this to equal n in the limit where n becomes large, we require $\delta n \rightarrow \frac{1}{2q}$. In the same limit, $p \rightarrow q/(q+1)$, which is approximately q for small q . This is therefore consistent with $\frac{1}{\delta n} = v \approx 2p$

each individual patent. The first element is that of patent citations, which identify one to one, directional content relations between patents. This enables us to study the flow of knowledge (A. Jaffe, 1989; A. Jaffe et al., 1993). The second element is that of the patent classifications, which hierarchically groups patents on the basis of their content. This enables us to focus specifically on the development of a particular technology, distinguishing between internal and external knowledge. A basic assumption of our work is that cumulateness is an intrinsic property of technology, which is independent from the way the technology is patented. It is therefore important to keep in mind the limitations of representing technological knowledge by patent data, which will henceforth discuss. For each limitation, we mention how we attempt to account for it.

1. Not all technology is or can be patented, (Jaffe Adam B. & de Rassenfosse Gaétan, 2017) and the 'quality' of patents (evaluated against the patentability requirements) varies (de Rassenfosse et al., 2016; A. B. Jaffe & Lerner, 2004). Especially when the number of patents involved is small, without a detailed examination of the content we risk misrepresenting a technology. In this analysis, we therefore choose technologies for which the number of patents is relatively large. Also, we only consider *granted* patents, which have withstood the critical assessment of patent examiners.
2. Citations may not always represent actual knowledge flows (Criscuolo & Verspagen, 2008). Citations may be provided by inventors but may also be added by examiners, and while the first may be more indicative for knowledge flow, the distinction was not always documented by all patent offices (Azagra-Caro & Tur, 2018). We therefore include an additional analysis in Appendix 3.D of the effect of both types of citations (examiner or inventor added) to the knowledge dynamics.
3. There are institutional differences between patent offices around the globe, which may affect the way inventions and linkages to prior art are documented. (Bacchiocchi & Montobbio, 2010). An important difference is for example is the greater tendency to cite in the United States patent system than in the European patent system (Criscuolo & Verspagen, 2008), which may impact the value of our indicators. To account for these differences we therefore do this analysis for patents from two different patent systems, choosing the US system (organized by the US patent office USPTO) and European system (organized by the European Patent Office EPO).

To aggregate patents of which the technological content is the same, we choose a patent family as a basic unit or node, creating a US data set selecting families with at least one USPTO member and a European data set selecting families with at least one EPO member⁷. In the US data set each unique reference (backward citation) of a US member of each family to any member of another family in our data set represents a unique link (hence we do not limit our selection to US-US citations only).⁸ Our European data set is created analogously. Henceforth by 'US patent'

⁷To be precise, we choose the DOCDB type of patent family, where all family members have exactly the same priorities

⁸Note that if we had selected *any* family citation we effectively take the union of all citations, hence failing to distinguish between the citing tendencies of different patent systems.

we actually refer to an patent family containing a US member which is granted, and similar for 'European patent' or 'EP patent'.

In order to select and demarcate technologies, we used the Cooperative Patent Classification (CPC) (CPC, 2018). In this analysis, we consider technologies on two levels of classification: the CPC group/subgroup level and a more aggregated level of classification. For the group/subgroup analysis we choose a set of 24 arbitrary technologies, yet making sure that (i) the set is diverse (including technologies from each main CPC section and from mostly different subclasses) and (ii) each technology contains a reasonably large number of patents (for US >700 and EP >200). Table 3.5.1 and Table 3.B.1 indicate the CPC codes and number of patents of these technologies. Table 3.5.1 singles out four 'focus technologies' which we will analyze in more detail. The sub-selection of the focus technologies was made choosing considerable variation in (a) knowledge base size (where nuclear fission has 3608 US patents, photovoltaics has over 9000), (b) age (where nuclear fission started developing in the 1960's, the main development of wind turbines starts from the 1990's), (c) the working (theoretical) principles behind the technologies (varying from nuclear physics to aerodynamics). From both Table 3.5.1 and 3.B.1 it is clear there are generally more US than European patents, even taking into account that the EP patents do not go back further than 1978. As the column with the number of patents in the same family indicates, most European patents (around 75 percent) have a US equivalent as well.

For the more aggregated level of classification, we grouped together patent classes analogous to the approach by Malerba and Orsenigo (Malerba & Orsenigo, 1996). However, given that their publication now dates more than 20 years back, and the patent classification system is subject to constant change, some differences between their grouping of classes and ours is inevitable⁹. In Table 3.B.2 in Appendix 3.B we present an overview of our grouping, note that we take the union of CPC classes (hence counting each patent once). The data in this research comes from the Patstat 2019 spring edition. Time is not adopted as an explicit variable in our models, yet

Table 3.5.1: *Description of the four focus technologies. The selected patents have an earliest filing year < 2009.*

Technology short name	CPC code	CPC description	#US granted patents	# EP granted patents	# same family
Nuclear Fission	Y02E 30/3	Energy generation of nuclear origin: nuclear fission reactors	3608	745	558
Photovoltaics	Y02E 10/5	Energy generation through renewable energy sources: photovoltaic energy	9088	2599	1947
Wind Turbines	Y02E 10/7	Energy generation through renewable energy sources: wind energy	5405	1767	1323
Combustion Engines	F02B 3/06	Engines characterised by air compression and subsequent fuel addition with compression ignition	6466	2089	1344

we check for the consistency of our models over time and at a later point consider the invention rate over time. We do that by using the earliest filing date of the patent, as it is the closest point in time to the actual invention and therefore helps to establish a chronological ordering of inventions. It generally takes several years however before filed patents are actually granted: the European patents granted in 2012 were on average first filed 6.5 years earlier, for US patents this was about 5 years. Likewise,

⁹As a matter of fact, the CPC did not yet exist at the time of the Malerba and Orsenigo paper, yet the closely related International Patent Classification (IPC) did.

from all patents eventually granted which were filed earliest in 2005, it took 50 percent 6.9 years to be granted, and it took about 12.5 years for 95 percent of them to be granted. For US patents filed earliest in 2005 the same percentages correspond to about 5 and 10 years respectively. To be relatively confident to include 95 percent of the patents for each year considered, hence avoiding a 'truncation effect' as much as possible, calculating back from 2019, we should therefore not consider earliest filing years later than 2008.

3.5.2 Id and ipl for the focus technologies

In Figure 3.5.1 we plot the id and ipl of the four focus technologies for the number of patents. We include the results from both the US and EP patents. We observe for all four technologies a linear increase of both the id and ipl, yet the rate of increase varies considerably across technologies. In the US data set, where wind turbines is after 2000 patents already at an ipl of 10, combustion engines reaches the same ipl only after 6000 patents. These variations are also found considering the id or EP dataset instead. It is therefore instructive to consider not only the absolute cumulateness of a technology, but also its cumulateness relative to the size of its knowledge base.

To obtain a more detailed understanding of the linear relationship between cumulateness and the number of inventions, we consider the coefficients of the linear fits in Figure 3.5.1 for US patents in Table 3.5.2 and for the EP patents in Table 3.5.3, the statistical details of these fits can be found in Appendix 3.A. The coefficients in Table 3.5.2 indeed vary considerably across technologies, and high values for m_0 correspond to high values of q . This suggests that if the need for specialized knowledge is high at the initial stages of a technology, it also increases faster as the technology develops. More importantly, Table

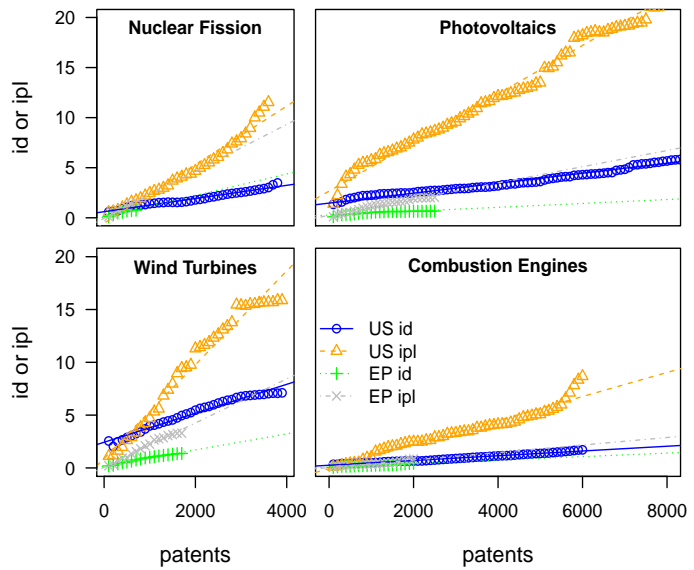


Figure 3.5.1: *Id and ipl for number of patents (US and European patents)* We plot the id and ipl for every 100 patents (represented by symbols) and linear fits (represented by lines). For statistical details of the fit see Appendix 3.A.

3.5.2 shows that the fitted ipl coefficients p (on the left, determined empirically) are in reasonable agreement with the predicted values (on the right, calculated). This suggests that the id and ipl are interrelated in accordance with the simple model described in section 3.4.3. This implies that the relation between id and ipl is rather predictable for each technology, suggesting that the transversal and longitudinal dimensions of cumulateness (using a proper re-scaling) can be used interchangeably. In Table 3.5.3 the variation across technologies for EP patents is largely similar to the variation for US patents, and again shows reasonable agreement between the

Table 3.5.2: *Coefficients of id and ipl for US.* On the left we present the fitted id and ipl coefficients and the id constants from Figure 3.3.1 for US patents. On the right we present the predicted ipl coefficients, where p'_a is directly based on the id data and is p'_b calculated using the fitted m_0 and q (see Section 3.4.3). With the exception of US wind turbines, these predictions agree rather well with the fitted ipl coefficients. As expected p'_a is generally more accurate than p'_b . For statistical details see Appendix 3.A

	id const m_0 (ref/pat)	id coeff q (ref/pat ²)	ipl coeff p (1/pat)	p'_a (1/pat)	p'_b (1/pat)
Nuclear Fission	0.65	0.0006	0.0029	0.0029	0.0022
Photovoltaics	1.45	0.0005	0.0024	0.0024	0.0020
Wind Turbines	2.42	0.0014	0.0044	0.0067	0.0067
Combustion Engines	0.26	0.0002	0.0011	0.0011	0.0006

Table 3.5.3: *Coefficients of id and ipl for EP.* Same as Table 3.5.2, but then for EP patents.

	id const m_0 (ref/pat)	id coeff q (ref/pat ²)	ipl coeff p (1/pat)	p'_a (1/pat)	p'_b (1/pat)
Nuclear Fission	0.07	0.0011	0.0023	0.0024	0.0017
Photovoltaics	0.25	0.0002	0.0008	0.0008	0.0010
Wind Turbines	0.18	0.0008	0.0021	0.0019	0.0016
Combustion Engines	0.07	0.0002	0.0004	0.0005	0.0005

fitted and predicted ipl coefficients. There are however also some overall differences with the US patents. The constants m_0 are generally smaller and, as a consequence, the ipl coefficients p are also smaller. There are minor differences per technology, the id coefficient q of nuclear fission being remarkably higher for the EP patents than for the US patents. We will revisit cross technology differences more systematically at the end of this chapter.

Finally, we observe in Figure 3.5.1 some minor deviations from the linear developments, in particular the ipl of nuclear fission and combustion engines speeding up for higher number of patents, and that of wind turbines slowing down. Additionally, the ipl of combustion engines and photovoltaics increases fast at a lower number of patents. (A closer analysis of the id leads to similar observations, though this is less clear in Figure 3.5.1). We will come back to these deviations in our discussion of Figure 3.5.2.

In Figure 3.5.2 we plot for the US patents the id and ipl over time, together with the total number of patents over time.¹⁰ The ipl values (shifted by k_0) are re-scaled by the corresponding factor p and the id values (shifted by m_0) are re-scaled by the corresponding factor q from Table 3.5.2. We observe for all four technologies that the time development of all three quantities largely coincides. In hindsight, this should not be a surprise given the observed linear relations in Figure 3.5.1: the id and ipl are mainly a function of the total number of patents and hence their developments are synchronized. The synchronization indicates that our modeling of the knowledge dynamics consistently applies over time, i.e. that it is to some extent time-independent. Still, we note the synchronization is not always perfect: towards 2009, we observe that the ipl of nuclear fission and especially combustion engines increases faster than the number of patents, and vice versa for wind turbines. Further, in the 1960s, the ipl of photovoltaics and combustion engines is somewhat lower than the number of patents.

¹⁰The number of backward citations only starts to become substantial from 1940 onward for all considered technologies, which is why we choose this as a starting point. We note that wind turbines have a substantial number of patents (about 1300) before 1940, yet citations before that period are either rare or not recorded in our data set.

Note that these asynchronous developments correspond exactly to the previously mentioned deviations from linearity in Figure 3.5.1. Note in Figure 3.5.2 that the 'fast ipl' deviations correspond to periods in time where the number of patents increase very slow, (nuclear fission and combustion engines towards 2009, photovoltaics and combustion engines in the 1960s) and that the 'slow ipl' deviations correspond to periods in time where the number of patents increase very fast (wind turbines towards 2009). To some extent, but less clearly in Figure 3.5.2, this also counts for the id developments. These observations are therefore in agreement with the hypothesized inverse relation between the rate of invention over time and cumulateness coefficients.

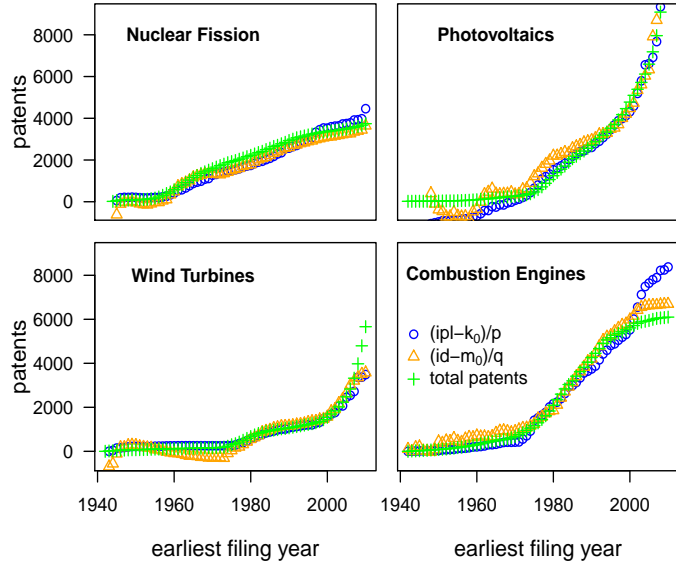


Figure 3.5.2: **Total patents and re-scaled id and ipl over time** For each earliest filing year we plot the total number of patents and the ipl and id, where $ipl-k_0$ is re-scaled by a factor $1/p$ and $id-m_0$ is re-scaled by a factor $1/q$ (the factors are taken from Table 3.5.2 and appendix 3.A). Both cumulateness indicators closely follow the development of the total patents over time.

3.5.3 Distributions of backward links and path length

The measured linear relationships between the id, ipl and number of patents are in line with the model predictions of Section 3.4, yet linear relationships may also arise in various other models. Additionally, we therefore study the empirical backward link and path length distributions and compare these to the predicted distributions. For brevity we focus on the US patents in this section, as the analysis for EP patents is largely similar.

In Figure 3.5.3 we plot the internal backward link distribution for the four focus technologies for US patents, plotting the distribution for each technology for every 1000 patents. We observe two characteristics: (1) the frequency drops exponentially (note the logarithmic axis) with the number of references and zero references occurring most frequently, (2) as the number of patents increase, the skewness decreases. Where (1) is indicative for a geometric distribution, (2) indicates that the parameter of this distribution depends on the number of patents. To test if these distributions agree with the predictions of Section 3.4.2, we in Figure 3.5.3 simultaneously plot the predicted distributions using the parameters q and m_0 from Table 3.5.2. We observe the predicted distributions fit the empirical distributions rather well. In appendix 3.C we compare these fits to a number of alternative distributions using probability plots, which again confirm the data is reasonably well described by geometric distributions with parameters from Table 3.5.2.

In Figure 3.5.4 we consider the path length distribution (for each internal path)

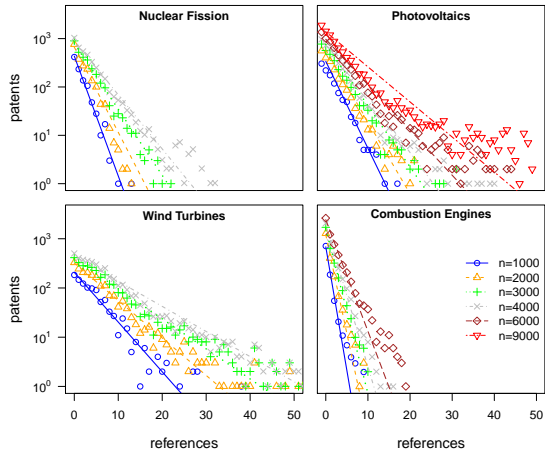


Figure 3.5.3: **Backward link distribution (US patents)** With symbols we plot the empirical distribution each time the number of patents increase by a 1000. With lines we plot the predicted geometric distributions using the parameters q and m_0 from Table 3.5.2. For clarity, we omit $n = 5000, 7000$ and 8000 (applicable to combustion engines and photovoltaics only).

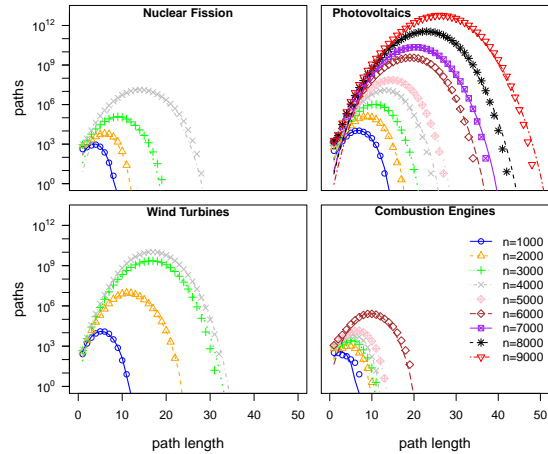


Figure 3.5.4: **Path length distribution (US patents)** With symbols we plot the empirical distribution each time the number of patents increase by a 1000. With lines we plot the predicted distributions from Equation 3.9, where the maximum path length values plotted by numerals in Figure 3.5.5 are used as values of n' .

for US patents, plotting the distribution for every 1000 patents. We observe two characteristics: (1) as the number of patents increase, the distribution becomes less skewed, approximating a parabolic shape (on a log scale) (2) the most frequent path length shifts right as the number of patents increase. Before we discuss the fitting of the path length distributions, we shortly consider the development of the maximum internal path length (mipl) for the number patents in Figure 3.5.5. The patterns are for each technology rather similar to those of the ipl in Figure 3.5.1, except that the mipl increases at about double the pace.

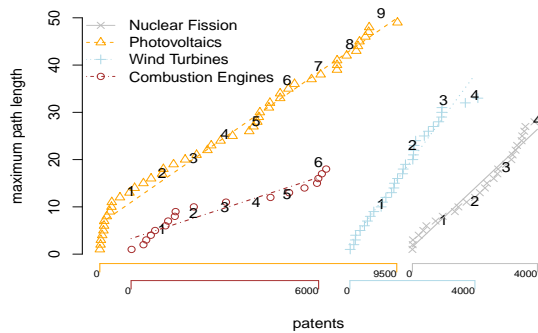


Figure 3.5.5: **Maximum path length for US patents** We plot the maximum internal path length for the number of patents (in symbols) and include linear fits (in lines) of the development. Details of the fits are included in Appendix 3.A. The plotted numerals each correspond to one of the fitting distributions plotted in Figure 3.5.4, the x-coordinate representing the number of patents n , the y-coordinate representing the n' used in the fitting distribution (see equation 3.9).

Indeed, linear fits of the mipl, see Appendix 3.A for the details, yield the coefficients 0.0062 (nuclear fission), 0.0046 (photovoltaics), 0.0089 (wind turbines) and 0.0021 (combustion engines), which are as expected very close to $2p$ (using the values p from Table 3.5.2). As explained at the end of Section 3.4.3, we can use the maximum path lengths as estimates for n' in Equation 3.9. The values of n' used in the fitting path length distributions in Figure 3.5.4 are the y-coordinates of the plotted numerals in Figure 3.5.5. We note that the empirical distributions in Figure 3.5.4 are very well fitted, and at the same time the numerals in Figure 3.5.5 fall neatly in the line of development of each technology. We therefore conclude that Equation 3.9 provides a rather ac-

curate description of the path length distributions. A closer examination of the distribution fits is provided in Appendix 3.C.

3.5.4 Cross technology variations

Finally, we discuss the variation in *id* and *ipl* across technologies in more detail. As we have two indicators for cumulativeness and two data sources (EPO and USPTO), we can identify cross cumulativeness variations along 4 dimensions. Figure 3.5.6 presents a systematic overview of the 24 technologies from Table 3.5.1 and 3.B.1. We observe positive trends for each comparison in Figure 3.5.6, and the technologies for each comparison remain rather consistently in a characteristic (high or low) range of cumulativeness. These observations are supported by the reasonable values of the squared correlation coefficient R^2 and the statistical significance we find for each comparison (for the statistical details see Appendix 3.A). The positive association between the two indicators in Figure 3.5.6 provides some evidence that the relation established in Equation 3.6 applies across a wider range of technologies than the four focus technologies. This suggests again that the degree of cumulativeness measured along the transversal and longitudinal dimensions largely agree for each technology, i.e. that both dimensions can be used more or less interchangeably. As expected from the greater citation tendency in the US system, the values of the US indicators are a factor 3-4 greater than their EP counterparts (see also Appendix 3.A). However, the positive association we find between US and EP patents for both the *id* and *ipl* indicates that this factor is approximately constant across different technologies. This suggests that, despite institutional differences, both indicators can be applied consistently within different patent systems, confirming that we can think of cumulativeness as a technology-specific characteristic.

In our discussion of the four focus technologies we provided some evidence for the hypothesized inverse relation between the time rate of invention and the *id* coefficient q (i.e. the rate at which the *id* increases per patent). The joint consideration of 24 technologies allows us to test this relation for a wider range of technologies. In Figure 3.5.7 the invention rate (measured by the average number of patents per year) is plotted for the *id* coefficients q (determined by the number of references per patent squared) for both the US and EP patents. In line with expectation, the two quantities are negatively associated (best fitted by a power law with a power ≈ -0.6 for US patents and ≈ -0.9 for EP patents). Again see Appendix 3.A for the statistical details. Figure 3.5.7 therefore confirms that the linear coefficient determining the increase of the *id* per patent (and indirectly the *ipl*) is related inversely with the rate of invention over time. Note that this does not mean the rate of cumulativeness development is exclusively determined by the rate of invention, as there may still be other factors at play related to the type of technology or technological knowledge. From Figure 3.5.6 it is not directly clear however what type of technologies we can typically associate with high and low cumulativeness: we observe technologies from various disciplines both on the higher and lower end of the spectra. In the final subsection, we therefore consider the differences between technologies on a more aggregated level of classification.

Figure 3.5.6: *Id and ipl for 24 technologies in EPO and USPTO. Linear fits are included based on the pairwise regressions in Appendix 3.A.*

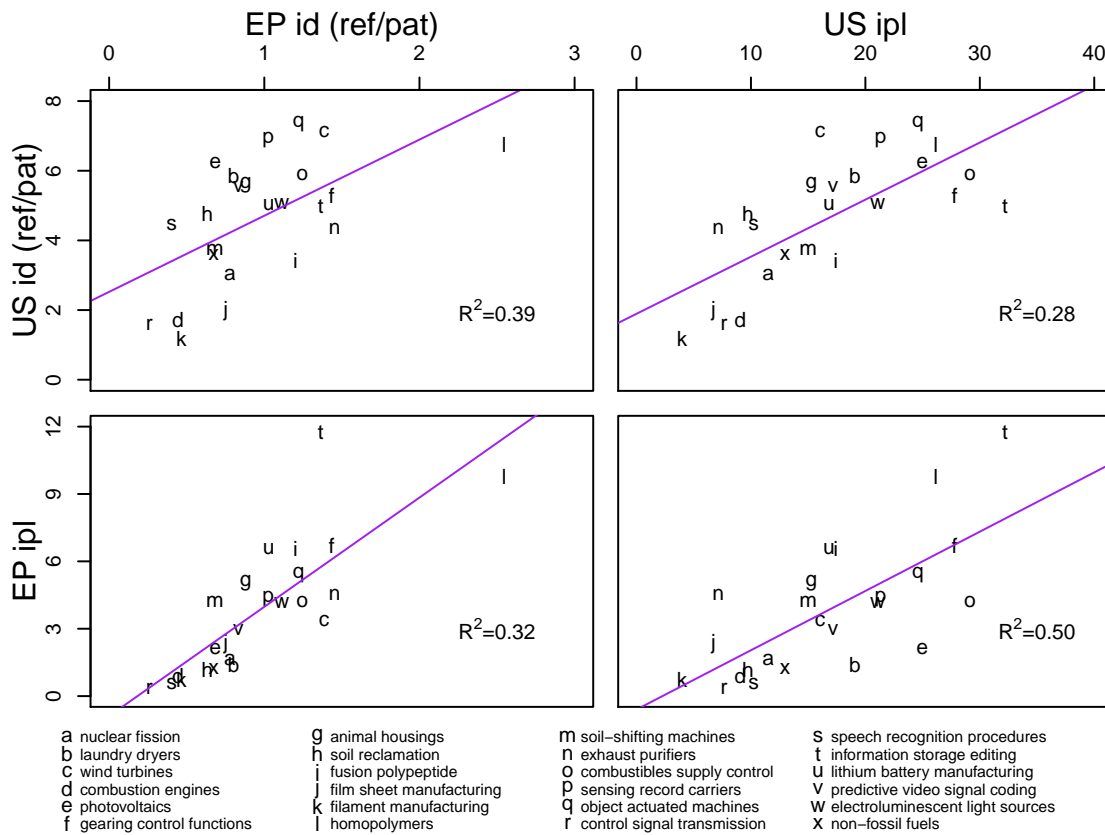
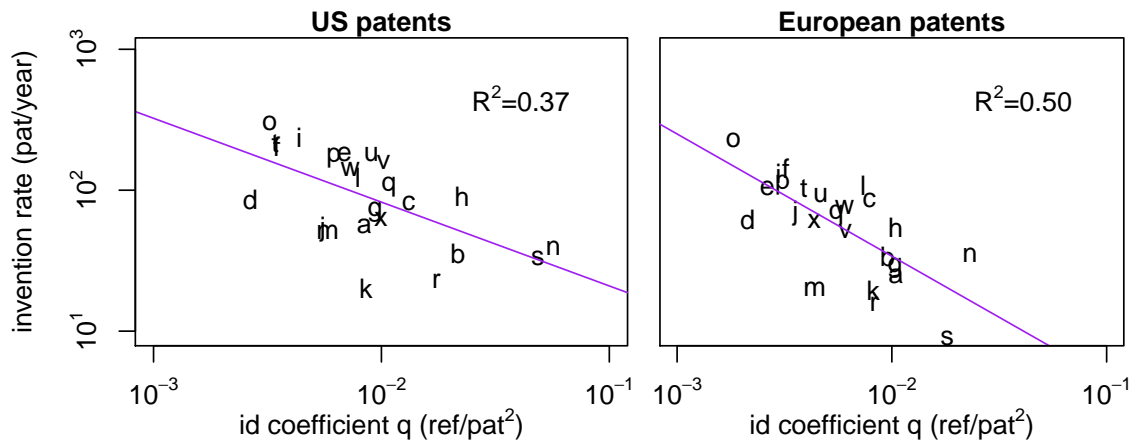


Figure 3.5.7: *Invention rate and id rate q* The symbols correspond to the technologies in the legend of Figure 3.5.6. Note the axes are logarithmic, hence the fitted line is a power-law (for details see Appendix 3.A).



3.5.5 Cumulativeness across technological fields

Finally we consider the cumulativeness of technologies on a more aggregated classification level, which we will henceforth refer to as 'technological fields', (for an overview see Table 3.B.2 in Appendix 3.A). This allows us to develop an overall understanding of which technologies can typically be associated with high or low cumulativeness. Furthermore, it allows us to check if our approach to cumulativeness is in line with earlier approaches (Breschi et al., 2000; Malerba & Orsenigo, 1996)¹¹, thus to some extent validating the indicators for cumulativeness suggested this contribution. However, as determining the ipl is computationally challenging for very large numbers of nodes (i.e. >100,000), we limit this analysis to determining the id of these technological fields. We plot the id for the number of patents for these fields for the US patents in Figure 3.5.8, where we also include a legend. Note that the different icon colors correspond to the different CPC main sections. Figure 3.5.9 shows a similar plot for the European patents.

For a deeper understanding of a technology's cumulativeness, we again stress the need to additionally consider the cumulativeness relative to the size of the knowledge base. For example, in Figure 3.5.8, while the knowledge base size is similar for the field Packing & Transporting and the field Optics & Photography, the latter has reached a far greater level of cumulativeness. Similarly, Nucleonics reaches the same cumulativeness level as Packing & Transporting while the knowledge base is about 15 times larger in the latter. The cumulativeness therefore appears to increase faster with each patent for Nucleonics and Optics & Photography than for Packing & Transporting. The expected increase in cumulativeness for the knowledge base size is indicated by the fits (dashed line) in Figures 3.5.8 and 3.5.9 and may depend on the level of classification. For this level of classification we can use these fits to distinguish *relatively high* cumulativeness (above the line) from *relatively low* cumulativeness (below the line). Using this distinction we see for the US patents the fields belonging to CPC sections Physics (red icons), Electricity (yellow icons), and Chemistry (purple icons) show relatively high levels of cumulativeness. Fields belonging to Sections Human Necessities (brown icons) and Performing operations & Transporting (blue icons) show relatively low levels of cumulativeness. The larger fields in the Sections Textiles (pink icons) and Fixed Constructions (black icons) too show relatively low levels of cumulativeness.

The study by Malerba and Orsenigo (M&O) distinguishes a number of highly aggregated technologies as Schumpeter Mark I (associated with low cumulativeness) and Schumpeter Mark II (associated with high cumulativeness). Our observations are in overall agreement with the general conclusion of M&O that *"Schumpeter Mark I technological classes are to be found especially in the 'traditional' sectors, in the mechanical technologies, in instruments as well as in the white electric industry. Conversely, most of the chemical and electronic technologies are characterized by the Schumpeter Mark II model."*¹². To make a more detailed comparison, we individually consider 23 technological fields which occur both in the M&O and our own set of fields and which M&O classify as either Schumpeter I or II. For the purpose of this

¹¹The contribution by Breschi is largely consistent with the one by Malerba and Orsenigo. As the latter consider more detailed technological classes and a wider geographical range of patents, we will focus on the latter.

¹²We interpret M&Os 'traditional' sectors to correspond to the early industrial and craft-like sectors such as Textiles, Domestic Articles and Wearables

Figure 3.5.8: *Cumulativeness versus size of knowledge base for US patents* We plot the cumulativeness (measured by the internal dependence) for the knowledge base size (measured by the number of patents) for 40 technological fields based on USPTO data. Fields in the same CPC section are colored similarly. Note both axes are logarithmic, hence the fitted regression line is a power law. The cumulativeness of technologies appearing substantially above (below) the fitted line can be identified as relatively high (low).

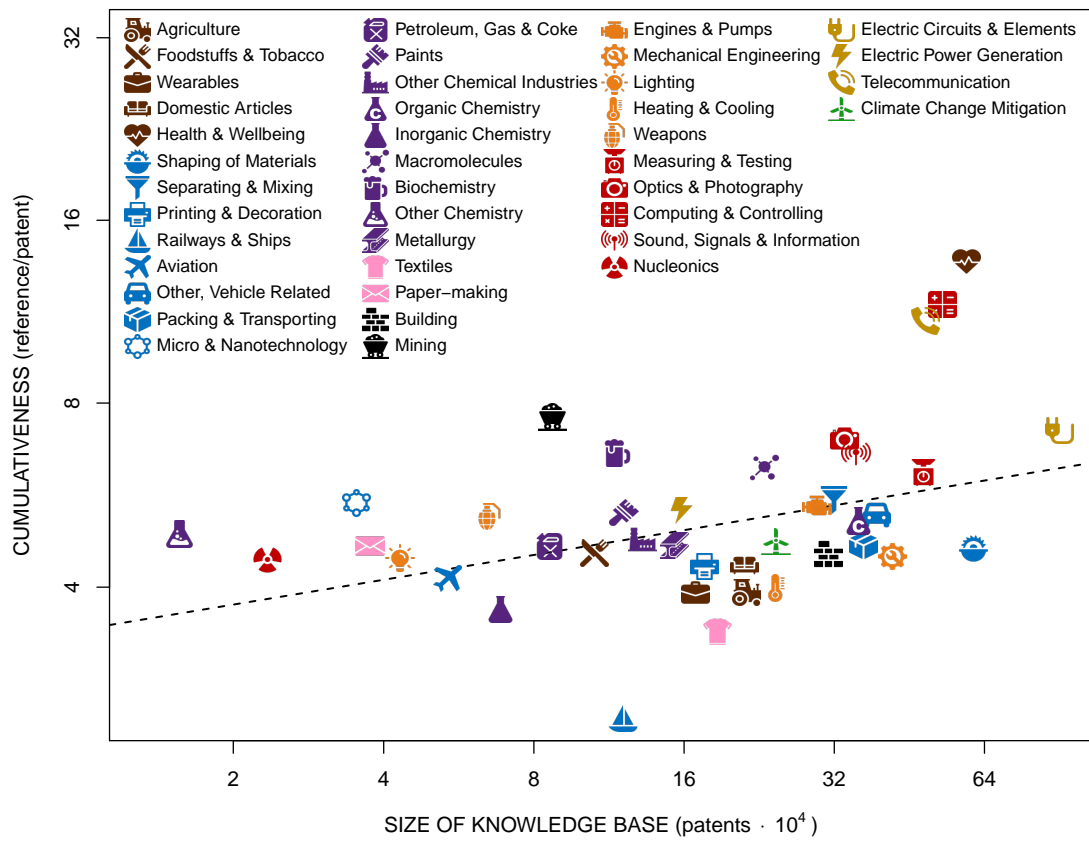
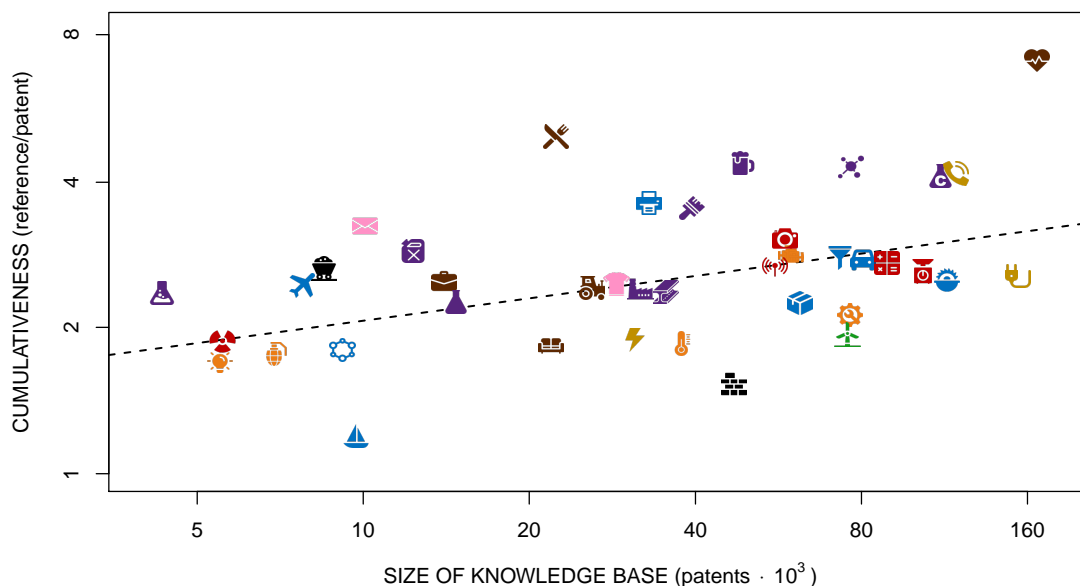


Figure 3.5.9: *Cumulativeness versus size of knowledge base for European patents* Same as Figure 3.5.8 but then for European patents. A legend for the icons is included in Figure 3.5.8.



comparison, we associate a technological field below the fitted line with low cumulativenness (which should correspond to M&O's Schumpeter Mark I) and technologies on or above with high cumulativenness (corresponding to M&O's Schumpeter Mark II). From the 23 thus considered technologies, 18 are identified correctly: 7 as low cumulativenness (Wearables, Domestic Articles, Agriculture, Shaping of Materials, Railways & Ships, Building, Mechanical Engineering) and 11 as high cumulativenness (Aviation, Petroleum, Gas & Coke, Macromolecules, Biochemistry, Engines & Pumps, Weapons, Photography & Optics, Nucleonics, Telecommunications, Computing & Controlling, Electronic Components & Circuitry). 5 Technological fields do not correspond to M&O's labeling: Inorganic Chemistry (Mark II), Printing & Decoration (Mark II), Lighting (Mark I), Measurement & Testing (Mark I) and Health and Wellbeing (Mark II). Note the first four are rather close to the line, however. The cumulativenness of Health and Wellbeing is exceptionally high though in our analysis. The reason for these deviations is not directly clear. We emphasize that M&O's Schumpeter Mark I or II labels are based on various aspects of the organization of innovation, and are therefore only an indirect indication of cumulativenness. Also, there might be some variation between the grouping of patent classes by M&O and ours. Finally, some technologies may have developed substantially between the M&O study (1996) and the final year we consider (2009).

The variations found across technological fields using the European patents in Figure 3.5.9 are largely similar to those we observed for the US patents. Notable differences are that for the European patents, the chemistry fields show relatively high cumulativenness and the physics and electricity fields show relatively low cumulativenness (as compared to the US patents). In general, the variations across technological fields are less for the European patents than for the US patents, which is likely related to the fact that the number of patents is substantially lower for the former. Although there are some differences with the US patents, the European

patents too show overall agreement with the results of M&O (of the 23 fields, 18 are identified correctly). The agreement between the M&O approach to cumulateness and our results provide a validation for the use of the id to measure the cumulateness of a technology, and indirectly for the ipl, given the earlier established close relation between both indicators.

3.6 Discussion

In this paper, we established an approach to interpret, model and measure the cumulative nature of technological knowledge development. We can identify a number of deeper implications and possible extensions of the theoretical model developed in this contribution.

A main point in the search model is the increasing difficulty to invent without any prior knowledge of the field, which leads to a geometric distribution for the number of backward links. In a number of other approaches, invention is perceived as a process of (re)combining existing pieces knowledge (Arthur, 2009; Fleming, 2001; Fleming & Sorenson, 2001). When we would focus on the number of combinations allowed by the number of existing inventions, a reasonable suggestion for the distribution of backward links would be a binomial type of distribution. This option may seem attractive, as assigning equal probability to each combination would lead the expected value of the number of backward links per invention to increase proportionally with the number of inventions, in agreement with observation. However, for the id we fail to observe characteristics of a binomial type of distribution. The fact that we obtain stronger evidence in Appendix 3.C for a geometric distribution suggests therefore that the mechanism of combination plays a lesser role than we might expect, or at least that we are dealing with a special type of combination, where for example only a small subset of the combinations is allowed.

While linear relations are common in descriptions of social phenomena, we emphasize that the linearity of the id and ipl in the number of inventions is neither an obvious nor an expected result. In a number of network approaches to knowledge dynamics it is instead supposed that the number of backward links per node is on average constant as the number of nodes increase (Albert & Barabasi, 2002; Price, 1976; D. Wang et al., 2013). It can be demonstrated this would imply a constant id and a logarithmically increasing or even constant ipl. These mechanisms would thus predict a stagnating cumulateness, even though the number of inventions keeps increasing. One may raise the objection that the external nodes are not included in our analysis, and that the id linearity may disappear once these are included. Additional checks on the four focus technologies in Table 3.5.1 however reveal that the external dependence (i.e. the average number of external nodes each node builds on) equally well shows a linear increase. Although considering only four technologies gives no guarantee, it is an indication that the linearity is a more general phenomenon. In this contribution we explored some possible mechanisms driving the increase of id and ipl. At the same time, we acknowledge that there may be other societal factors driving the increase, such as increased computerization or other factors improving the availability of search results. Accounting for the effect of these factors is however challenging, as it would require us to compare similar technological developments over different time periods.

Our approach suggests that the cumulateness of technologies develops largely

in sync with the size of the respective knowledge base, which suggests that these knowledge dynamics are to some extent time-independent, i.e. less impacted by historical events. Likewise, the description was formulated independently of spatial (geographical) factors (and appeared consistent between the US and Europe). This appears to contradict a commonly held notion that technology development is highly path dependent, i.e. that history and local circumstances crucially matter. However, the time and space independence here only applies to the relation between cumulativeness and the size of the knowledge base, hence the crucial choices determining particular *technological content* may still largely depend on historical or local events. Furthermore, we observed that the rate of invention of this technology over time is inversely related to the rate of proportionality between the cumulativeness and size of the knowledge base. If there is causation from former to latter, then technological cumulativeness may in the end be less determined by intrinsic knowledge properties than generally understood. If there is causation from latter to former, then the cumulativeness rate of a technology can be interpreted as key determinant and predictor of its rate of invention. Alternatively, a simultaneous effect of both causalities may also be the case. Regardless of a possible causality direction, it would for later work be interesting to compare the deviations from linearity in Figure 3.5.1 with different phases in the technology life cycles (Abernathy & Utterback, 1978; Anderson & Tushman, 1990). The development of combustion engines and nuclear fission indeed show hints of typical life-cycle s-shapes in Figure 3.5.2, the points of acceleration and deceleration corresponding to the deviations in Figure 3.5.1. While the present model does not account for these deviations, we note that, at least for the technologies here considered, the deviations are minor, and linearity remains the dominant pattern.

In this contribution we focused on the cumulativeness of *technological* knowledge. It would be interesting to compare this to cumulativeness in other fields of knowledge such as science or art. The indicators and models discussed in this contribution can reasonably well be generalized to these areas. Also, it would be interesting to look at science-technology or art-technology dependencies, which then allow us to consider the cumulativeness of technology as a whole, i.e. consider all technology as internal and the influence of science and/or art as 'external'. These questions are however beyond the scope of this work.

Finally we mention two limitations to our approach. First, our results critically depend on a particular choice for a demarcation/classification of different technologies, in our case the CPC. Even though a validated classification, innovation researchers should keep in mind the CPC is in the first place designed to aid patent examiners in their search for prior art, which may not always align with the technology definitions and level of detail researchers require. Furthermore, as new technologies develop the CPC is continuously restructured, causing possible misalignment with the researcher's time perspective of a developing technology. To allow for a more detailed classification or a more sophisticated internal-external distinction researchers may consider alternatives based on textual analysis of patents (Kelly et al., 2018), technological relatedness (Castaldi et al., 2015) or distance measures (Gilsing et al., 2008; A. Jaffe, 1989). While we acknowledge these points, we note that the main focus of this work was on developing a methodology to determine a technology's cumulativeness, which is generally applicable once the internal-external distinction is in place. In general, we emphasize that a better understanding of the

applicability of our analysis requires us to research a greater number of technologies. This would also help us understand if more closely related technologies also differ less in cumulativeness (hints of which we observe in Figures 3.5.8 and 3.5.9). Second, we kept the models in this contribution as simple as possible, thereby excluding a number of arguably relevant factors, amongst others: (i) the average time lag between the appearance of knowledge and the usage of that knowledge (ii) more advanced mechanisms in patent networks such as preferential attachment effect (Albert & Barabási, 2000; Érdi et al., 2013; Valverde et al., 2007), (iii) linkage to external inventions, which allows paths to start directly from external nodes. Though we can think of possible extensions of the model including these factors, we preferred a simple version for clarity.

3.7 Conclusions

This paper presents both a theoretical and an empirical investigation of technological cumulativeness. Theoretical perspectives agree that technological cumulativeness involves a series of developmental steps within a technology, where the cumulativeness is higher (i) when the dependence between subsequent steps is larger, and (ii) when the total number of subsequent steps is higher. We capture these transversal (i) and longitudinal (ii) dimensions of cumulativeness through our indicators *internal dependence* (id) and *internal path length* (ipl).

We then analytically derive how the id and ipl interrelate, and how they change as the size of the knowledge base of a technology increases (as measured by the total number of inventions). To this end, we model the invention process as a series of searches. A relevant parameter in this process is the technology-specific rate q at which it becomes harder to invent without using the existing knowledge in the field. We expect q to be inversely related to the rate of invention over time, as there tends to be more specialization (and hence less need for complete knowledge) at greater rates of invention. From this model we deduce that the id and ipl, while following different distributions, are both expected to increase linearly with the size of the knowledge base. The coefficients of these linear relations are predicted to approximate q as the knowledge base becomes larger.

Empirical tests on several technologies, using patent and citation data from both USPTO and EPO as proxies for invention and knowledge flow, provide empirical support for these expectations and show that the id and ipl can be used consistently for both patent systems. Further, the variations in cumulativeness across technological fields are found to be largely consistent with earlier contributions that used different approaches to technological cumulativeness: chemistry, physics and to some extent electronics are generally characterized by relatively high cumulativeness, while the craft-like and mechanical engineering fields show relatively low cumulativeness.

Our study leads to a number of new insights about technological cumulativeness and its relation to technological knowledge:

1. The cumulativeness of a technology develops proportionally with the size of its knowledge base, with a technology-specific *cumulativeness rate*. A thorough understanding of a technology's cumulativeness therefore considers the

cumulativeness both absolute as well as relative to the size of its knowledge base.

2. The measurements of cumulativeness along the transversal dimension and the longitudinal dimension are found to be consistent for various technologies. It appears therefore that both provide an equivalent description of a technology's cumulativeness. Measuring the transversal dimension by means of the internal dependence is (computationally) simple, and therefore provides a relatively fast and reliable indication of a technology's cumulativeness.
3. The time development of the cumulativeness indicators is largely synchronized with the time development of the knowledge base size. This suggests that short term, immediate effects have a limited influence on the relation between cumulativeness and knowledge base size (meaning that the cumulativeness rate remains constant). However, across technologies we observe an inverse relation between the cumulativeness rate and the rate of invention over time. This suggests that effects acting over long periods of time, such as the gradual acceleration or deceleration of inventive efforts, may therefore affect the cumulativeness rate.
4. Technological cumulativeness is understood to be a mechanism for the emergence of technological complexity. For a comprehensive understanding of the dynamics of technological complexity, it is important to take into account both the transversal and the longitudinal dimension of cumulativeness. Our study shows that cumulativeness increases along both these dimensions (for the considered technologies), which suggests an overall increase of technological complexity as well, yet this partially depends on the chosen measure of complexity.

These insights lead to a number of implications for innovation policies that benefit from a detailed understanding of the cumulativeness of technologies, such as smart specialization. In their consideration of various technologies, these policies are advised to choose a comprehensive approach, including both the absolute cumulativeness as well the cumulativeness relative to the size of the knowledge base. Where the first is indicative for the overall difficulty of entry in a technology, the second is indicative for the relative difficulty of entry as compared to technologies with similar-sized knowledge base. Furthermore, given that near future inventive activity (and with that knowledge output) allows for some estimation or planning, these policies are advised to additionally take into account the expected development of the cumulativeness of these technologies. Although these developments are sometimes considered a black box, we have demonstrated that the cumulativeness in fact develops rather predictably with the size of the knowledge base. In the longer run, policymakers should be aware that the rate of invention over time of a technology, usually a direct or indirect subject of policy interventions, is inversely related to the cumulativeness rates. Although the possible causality in this relation is as of yet unclear, the consequences are either way considerable. In the most extreme cases, it either implies a certain 'counter effect': that a substantial acceleration of inventive activities indirectly slows down the cumulativeness rate of a technology, or it implies that, despite efforts of acceleration or deceleration, the inventive rate is largely conditioned by the cumulativeness rate alone.

3.8 Acknowledgments

We would like to thank Anton Pichler, Thomas Schaper and three anonymous reviewers for helpful comments on the script. The icons in Figures 3.5.8 and 3.5.9 are made by Freepik, Eucalyp, fjstudio, Those Icons, Pixel perfect, Kiranshastry, Becris, Smashicons, Prosymbols and Good Ware from www.flaticon.com. This work was supported by NWO (Dutch Research Council) grant nr. 452-13-010

3.9 Data availability

The data used in this contribution originate from the Patstat patent database, which is available for licensing by the EPO (European Patent Office). In the Appendix 3.F we included the codes that allow for the replication of our results using Patstat data.

Appendix

In the following appendices, we include more detailed information about various aspects of the paper. In Appendix 3.A we discuss the statistics of the linear fits applied throughout the paper. Subsequently, in Appendix 3.B we present an overview of the selected technologies and their corresponding CPC classifications. Then, in Appendix 3.C we discuss the probability plots of the fitted distributions in Figures 4 and 5. This is followed by Appendix 3.D, in which we discuss and investigate the difference between applicant and examiner added citations and Appendix 3.E where we provide the detailed derivations of the equations appearing in the paper. Finally we include the T-SQL scripts in Appendix 3.F which allow for the reproduction of our results using Patstat.

3.A Statistics of the linear fits

In this section we present in the statistical details of the linear fits in Figures 3.5.1,3.5.5,3.5.6 and 8 of the paper. The fits are estimated using an ordinary least squares approach. We start with the linear fits *id* and *ipl* in Figure 3.5.1 of the paper. Tables 3.A.1 and 3.A.2 respectively represent the *id* and *ipl* for US patents, and Tables 3.A.3 and 3.A.4 respectively represent the *id* and *ipl* for EP patents. Next Table 3.A.5 present the results of the linear regressions in Figure 3.5.5. Then, Table 3.A.6 presents the outcomes of the pairwise regressions of Figure 3.5.6. Finally Table 3.A.7 presents the outcomes of the fits in Figure 3.5.7.

Table 3.A.1: *Estimated linear models for Internal Dependence (US patents)*

	<i>Dependent variable:</i>			
	Nuclear Fission (1)	Photovoltaics (2)	Wind Turbines (3)	Combustion Engines (4)
Patents	0.0006*** (0.000002)	0.0005*** (0.000001)	0.0014*** (0.000004)	0.0002*** (0.000000)
Constant	0.6462*** (0.0032)	1.4492*** (0.0045)	2.4157*** (0.0096)	0.2560*** (0.0012)
Observations	3,595	9,066	3,979	6,068
R ²	0.98	0.97	0.96	0.98
Adjusted R ²	0.98	0.97	0.96	0.99
Residual Std. Error	0.096 (df = 3593)	0.214 (df = 9064)	0.303 (df = 3977)	0.047 (df = 6066)
F Statistic	166,232*** (df = 1; 3593)	340,069*** (df = 1; 9064)	108,235*** (df = 1; 3977)	422,204*** (df = 1; 6066)

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 3.A.2: *Estimated linear models for Internal Path Length (US patents)*

	<i>Dependent variable:</i>			
	Nuclear Fission	Photovoltaics	Wind Turbines	Combustion Engines
	(1)	(2)	(3)	(4)
Patents	0.0029*** (0.00001)	0.0024*** (0.000003)	0.0044*** (0.00001)	0.0011*** (0.000004)
Constant	-0.4162*** (0.0198)	2.6356*** (0.0139)	0.9349*** (0.0302)	-0.0644*** (0.0127)
Observations	3,595	9,066	3,979	6,068
R ²	0.96	0.99	0.97	0.94
Adjusted R ²	0.96	0.99	0.97	0.94
Residual Std. Error	0.595 (df = 3593)	0.664 (df = 9064)	0.951 (df = 3977)	0.496 (df = 6066)
F Statistic	90,488*** (df = 1; 3593)	828,425*** (df = 1; 9064)	113,419*** (df = 1; 3977)	98,777*** (df = 1; 6066)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3.A.3: *Estimated linear models for Internal Dependence (European patents)*

	<i>Dependent variable:</i>			
	Id Nuclear Fission	Id Photovoltaics	Id Wind Turbines	Id Combustion Engines
	(1)	(2)	(3)	(4)
patents	0.0011*** (0.00001)	0.0002*** (0.000002)	0.0008*** (0.000003)	0.0002*** (0.000001)
Constant	0.0665*** (0.0042)	0.2522*** (0.0029)	0.1812*** (0.0028)	0.0690*** (0.0009)
Observations	744	2,598	1,766	2,092
R ²	0.95	0.81	0.98	0.97
Adjusted R ²	0.95	0.81	0.98	0.9616
Residual Std. Error	0.057 (df = 742)	0.073 (df = 2596)	0.059 (df = 1764)	0.021 (df = 2090)
F Statistic	13,048*** (df = 1; 742)	11*** (df = 1; 2596)	76*** (df = 1; 1764)	52,396*** (df = 1; 2090)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3.A.4: *Estimated linear models for Internal Path Length (European patents)*

	<i>Dependent variable:</i>			
	Ipl Nuclear Fission	Ipl Photovoltaics	Ipl Wind Turbines	Ipl Combustion Engines
	(1)	(2)	(3)	(4)
patents	0.0023*** (0.00001)	0.0008*** (0.000003)	0.0021*** (0.00001)	0.0004*** (0.000002)
Constant	-0.0473*** (0.0046)	0.2337*** (0.0046)	0.0913*** (0.0070)	0.0274*** (0.0023)
Observations	744	2,598	1,766	2,088
R ²	0.98	0.96	0.98	0.95
Adjusted R ²	0.98	0.96	0.98	0.95
Residual Std. Error	0.062 (df = 742)	0.117 (df = 2596)	0.147 (df = 1764)	0.052 (df = 2086)
F Statistic	47,153*** (df = 1; 742)	69*** (df = 1; 2596)	92*** (df = 1; 1764)	37,728*** (df = 1; 2086)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3.A.5: *Development of Maximum Internal Path Length (mipl) (US patents)* The results of this regressions are plotted in Figure 3.5.5.

	<i>Dependent variable:</i>			
	Mipl Nuclear Fission (1)	Mipl Photovoltaics (2)	Mipl Wind Turbines (3)	Mipl Combustion Engines (4)
patents	0.0062*** (0.0002)	0.0046*** (0.0001)	0.0089*** (0.0003)	0.0021*** (0.0002)
Constant	1.54*** (2.6496)	8.31*** (0.4073)	2.07*** (1.6830)	3.34 (0.7232)
Observations	28	49	33	17
R ²	0.97	0.98	0.97	0.91
Adjusted R ²	0.97	0.98	0.97	0.91
Residual Std. Error	1.48 (df = 26)	1.84 (df = 47)	1.65 (df = 31)	1.54 (df = 15)
F Statistic	814*** (df = 1; 26)	2,836*** (df = 1; 47)	1,062*** (df = 1; 31)	157*** (df = 1; 15)

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 3.A.6: *Pair wise regression between id and ipl for EP and US patents* The results of these regressions are plotted in Figure 3.5.6.

	<i>Dependent variable:</i>			
	id US patents		ipl EP patents	
	(1)	(2)	(3)	(4)
id EP patents	3.284*** (0.870)		4.625*** (1.449)	
ipl US patents		0.127*** (0.043)		0.264*** (0.057)
Constant	1.509* (0.828)	2.278*** (0.795)	-0.256 (1.378)	-0.599 (1.045)
Observations	24	24	24	24
R ²	0.39	0.28	0.32	0.50
Adjusted R ²	0.37	0.25	0.29	0.47
Residual Std. Error (df = 22)	1.49	1.62	2.48	2.13
F Statistic (df = 1; 22)	14.25***	8.73***	10.19***	21.76***

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 3.A.7: **Invention rate over time for the id coefficients q** The log of US invention rate (patent/year) is regressed for the log of the id coefficient q (reference/patent²) for the 24 technologies in Tables 3.5.1 and 3.B.1, for both the US and EP patents. The q coefficients are calculated by dividing the total references by the total patents squared. The results of these regressions are plotted in Figure 3.5.7

	<i>Dependent variable:</i>	
	Log Invention Rate US	Log Invention rate EP
	(1)	(2)
Log US id coefficient q	-0.595*** (0.166)	
Log EP id coefficient q		-0.867*** (0.186)
Constant	0.302 (1.174)	-2.463* (1.398)
Observations	24	24
R ²	0.368	0.496
Adjusted R ²	0.339	0.473
Residual Std. Error (df = 22)	0.625	0.579
F Statistic (df = 1; 22)	12.791***	21.635***

Note:

*p<0.1; **p<0.05; ***p<0.01

3.B Overview of selected technologies and corresponding CPC classifications

Earlier we included more detailed information on the four focus technologies in Table 3.5.1. A more detailed description of the other 20 technologies in Figure 3.5.6, including their CPC classifications and number of granted patents, can be found in Table 3.B.1. While the choice for these technologies was mostly arbitrary, we took care to include technologies from each main CPC section (indicated by first letter A,B,C,D,E,F,Y) and from mostly different CPC subclasses (indicated by first 4 symbols e.g. C10J). To limit the scope of the technologies we selected the technologies on the CPC groups and subgroup level (the most dis-aggregated two levels of the CPC). Even though this selection of the group and subgroups was mostly arbitrary, we took into account that we require a substantial number of patents for each technology (>200). Then, in Table 3.B.2 we present an overview of technologies on a more aggregated level of classification, which we earlier referred to as 'technological fields' (and appear in Figures 3.5.8 and 3.5.9). The choice for this level of classification and the particular grouping of CPC classes was done such that the technological fields correspond as much as possible to the technologies appearing in Malerba and Orsenigo (Malerba & Orsenigo, 1996). We also include the number of unique US and EP patents in these fields.

3.C Evaluating the distribution fits

In this appendix we discuss the fits of the distributions in Figures 3.5.3 and 3.5.4. For the distributions where less data is available (i.e. $n=1000, 2000$), χ^2 tests indicate there is not enough evidence to reject the null hypothesis that the backward link distributions are described by geometric distributions with parameters from Table 3.5.2. However, for larger n , the p -values quickly get very small for virtually any distribution we try, which suggests that the χ^2 test is rather strict for our purpose. Instead, we therefore consider probability plots instead, where we compare the performance of the predicted distribution to a number of other possible candidates, such as the binomial distribution for the backward links and the normal distribution for the path length. The x-value of each point in the probability plot represents the empirical probability of a certain occurrence and its y-value represent the predicted probability of its occurrence. The closer the points to the $x = y$ line, the better the distribution fit therefore. The Figure 3.C.1 shows the probability plots for the (empirical) backward link distribution of the four focus technologies and three candidate distributions: the geometric, normal and binomial distributions. We choose to show the $n = 3000$ case, yet the other cases are largely comparable. The parameters of each distribution are chosen such that the fit with the empirical distribution is optimized. We observe that the geometric distribution is for all technologies very close to the $x = y$ line, more so than the other distributions.

Similarly, Figure 3.C.2 shows the probability plots for the path length distribution and three candidate distributions: the Poisson, normal and binomial type of distribution of Equation 3.9¹³. For the path length distributions too we observe that the distribution from Equation 3.9 is generally close to the $x = y$ line for each

¹³We choose again the $n = 3000$ cases, except for computational reasons we chose $n = 2000$ for

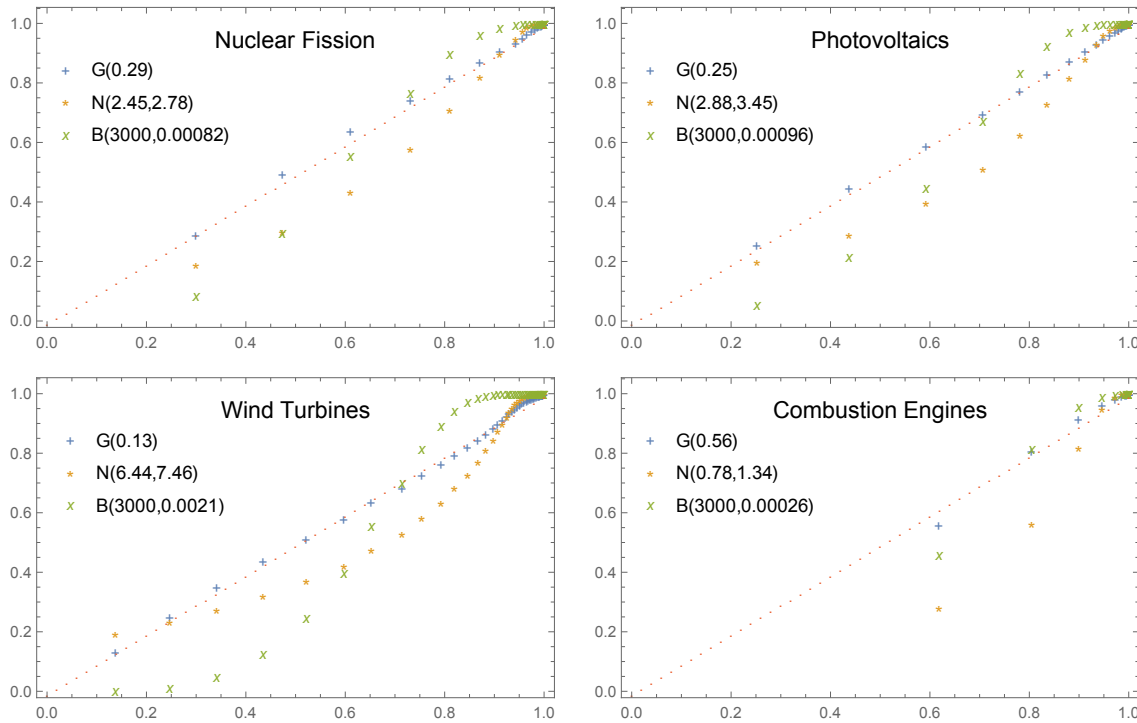
Table 3.B.1: *CPC code and description of additional technologies*

Technology name	CPC description	CPC code	# US granted patents filed earliest<2009	# EP granted patents filed earliest<2009
animal housings	animal husbandry: equipment for housing animals	A01K 1	5974	851
soil reclamation	reclamation of contaminated soil	B09C 1	2118	608
fusion polypeptide	fusion polypeptide	C07K2319	7847	4049
film sheet manufacturing	manufacture of articles or shaped materials containing macromolecular substances: films or sheets	C08J 5/18	3570	2114
filament manufacturing	general methods for the manufacture of artificial filaments or the like	D01F 1	1383	570
laundry dryers	domestic laundry dryers	D06F 58	2701	839
homopolymers	Homopolymers and copolymers of unsaturated aliphatic hydrocarbons having only one carbon-to carbon double bond	C08F 10	8554	3461
soil-shifting machines	soil-shifting machines	E02F 9	6456	1560
exhaust purifiers	exhaust or silencing apparatus having means for purifying, rendering innocuous, or otherwise treating exhaust for extinguishing sparks	F01N 3/2066	766	629
combustibles supply control	electrical control of supply of combustible mixture or its constituents	F02D 41	18187	6816
gearing control functions	Control functions within control units of changespeed- or reversing-gearings for conveying rotary motion	F16H 61	15253	4463
sensing record carriers	methods or arrangements for sensing record carriers,	G06K 7	11142	3316
object actuated machines	mechanisms actuated by objects other than coins to free or to actuate vending, hiring, coin or paper currency dispensing or refunding apparatus	G07F 7	6799	2214
control signal transmission	arrangements for transmitting signals characterised by the use of a wireless electrical link	G08C 17/02	907	313
speech recognition procedures	procedures used during a speech recognition process, e.g. man-machine dialogue	G10L 15/22	919	223
information storage editing	editing; indexing; addressing; timing or synchronising; monitoring; measuring tape travel	G11B 27	14607	3501
lithium battery manufacturing	manufacturing of secondary cells, accumulators with non-aqueous electrolyte, lithium accumulators	H01M 10/052	5550	2203
predictive video signal coding	methods or arrangements for coding, decoding, compressing or decompressing digital video signals using transform coding in combination with predictive coding	H04N 19/61	5411	1371
electroluminescent light sources	electroluminescent light sources	H05B 33	6933	1848
non-fossil fuels	technologies for the production of fuel of nonfossil origin	Y02E 50	3627	1549

Table 3.B.2: Aggregation of CPC classes and number of unique patents

AGGREGATED CLASS	CPC classes	US patents earliest filed <2009	EP patents earliest filed <2009
AGRICULTURE	A0	199829	62274
FOODSTUFFS & TOBACCO	A21-A24	99283	99283
WEARABLES	A41-A46	158168	31274
FURNITURE & DOMESTIC ARTICLES	A47	198137	35881
HEALTH & WELLBEING	A61-A63	551343	1057394
SEPARATING & MIXING PROCESSES	B01-B09	299444	188607
SHAPING OF MATERIALS	B21-B33	569035	257470
PRINTING & DECORATION	B41-B44	164956	106699
RAILROADS & SHIPS	B61,B63	113111	9208
AVIATION	B64	50379	7350
OTHER, VEHICLES RELATED	B60,B62	364114	76396
PACKING & TRANSPORTING	B65-B68	343185	124825
MICRO & NANOTECHNOLOGY	B81,B82	33116	14804
INORGANIC CHEMISTRY	C01	64289	29767
OTHER CHEMICAL INDUSTRIES	C02-C06,C13,C14	130118	69540
ORGANIC CHEMISTRY	C07	335300	105630
MACROMOLECULES	C08	217371	72606
PAINTS	C09	113624	124642
PETROLEUM, GAS & COKE	C10	80664	31815
BIOCHEMISTRY	C11,C12	110121	190851
METALLURGY	C21-C25	143398	71321
OTHER CHEMISTRY	C30,C40	14622	9095
TEXTILES	D01-D10	174651	63344
PAPER-MAKING	D21	35187	29176
BUILDING	E01-E06	292294	64659
MINING	E21	81643	20124
ENGINES & PUMPS	F01-F05	276811	151200
MECHANICAL ENGINEERING	F15-F17	392610	145528
LIGHTING	F21	40427	8417
HEATING	F22-F28	229621	63076
WEAPONS	F41,F42	61036	11039
MEASURING & TESTING	G01,G04	452562	229884
OPTICS & PHOTOGRAPHY	G02,G03	314480	159094
COMPUTATION & CONTROLLING	G05-G07	493428	200383
SOUND, SIGNALLING & INFORMATION	G08-G16	331302	133285
NUCLEONICS	G21	21952	9357
ELECTRIC ELEMENTS & CIRCUITRY	H01-H03, H05	846273	351941
ELECTRIC POWER GENERATION	H02	148028	52661
TELECOMMUNICATION	H04	456994	429038
CLIMATE CHANGE MITIGATION	Y02	229074	131096

Figure 3.C.1: **Probability plots for the backward link distributions (US patents)** We plot the probability plots for the backward link distributions for the four focus technologies, for the geometric distribution $G(\rho)$, the normal distribution $N(\mu, \sigma^2)$ and the binomial distribution $B(n, p)$. The parameters ρ, μ, σ^2 and p are optimized to obtain the best fit.



technology. Only for the lower path length values of combustion engines this distribution deviates slightly more than the other distributions, yet overall it presents still the best fit. Note that the quality of fits provided by the geometric distributions in Figure 3.C.1 and the distribution from Equation 3.9 in Figure 3.C.2 (both single parameter distributions) is quite remarkable, especially in comparison to the normal distribution, which allows us to fit two parameters instead.

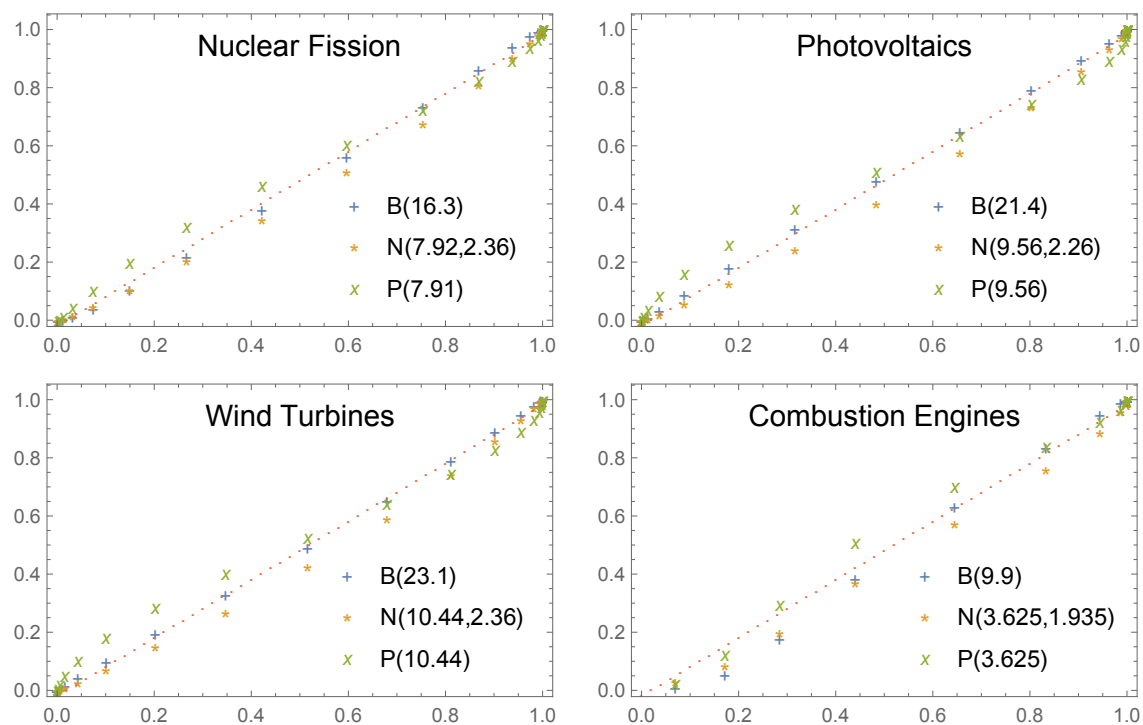
3.D Examiner versus applicant citations

In most patent offices, a citation can be introduced both by the applicant and the patent examiner. The citations added by the applicant are often perceived to be as the better indicator for knowledge flows (Criscuolo & Verspagen, 2008; A. Jaffe et al., 2000), yet the difference between the types of citation is/was not always recorded for each patents office, not the least being USPTO before the year 2000. This research therefore uses a general citation instead to represent a knowledge flow. As a justification for this choice we investigate in this appendix (when possible) the similarity between the knowledge dynamics based on applicant citations ("type APP") to the overall dynamics.

In line with (Azagra-Caro & Tur, 2018; Criscuolo & Verspagen, 2008), we determine for the European patents the type APP citations as those with Patstat's *citn_categ*='D'. For the US patents the type APP citations are selected as those

wind turbines.

Figure 3.C.2: **Probability Plots for the path length distributions (US patents)** We plot the probability plots for the path length distributions for the four focus technologies, for the Poisson distribution $P(\eta)$, the normal distribution $N(\mu, \sigma^2)$ and the binomial type of distribution $B(n')$ from Equation 3.9, (where the values of n' correspond to those in in Figure 3.5.5). The parameters η, μ, σ^2 and are optimized to obtain the best fit. Each distributions is plotted for $n = 3000$, except for wind turbines $n = 2000$.



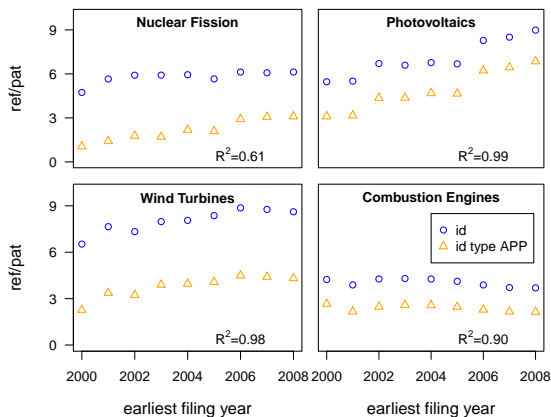


Figure 3.D.1: *Type APP id for US patents* We plot the 'normal' id and the id based on type applicant citations (type APP) for the earliest filing year. Both develop largely similar, which is also reflected by the relatively high correlation coefficients R^2

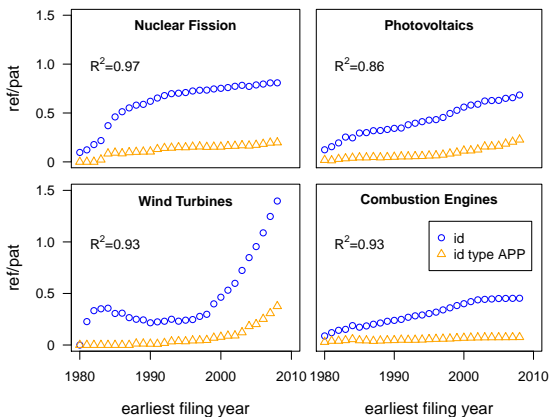


Figure 3.D.2: *Type APP id for EP patents* Similar to Figure 3.D.1 but then for EP patents. Contrary to the US data, the distinction between applicant and examiner citations is recorded for the European data for the entire period of operation of the EPO.

with Patstat's `citn_origin='APP'`.¹⁴ As mentioned earlier, for the US patents we only include data from about 2000 onward, as only after 2000 the distinction was made between examiner and applicant citations by USPTO (EPO, n.d.). In Figures 3.D.2 and 3.D.1 we plot the internal dependence based on Type APP citations both for the US and EP patents for different years. We observe that both dependences develop rather similarly over time, which is confirmed by the high correlation coefficients R^2 for each technology. The type APP id's are consistently a fraction lower as the citations added by the application are a subset of the total citations. Where for the EP patents the type APP citations are about a quarter of the total citations, for the US patents it is about a half. For both cases however the fraction varies somewhat per technology. We therefore systematically compare both id's for the entire set of 24 technologies in Figure 3.D.3. While a power law provides the best fit (as illustrated in 3.D.3), the relation is also rather well fitted by a simple linear relation. Regardless of the exact form, both id's are rather closely (and positively) related across technologies. As the id is closely related to the internal path length, this suggests that we can draw a similar conclusion for the latter indicator.

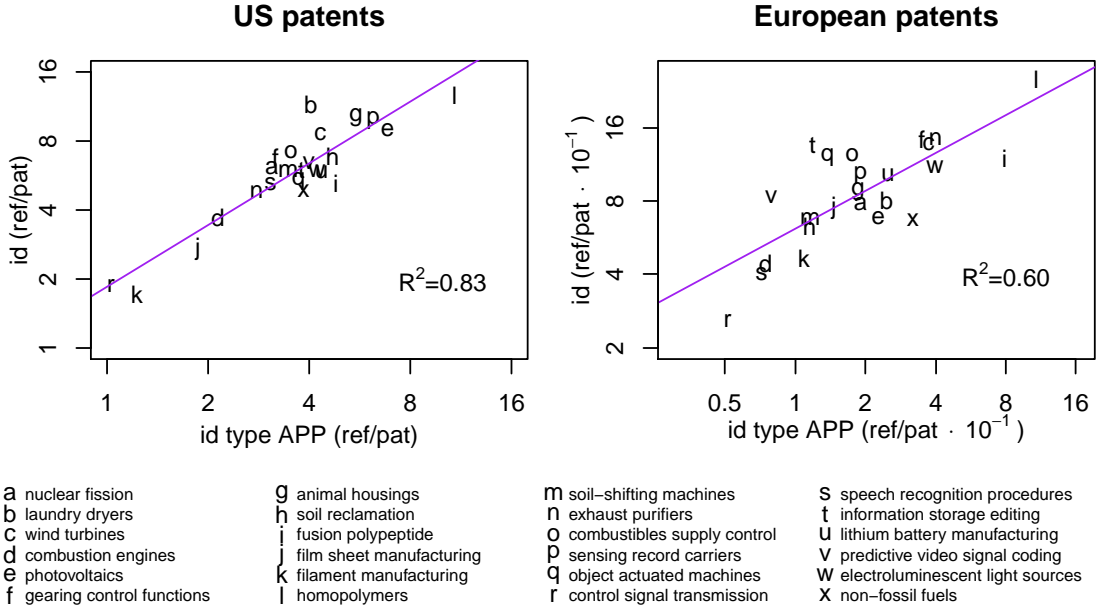
In conclusion, while there is no guarantee that both types of citation result in similar knowledge dynamics, these results suggest that the knowledge dynamics based on type app citations are closely (and positively) related to the knowledge dynamics based on general citations.

3.E Mathematical appendix

In this appendix we explain the number of mathematical derivations appear in Section 3.4.3 in more detail. We start by explaining how the assumption that $n_0 = rn$ is compatible with the geometric distribution found for the number of backward links in Section 3.4.2 if q is small compared to m_0 . Using that the probability to obtain

¹⁴We supplemented the 2019 Patstat dataset with the 2018 Patstat dataset, which contains a more complete recording of the `citn_cat` 'D'.

Figure 3.D.3: ***Id versus id type APP for 24 technologies*** In the left panel we compare the two *id*'s for the US patents and on the right for European patents. Note that both axes are logarithmic, hence the fitted line is a power law.



an initial node is $P_n(m=0) = \frac{1}{qn+m_1}$, we note that the expected number of initial nodes $\langle n_0 \rangle$ after n inventions is

$$\sum_{n'=1}^n P_{n'}(m=0) = \frac{1}{q} H\left(n + \frac{m_1}{q}\right) - \frac{1}{q} H\left(\frac{m_1}{q}\right) \quad (3.12)$$

$$\approx \frac{1}{q} \log\left(n + \frac{m_1}{q}\right) - \frac{1}{q} \log\left(\frac{m_1}{q}\right) \quad (3.13)$$

$$\approx \frac{1}{q} \log\left(1 + \frac{qn}{m_1}\right), \quad (3.14)$$

where we approximated the harmonic numbers $H(n)$ by logarithms. When $\frac{qn}{m_1}$ is small, i.e. when $q \ll m_1$, we can approximate the last expression as $\langle n_0 \rangle \approx \frac{n}{m_1}$. This suggests that, for $q \ll m_1$ we can approximate the coefficient $r \approx \frac{1}{m_1} = \frac{1}{m_0+1}$.

Next we discuss the steps leading to Equation 3.4, 3.5, 3.6 and 3.10. To see that the expression in Equation 3.4 satisfies Equation 3.3, first note that $\binom{n}{k} = 0$ for $k > n$ and that $r \binom{n}{1} = rn$, as the initial conditions require. Then, start from the recursive property of the binomial coefficient $\binom{n+1}{k+1} - \binom{n}{k+1} = \binom{n}{k}$ and multiply left and right by rq^k . We then obtain

$$rq^k \binom{n+1}{k+1} - rq^k \binom{n}{k+1} = rqq^{k-1} \binom{n}{k} \quad (3.15)$$

$$f_k(n+1) - f_k(n) = qf_{k-1}(n), \quad (3.16)$$

which is Equation 3.3. To sum $f_k(n)$ from k to n (we need not sum further as all $f_k(n) = 0$ for $k > n$), we can use the binomial theorem $\sum_{k=0}^n \binom{n}{k} x^k y^{n-k} = (x+y)^n$, which counts for any real (or complex) number x and y . Taking $x = q$ and $y = 1$

we get

$$\sum_{k=0}^n \binom{n}{k} q^k = (q+1)^n \quad (3.17)$$

$$\sum_{k=1}^n \binom{n}{k} q^k = (q+1)^n - 1 \quad (3.18)$$

$$r \sum_{k=0}^n \binom{n}{k+1} q^{k+1} = r(q+1)^n - r \quad (3.19)$$

$$\sum_{k=0}^n f_k(n) = \frac{r(q+1)^n - r}{q}. \quad (3.20)$$

To obtain the expression in Equation 3.5 we divide Equation 3.4 by the right-hand side of Equation 3.20. To obtain Equation 3.6, we have to calculate $\langle k \rangle = \sum_{k=0}^n k \tilde{f}_k(n)$. We first calculate $\sum_{k=0}^n k f_k(n) = r \sum_{k=0}^n k \binom{n}{k+1} q^k$ instead. Note that this is exactly the expression we get if we differentiate $r \sum_{k=0}^n \binom{n}{k+1} q^k$, i.e. the left-hand side of Equation 3.20, with respect to q and then multiply once by q . Equivalently, we do this to the right-hand side of Equation 3.20, obtaining

$$\sum_{k=0}^n k f_k(n) = q \frac{\partial}{\partial q} \frac{r(q+1)^n - r}{q} = r \left(n(1+q)^{n-1} - \frac{(1+q)^n - 1}{q} \right). \quad (3.21)$$

To obtain $\sum_{k=0}^n k \tilde{f}_k(n)$ from $\sum_{k=0}^n k f_k(n)$, we divide simply by the total number of paths (the right-hand side of Equation 3.20),

$$\sum_{k=0}^n k \tilde{f}_k(n) = \frac{qn(1+q)^{n-1} - 1}{(1+q)^n - 1} - 1 \quad (3.22)$$

$$= \frac{nq}{q+1} \cdot \frac{1}{1 - (1+q)^{-n}} - 1. \quad (3.23)$$

For large n , this expression quickly goes to $\frac{nq}{q+1} - 1$. Hence we obtain Equation 3.6 for $k_0 = -1$.

Finally, we note that solving the expression $f'_{k+1}(n'+1) - f'_{k+1}(n) = f'_k(n')$ goes completely analogous, (where $f \rightarrow f'$ and $n \rightarrow n'$), except that there is no coefficient q (or the coefficient is '1'). Equivalently, therefore, we redo the approach in this appendix and do the substitutions $f \rightarrow f'$, $n \rightarrow n'$ and $q \rightarrow 1$. Substitutions in Equation 3.5 and 3.6 directly lead to respective Equations 3.9 and 3.10. Note that the initial conditions change accordingly: (a) with a maximum speed $v = 1/\delta n$, $f'_k(n') = 0$ for $k > nv = n'$ and (b) at n' nodes we expect to have $n'r$ initial nodes.

3.F Data scripts

To facilitate the reproducing of key findings and conclusions of our manuscript, and to enable building upon our work, we below provide information on how our data set was created. The provided T-SQL scripts are to be used in combination with the Patstat patent database, which is available for licensing by the European Patent

Office (EPO). While our paper used the Patstat 2018 and 2019 autumn release, the scripts can be modified to be used with other Patstat releases.

The below overview includes the queries used to retrieve the relevant data sets. The technologies were selected on two levels of classification, (A) the CPC group / subgroup level and (B) the CPC class level (where some classes were aggregated). For a table with the exact coding of these groups and classes, see Section 3.B.

3.F.1 Technology selection using CPC group/sub group level

The below query selects patents for defined technologies on the CPC group or subgroup level (the provided example is for EP patents for the group “F16H 61”)

```
select distinct a.docdb_family_id,
MIN(a.earliest_filing_date) as earliest_date,
MIN(a.earliest_filing_year) as earliest_year
into F16H_61_EP
from Patstat2019b.dbo.tls201_appln as a
inner join Patstat2019b.dbo.tls224_appln_cpc as d
on a.appln_id=d.appln_id
where a.earliest_filing_year<2009
and a.granted='Y'
and a.appln_auth='EP'
and a.appln_kind='A'
and d.cpc_class_symbol like 'F16H 61/%'
group by a.docdb_family_id
order by earliest_date
```

As discussed in the methodology section of the paper, an ‘EP patent’ in our study implies a DOCDB family with at least one EP member. The internal citations for EP patents are determined by considering the references in the EP members of a technology X to other patents in X, the references thereby need not be to other EP members per se. When there are multiple of such references between two families, one reference is counted. The above is performed by the following query (the provided example is for EP patents group “F16H 61”).

```
select distinct a.docdb_family_id as cited_family,
b.docdb_family_id as citing_family,
min(a.earliest_filing_date) as earliest_date_cited,
min(b.earliest_filing_date) as earliest_date_citing
into F16H_61_EP_citations
from Patstat2019b.dbo.tls201_appln as a
inner join Patstat2019b.dbo.tls228_docdb_fam_citn as c
on a.docdb_family_id=c.cited_docdb_family_id
inner join Patstat2019b.dbo.tls201_appln as b
on c.docdb_family_id=b.docdb_family_id
inner join Patstat2019b.dbo.tls224_appln_cpc as d
on a.appln_id=d.appln_id
inner join Patstat2019b.dbo.tls224_appln_cpc as e
```

```

on b.appln_id=e.appln_id
inner join Patstat2019b.dbo.tls211_pat_publn as f
on b.appln_id=f.appln_id
inner join Patstat2019b.dbo.tls212_citation as g
on f.pat_publn_id=g.pat_publn_id
inner join Patstat2019b.dbo.tls211_pat_publn as h
on h.pat_publn_id=g.cited_pat_publn_id
inner join Patstat2019b.dbo.tls201_appln as i
on h.appln_id=i.appln_id
where
i.docdb_family_id=c.cited_docdb_family_id
and a.earliest_filing_year<2009
and b.earliest_filing_year<2009
and a.granted='Y'
and b.granted='Y'
and a.appln_auth='EP'
and b.appln_auth='EP'
and a.appln_kind='A'
and b.appln_kind='A'
and e.cpc_class_symbol like 'F16H 61/%'
and d.cpc_class_symbol like 'F16H 61/%'
group by a.docdb_family_id, b.docdb_family_id
order by earliest_date_citing, earliest_date_cited

```

3.F.2 Technology selection using CPC class level

In order to select technologies on the CPC class level, the below query first creates a list that assigns each DOCDB family to one or more CPC classes. The provided example is for EP patents.

```

select distinct a.docdb_family_id,
MIN(a.earliest_filing_year) as earliest_year,
LEFT(d.cpc_class_symbol, 3) as cpc_class
into all_nod_EP_class
from Patstat2018b.dbo.tls201_appln as a
inner join Patstat2018b.dbo.tls224_appln_cpc as d
on a.appln_id=d.appln_id
where a.earliest_filing_year<2009
and a.granted='Y'
and a.appln_auth='EP'
and a.appln_kind='A'
group by a.docdb_family_id
order by earliest_year

```

The resulting dataset can be used to count the number of internal citations for a given class. The following query does so specifically for the combined class “Other, vehicle related” (combination of CPC classes B60 and B62), for EP patents:

```

select distinct count(distinct a.docdb_family_id) as int_references,
b2.docdb_family_id, b2.earliest_year
into OTH_VEH_EP
from all_nod_EP_class as a
inner join Patstat2019b.dbo.tls228_docdb_fam_citn as c
on a.docdb_family_id=c.cited_docdb_family_id
inner join Patstat2019b.dbo.tls201_appln as b
on c.docdb_family_id=b.docdb_family_id
inner join all_nod_EP_class as
on b.docdb_family_id=b2.docdb_family_id
inner join Patstat2019b.dbo.tls211_pat_publn as f
on b.appln_id=f.appln_id
inner join Patstat2019b.dbo.tls212_citation as g
on f.pat_publn_id=g.pat_publn_id
inner join Patstat2019b.dbo.tls211_pat_publn as h
on h.pat_publn_id=g.cited_pat_publn_id
inner join Patstat2019b.dbo.tls201_appln as i
on h.appln_id=i.appln_id
where
i.docdb_family_id=c.cited_docdb_family_id
and (a.cpc_group like 'B60%'
or a.cpc_group like 'B62%')
and (b2.cpc_group like 'B60%'
or b2.cpc_group like 'B62%')
group by b2.docdb_family_id,
b2.earliest_year,b2.cpc_group

```

3.F.3 Selecting applicant-added references

Finally, to identify all references added by the applicant, citations are selected with `citn_origin` “APP” and/or those with `citn_categ` “D”. As noted in the supplementary material, we complemented the data from the Patstat 2019 edition with data from the Patstat 2018 autumn edition, which was found to be more complete. In the example below the citations are selected for the 2018 edition, specifically for EP patents group “F16H 61”:

```

select distinct a2.docdb_family_id as cited_family,
a2.earliest_date as cited_earliest_date,
b2.docdb_family_id as citing_family,
b2.earliest_date as citing_earliest_date,
b2.earliest_year as citing_earliest_year
into F16H_61_EP_citations_type_app
from F16H_61_EP as a
inner join Patstat2018b.dbo.tls228_docdb_fam_citn as c
on a.docdb_family_id=c.cited_docdb_family_id
inner join Patstat2018b.dbo.tls201_appln as b

```



```

on c.docdb_family_id=b.docdb_family_id
inner join F16H_61_EP as b2
on b.docdb_family_id=b2.docdb_family_id
inner join Patstat2018b.dbo.tls211_pat_publn as f
on b.appln_id=f.appln_id
inner join Patstat2018b.dbo.tls212_citation as g1
on f.pat_publn_id=g1.pat_publn_id
inner join Patstat2018b.dbo.tls211_pat_publn as h
on h.pat_publn_id=g1.cited_pat_publn_id
inner join Patstat2018b.dbo.tls201_appln as i
on h.appln_id=i.appln_id
inner join Patstat2018b.dbo.tls211_pat_publn as f2
on b.appln_id=f2.appln_id
inner join Patstat2018b.dbo.tls212_citation as g2
on f2.pat_publn_id=g2.pat_publn_id
inner join Patstat2018b.dbo.tls215_citn_categ as j
on j.pat_publn_id=g2.pat_publn_id
inner join Patstat2018b.dbo.tls211_pat_publn as h2
on h2.pat_publn_id=g2.cited_pat_publn_id
inner join Patstat2018b.dbo.tls201_appln as i2
on h2.appln_id=i2.appln_id
where
and b2.earliest_filing_year<2009
and b.appln_auth='EP'
and ((g1.citn_origin='APP' AND
i.docdb_family_id=c.cited_docdb_family_id) or
(i2.docdb_family_id=c.cited_docdb_family_id AND
j.citn_id=g2.citn_id AND j.citn_categ like '%D%'))
group by b2.docdb_family_id,
b2.earliest_date,a.docdb_family_id,
a.earliest_date,b2.earliest_year
order by b2.earliest_date, a.earliest_date

```


Chapter 4

Cumulative structure and path length in knowledge networks

Peter Persoon *This chapter is currently being prepared for submission to a journal*

Abstract

An important knowledge dimension of science and technology is the extent to which their development is cumulative, that is, the extent to which later findings build on earlier ones. Knowledge structures can be studied using a network approach in which nodes represent findings and links represent knowledge flows. This network approach allows us to use the notion of network paths and path length to study cumulative knowledge structures. Starting from the Price model of network growth, we derive an exact solution for the path length distribution of all unique paths from a given initial node to each node in the network. We study the relative importance of the average in-degree and cumulative advantage effect and implement a generalization where the in-degree depends on the number of nodes. The cumulative advantage effect is found to fundamentally slow down path length growth. As the collection of all unique paths may contain many redundancies, we additionally consider the subset of the longest paths to each node in the network. As this case is more complicated, we only approximate the longest path length distribution in a simple context. Where the number of all unique paths of a given length grows unbounded, the number of longest paths of a given length converges to a finite limit, which depends exponentially on the given path length. Fundamental network properties and dynamics therefore characteristically shape cumulative structures in those networks, and should therefore be taken into account when studying those structures.

4.1 Introduction

Science and technology advance when scientists and inventors learn from earlier findings and use this knowledge to create new findings. A key element of theories of knowledge development is therefore the cumulative nature of discovery and invention (Basalla, 1989; Dean et al., 2014; Freeman & Soete, 1997; Trajtenberg et al., 1997), i.e. the building of new knowledge on earlier knowledge. A better understanding of this phenomenon may provide insight into what knowledge development needs to flourish, and how knowledge structures can be built robustly (Albert & Barabasi, 2002; Albert et al., 2000). Furthermore, a general understanding of cumulative knowledge structures can provide a framework to study how different fields or disciplines of knowledge vary in this dimension, which may help explain variations found across these fields in other knowledge dimensions. In the specific context of technological knowledge, for example, the 'cumulateness of knowledge' is conjectured to closely relate to the appropriability of that knowledge, as well as to the difficulty by which knowledge travels geographically (Breschi et al., 2000; Malerba & Orsenigo, 1996; Nelson & Winter, 1982). Understanding how cumulative structures develop is therefore not only relevant from a theoretical perspective, but of great importance as well to targeted science and technology policies aiming to strengthen the development of particular fields.

Approaches to cumulative knowledge structures that aim for a quantitative description may benefit from a network perspective on knowledge. In this perspective, nodes represent findings (which can be any element of knowledge, but usually a scientific finding or an invention) and links represent knowledge connections (indicating that a finding builds on another finding, i.e. knowledge flow in the system). While this may sound abstract, this perspective can, given some limitations¹, be approached empirically using data about publications and citations (Garfield, 1979; Price, 1965b; Trajtenberg, 1990). Many contributions studying knowledge networks in this fashion use - or are variations on - a model introduced by Price (Price, 1976). In this model, nodes are more likely to connect to nodes that already have a large number of knowledge connections, referred to by Price as the 'cumulative advantage effect'² also known as, in the context of un-directed links, 'preferential attachment' (Barabasi & Albert, 1999). In many applications of the Price model, the focus is on degree distributions, which describe how outgoing or incoming links are distributed over nodes (Barabasi & Albert, 1999; Steinbock et al., 2019; D. Wang et al., 2013). While these distributions to an important extent determine network structures, they are mainly revealing for the variation in the relative importance of nodes, and perhaps less useful to study to what extent there is knowledge flow in such networks. Yet these knowledge flows are an essential element of cumulative structures, in which findings build on findings, which build on other findings, etc. It may therefore be more useful to focus instead on the extent to which sequences of findings appear, which are defined naturally by the well-studied notions of network paths and path

¹For example, not all citations may represent knowledge flow. While acknowledging these limitations, we will not go into that discussion here. For an overview in the context of scientific citations see (Bar-Ilan & Halevi, 2017; Catalini et al., 2015) or patent citations see (Alcácer & Gittelman, 2006; Duguet & MacGarvie, 2005)

²The term 'cumulative' in this expression, coined by Price, simply means 'added up', and differs from earlier used meaning in 'cumulative knowledge structures', where it suggest the characteristic aspect of knowledge building on knowledge

length (Katzav et al., 2015; Newman, 2010; Watts & Strogatz, 1998). Yet, where most studies of network paths focus on distance metrics based on considering the *shortest* paths in the network (Caravenna et al., 2019; Dereich et al., 2012, 2017; Dommers et al., 2010), that choice is not at all obvious for knowledge networks. The shortest paths could be misleading in the context of cumulative structures, where one might want to take into account all necessary intermediate steps of development (Evans et al., 2020; Hu et al., 2011; Martinelli & Nomaler, 2014), which may not be included in the shortest paths.

As an alternative, one might therefore consider metrics based on the *longest* paths instead (see Figure 4.1.1), the length of which necessarily represents the maximum number of intermediate developmental steps. Yet, if we limit the analysis to the longest (or shortest) path between two findings, we ignore that there may be more paths between these findings, which may describe equally relevant sequences of developmental steps. Indeed a key element of invention and discovery is exactly the combination (or sometimes 'recombination') of different ideas (Arthur, 2009; Kaplan & Vakili, 2015; Strumsky & Lobo, 2015), which may be drawn from different sequences of development. To account for these, we may as another alternative consider metrics based on *all unique* paths (for an illustration see Figure 4.1.1), for example, the average length of these paths. A downside of considering all paths is that, especially when the average degree is large, there may be many paths between two findings, and not all of these may represent distinct knowledge flows leading to distinct recombined ideas. For example, when two paths leading to a finding largely overlap, the content conveyed in the knowledge flow they represent may largely be the same, and considering them separately is largely a redundant effort. As both alternatives therefore have advantages as well as disadvantages, it may be useful to consider both of them to study cumulative structures.

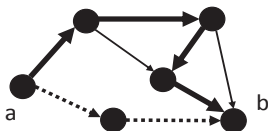


Figure 4.1.1: **Types of paths** Between node *a* and *b* we can distinguish between the shortest path (dashed links), the longest path (fat links) and all unique paths (the paths formed by dashed, fat or thin links or any combination thereof).

It is however not immediately clear how, in the context of knowledge networks, the metrics based on shortest paths can be generalized for the longest paths or all unique paths. Starting from the Price model, Evans et al. make an important contribution, deriving a lower bound for the length of the longest path in a network (Evans et al., 2020). While this is insightful about the longest stretch of knowledge flow in a network, as we argued earlier, there are usually many more paths in a network, some of them representing equally interesting sequences of findings. The longest path with length l might be exceptional, begging the question of how many paths there are of length $l - 1$, $l - 2$ etc, i.e. how the number of paths is distributed over various lengths.

A detailed understanding of the path length distributions in knowledge networks allows us to form well-founded expectations of the typical stretch of knowledge flows in cumulative structures and is therefore key to interpret variation in these structures across different scientific disciplines or technologies. In this contribution, we therefore explore the typical path length distributions we might encounter in knowledge networks, and how we can use these to calculate metrics such as the expected path length. We

will in a first way of counting network paths consider the distribution of all unique paths coming from a given initial node (see Section 4.2). Following the Price model, we thereby investigate the role of the cumulative advantage effect. Motivated by recent results which indicate that the average degree, which is usually kept constant, may in fact steadily increase with the number of nodes in a knowledge network (P. Persoon et al., 2021), we consider a generalization of the model allowing for this increase. In a second way of counting paths (see Section 4.3), we focus on a subset of all unique paths, by selecting only the *longest paths* from the initial node to each node in the network. As deriving an exact solution for this distribution is challenging, we will approximate it instead, thereby ignoring the cumulative advantage effect. Though simplified, this allows us to derive the main characteristics of the distribution, approximate the expected longest path length and compare it to the case of all unique paths.

4.2 All unique paths in the Price model

For each discrete step in time n , the Price model generates a directed acyclic graph $G(n)$ consisting of N nodes and M links (Price, 1976). Starting from some initial acyclic graph $G(1)$, at each step in time, a new node is added to the network, which is connected with incoming links to an average of $\langle m \rangle$ existing nodes in the network. The number of incoming links of a node l in $G(n)$, i.e. its in-degree, therefore does not change as n increases, yet the number of outgoing links of l , i.e. its out-degree, is however expected to gradually increase with n . In the context of knowledge networks, the incoming links of a node l represent the set of knowledge connections appearing at once when l is introduced (i.e. published, patented), hence l can be interpreted to 'build on' the set of nodes to which it is connected by the incoming links. Reversely, the set of nodes to which l is connected by its outgoing links can be interpreted to build on l . Note this implies that the links, (and thus the paths), are in the direction of knowledge flow, which is a convention in line with Evans, yet opposite to a number of others (Newman, 2010; Steinbock et al., 2019; Vazquez, 2001). In most applications of the Price model, it is assumed that the average in-degree $M/N = \langle m \rangle$ is approximately constant as the network grows.

In this contribution, our initial graph $G(1)$ consists of a single 'initial node' 1 and we number the subsequent nodes by the order of appearance: 2, 3, ..., n , hence at any time, $N = n$. While this choice for an initial graph allows for a simple description of the growth process, it also introduces two subtleties. First, for $n = 1$ there are no other nodes to connect to, so insisting that $\langle m \rangle > 0$ at that point appears problematic. As an exception, we will allow this node (and only this node) to connect to itself. Second, especially when n is small and $\langle m \rangle$ is large, new nodes may not have enough distinct nodes to connect to. Therefore, we allow multiple linkages to the same node, which should occur more rarely when the network becomes larger. We refer to Evans (Evans et al., 2020) for a more elaborate discussion of these subtleties.

In the Price model, the probability $\Pi(n, l)$ for a new node $n + 1$ to connect to an existing node l consists of two parts: (i) a part which is non-zero and equal for all nodes, and (ii) a part which is proportional to the out-degree $h(l, n)$ of l . Introducing the constant $c \geq 0$ which represents the strength of effect (ii) in proportion to effect

(i)³, we can thus write

$$\Pi(l, n) = \frac{1 + ch(l, n)}{\sum_t^n 1 + ch(t, n)} = \frac{1 + ch(l, n)}{n + c\langle m \rangle n}, \quad (4.1)$$

hence note that when $c = 0$, the 'cumulative advantage effect' is switched off, and we are left with the neutral case where new nodes link equally likely to any node in $G(n)$. For simplicity we will in this work only consider the paths in $G(n)$ starting from the initial node (in the Section 4.5 we discuss some generalizations of this choice), so when we mention in the following 'a path to node l ' we mean a unique path from the initial node to node l . The number of paths $f_k(n)$ are likewise defined as the total number of unique paths of length k in $G(n)$ starting from the initial node. We assume there is a single path of length zero from the initial node to itself, i.e. $f_0 = 1$ for all n , though this largely a matter of convention. We will derive an expression for the expected value $\langle f_k(n) \rangle$, yet for brevity we drop the $\langle \rangle$ notation, also for $\langle m \rangle$. Let $q_k(l)$ be the number of paths to node l with length k , hence when a new node connects to l , there are $q_k(l)$ new paths of length $k + 1$. The expected increase in the number of paths of length $k + 1$ is therefore

$$\Delta_n f_{k+1}(n) = m \sum_{l=1}^n q_k(l) \Pi(l, n) = m \sum_{l=1}^n \frac{q_k(l) + cq_k(l)h(l, n)}{(1 + cm)n}. \quad (4.2)$$

We note that each of the $q_k(n)$ paths going through l extend into $q_k(n)h(l, n)$ paths of length $k + 1$, therefore $\sum_l h(l, n)q_k(n) = f_{k+1}(n)$ and using that $\sum_l q_k(l) = f_k(n)$, we obtain

$$\Delta_n f_{k+1}(n) = m \frac{f_k(n) + cf_{k+1}(n)}{(1 + cm)n}. \quad (4.3)$$

Additionally, we have the initial condition that $f_k(1) = 0$ for all $k > 0$ as there are no paths of length $k > 0$ when $n = 1$. Before we discuss the general solution, let us focus briefly on the simple neutral case where we exclude the cumulative advantage effect.

4.2.1 Excluding the cumulative advantage effect

Excluding the cumulative advantage effect amounts to setting $c = 0$. Equation 4.3 then becomes $\Delta_n f_{k+1}(n) = mf_k(n)/n$. Noting that $f_0 = 1$, this basic relation is directly solved by

$$f_k(n) = \frac{m^k}{\Gamma(n)} S(n, k + 1) \quad (4.4)$$

where $\Gamma(n)$ is the gamma function and $S(n, k)$ is the n, k^{th} unsigned Stirling number of the first kind. The latter appear as coefficients in the rising factorial of a real number x to height n , defined in mathematics as

$$x^{\overline{n}} = x(x + 1)\dots(x + n - 1) = \sum_{k=0}^n S(n, k)x^k. \quad (4.5)$$

³We note that in the original model of Price, $c = 1$ and in the approach by Evans, the parameter $p = cm/(1 + cm)$ is instead introduced.

Stirling numbers can be expressed in terms of harmonic numbers and generalized harmonic numbers (Adamchik, 1997), for example allowing us to write for $f_1(n) = mH(n-1)$, where $H(n)$ is the n^{th} harmonic number. When n gets large, the leading term of $S(n, k+1)/\Gamma(n)$ is approximately $\log(n)^k/\Gamma(k+1)$ (Wilf, 1993). For large n , the number of paths of length k can for large n therefore be approximated as $m^k \log(n)^k/\Gamma(k+1)$, which we can recognize as a (not normalized) Poisson distribution of the variable k .

Using Equation 4.5 we can derive the expected total number of paths $K(n) = \sum_k f_k(n)$, to equal

$$K(n) = \frac{\Gamma(m+n)}{\Gamma(n)\Gamma(m+1)} \quad (4.6)$$

This expression increases approximately as n^m . To obtain the expected path length $\ell(n) = \sum_k k f_k(n)/K(n)$ note that we can differentiate $K(n)$ with respect to m and multiply by $m/K(n)$, resulting in

$$\ell(n) = m\psi(m+n) - m\psi(m+1), \quad (4.7)$$

where $\psi(m+n)$ is the digamma function, which increases logarithmically in n . We conclude therefore that the expected path length of all unique paths increases logarithmically with the number of nodes n , along with a coefficient m . To be able to compare this relation to later cases we can denote it more generally as

$$\ell(n) \approx d_m \log(n) + \ell_1, \quad (4.8)$$

where the coefficient d_m is some constant depending on m and ℓ_1 is another constant we are less interested in. For the case where there is no cumulative advantage effect we therefore have $d_m = m$.

4.2.2 Including the cumulative advantage effect

For general values of c the analysis becomes slightly more complicated. Going back to Equation 4.3, let us start by writing down the general solution (we refer to the supplementary material for a detailed derivation):

$$f_k(n) = \frac{1}{\Gamma(n)(-c)^k} \sum_{t=k+1}^n (t-k) S_p(n, t) (-p)^{t-1}. \quad (4.9)$$

where $p = cm/(1+cm)$ and $S_y(n, t)$ is the n, t^{th} non-central unsigned Stirling number of the first kind (Koutras, 1982; M. D. Schmidt, n.d.), which are defined for any real y by a slight variation of Equation 4.5, namely $x^{\bar{y}} = \sum_{k=0}^n S_y(n, k)(x-y)^k$ and in particular $S_0(n, k) = S(n, k)$. Note that for $c \rightarrow 0$, we have $p \rightarrow 0$, $p/c \rightarrow m$ and the only member in the sum of Equation 4.9 not going to zero is the first term $(p/c)^k S_p(n, k+1) \rightarrow m^k S(n, k+1)$, thus retrieving the solution for $c = 0$. We plot the distribution, for a number of values of m and c , in Figure 4.2.1 (left two panels), including the case $c = 0$. We observe the distributions for greater c are more skewed towards lower path length values, it appears therefore that the cumulative advantage effect tempers the path length growth. Specifically considering

$$f_1(n) = \frac{1}{c} \left(\frac{\Gamma(p+n)}{\Gamma(p+1)\Gamma(n)} - 1 \right), \quad (4.10)$$

we see that $f_1(n)$ is initially smaller than $mH(n-1)$ (i.e. the value for $f_1(n)$ when $c=0$), yet for a given n , it will overtake $mH(n-1)$ and subsequently grow much larger. Where in the limit of large n , $mH(n-1)$ increases logarithmically, the expression in Equation 4.10 increases as n^p . We can show that the $f_k(n)$ for $k > 1$ show similar behavior. This leads us to the conclusion that, up to a given length k , there are many more paths when there is a cumulative advantage effect, yet beyond that length k , there are actually fewer paths (compared to the $c=0$ case). In other the words, there tend to be more shorter paths when there is a cumulative advantage effect. Finally, in the supplementary material we show that the leading order of $f_k(n)$ for large n can be approximated as

$$f_k(n) \approx \left(\frac{p}{c}\right)^k \frac{\Gamma(n+p)}{p\Gamma(1+p)\Gamma(n)\Gamma(k)} \log\left(\frac{n+p}{1+p}\right)^{k-1}, \quad (4.11)$$

which, up to a factor depending on n , we may again recognize as a (not normalized) Poisson distribution of the variable k .

Again summing $f_k(n)$ over all k , we obtain for the total number of paths

$$K(n) = \frac{\Gamma(m_c+n)}{(1+c)\Gamma(n)\Gamma(m_c+1)} + \frac{c}{1+c}, \quad (4.12)$$

where $m_c = m(1+c)/(1+cm)$. For large n we can conclude this expression grows approximately as n^{m_c} . Note that $m_c < m$ for $m > 1$, hence the power of n by which the number of paths increase is here smaller than the one derived in the $c=0$ case. In line with the observations with Figure 4.2.1, the cumulative advantage effect thus slows down the growth of the number of paths for $m > 1$. However, when $0 < m < 1$, m_c is actually larger than m , hence, in that case, the cumulative advantage effect somewhat accelerates the growth of the number of paths. This effect, apart from the fact that $m < 1$ may be rather uncommon in knowledge networks, is however limited: rewriting m_c as $1 - \frac{1-m}{1+cm}$, we see that, given $0 < m < 1$, it will still always be smaller than 1 for any c . Therefore, we conclude that m alone determines whether the number of paths increases faster than linear or not. We can divide $f_k(n)$ by $K(n)$ to obtain the normalized path length distributions, which we depict for a number of values in Figure 4.2.1 (right two panels). In line with the observations for the not-normalized distribution, these plots indicate that the shorter paths are more probable for lower m and greater c .

To obtain the expected path length $\ell(n)$, we show in the supplementary material how $K(n)$ can with a minor adaptation be approached as a generating function, which allows us to straightforwardly calculate $\ell(n) = \sum_k k f_k(n)/K(n)$, resulting in

$$\ell(n) = \frac{1+c+m_c\psi(m_c+n)-m_c\psi(m_c+1)}{cr(n)+1+c} - \frac{1}{1+c} \quad (4.13)$$

where $r(n) = (K(n) - c/(c+1))^{-1}$. In the limit of large n , $r(n) \rightarrow 0$. We can then approximate

$$\ell(n) \approx \frac{m_c\psi(m_c+n)-m_c\psi(m_c+1)+c}{1+c}. \quad (4.14)$$

This again shows that the expected path length increases logarithmically in n . The only difference with the $c = 0$ case is that the coefficient of $\psi(m_c + n)$, i.e. d_m , is here $m_c/(1 + c)$ instead of m . Noting that $m_c/(1 + c) = m/(1 + cm) < m$ for any $m > 0$, we conclude that, compared to the $c = 0$ case, the cumulative advantage effect slows down the development of the expected path length by a factor proportional to c . Furthermore, the cumulative advantage effect puts an upper limit on d_m of value $1/c$ (which is reached only for very large in-degree). This upper limit is therefore lower when the cumulative advantage effect is greater. Note that the upper limit on d_m disappears only when $c = 0$.

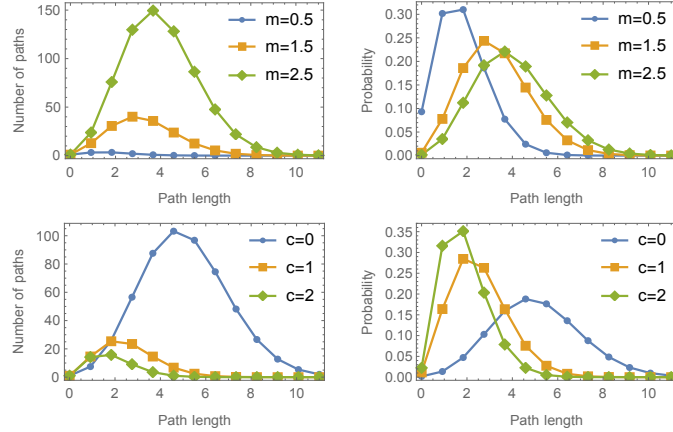


Figure 4.2.1: **Distribution of path lengths** In the left two panels we plot the distribution of the number of paths for various values of m and c , in the right two panels we plot the same distributions but then normalized for the number of paths. Unless otherwise specified, $n = 80$, $m = 1.5$ and $c = 0.5$. We observe that there are less paths for lower m and greater c and that the distributions are more skewed to lower path length values for lower m and greater c .

4.2.3 Generalization for increasing in-degree

Finally we discuss an extension of the model where we allow the in-degree $\mu(n)$ of node n to depend on n , i.e. considering a number of expressions for $\mu(n)$. Equation 4.2 then becomes

$$\Delta_n f_{k+1}(n) = \mu(n) \frac{f_k(n) + c f_{k+1}(n)}{n + c \sum_{l=1}^n \mu(l)}. \quad (4.15)$$

When we take $\mu(n)$ to be any linear combination of integer or non-integer powers of n , which is finite and positive for all n and in which the largest power of n has an exponent $\alpha > 0$, then in the limit of large n , Equation 4.15 reduces to

$$\Delta_n f_{k+1}(n) \approx (1 + \alpha) \frac{f_k(n) + c f_{k+1}(n)}{cn}. \quad (4.16)$$

This equation is similar to Equation 4.3 if we make the substitution $(1 + \alpha)/c = m/(1 + mc)$. For large n , we therefore have the same dynamics as in earlier model with $m = -\frac{\alpha+1}{c\alpha}$ (where $c \neq 0$). This substitution may at first seem odd, as when m was interpreted as the average in-degree, it was restricted to positive values. This assumption was used mainly to interpret the results however, and we see that as long as $m_c > 0$, the derivation leads to the same equations for negative m . In fact we obtain perfectly acceptable results when, using $m = -\frac{\alpha+1}{c\alpha}$, we note that the parameter p (appearing in Equation 4.9) becomes $\alpha + 1$ and m_c , (appearing in Equation 4.12) becomes $(\alpha + 1)(1 + c)/c$. Recalling that m_c is the power of n by which the total number of paths increase, we thus conclude that the smaller the cumulative advantage effect, the stronger the number of paths increase, but at least by a power $\alpha + 1$. For the expected path length we similarly conclude that the

coefficient $d_m = m_c/(1+c)$ appearing in Equation 4.14 becomes $(\alpha+1)/c$. We therefore conclude the expected path length still increases logarithmically in n , yet with a coefficient d_m which is (a) proportional to the largest power of n appearing in $\mu(n)$ and (b) inversely proportional to the strength of the cumulative advantage effect.

In the above generalization the assumption that $c \neq 0$ is rather crucial. As is shown in detail in (P. Persoon et al., 2021), the situation becomes rather different with $c = 0$ and $\mu(n) \propto n$. The number of paths then increases exponentially in n and the expected path length increases linearly in n .⁴ It can be demonstrated that when $\mu(n)$ grows faster than linear in n for $c = 0$, the number of paths increases even faster than exponentially, and likewise the expected path length increases even faster than linear in n . This suggests therefore that the cumulative advantage effect plays a crucial role in keeping the number of paths a power of n and the expected path length a logarithmic relation in n , thus fundamentally slowing down the path length dynamics for the case that $\mu(n)$ increases with n . Only when $\mu(n)$ increases *even* faster in n , namely exponentially, the sum appearing in the denominator of Equation 4.15 will be proportional to $\mu(n)$, thus leading for large n to the relation $\Delta_n f_k(n) \propto f_k(n) + c f_{k+1}(n)$, which can be demonstrated to result in expected path length growth linear in n . We conclude that, in order to break through the 'logarithmic barrier' imposed by the cumulative advantage effect, the in-degrees need to grow at least exponentially with the number of nodes.

4.3 Sub-selecting the longest paths

In Section 4.2 we derived that, when the average in-degree is larger than 1, the number of paths in the network increases rather fast. In the context of knowledge networks, not all of these paths may represent relevant knowledge flows, and there will be many redundancies when each unique path is considered separately. It may therefore make sense to focus instead for each node l on the *longest path* from the initial node to l . We will call these paths in the following 'longest paths', yet they should not be confused with the single, unique longest path in the whole network, which is the subject of work by Evans (Evans et al., 2020).

Note that the longest path from the initial node to a node l may not be unique. In the following, we will however assume we just choose one longest path from the initial node to each node in $G(n)$ and we are interested in deriving an expression for the number $f_k(n)$ of such longest paths of length k . As before we have $f_0 = 1$ for all n . For simplicity we will focus in this derivation on the $c = 0$ case and keep m constant, we leave those generalizations for later work.

We start by noticing that, when a new node $n+1$ connects to a node l in $G(n)$, and there is a longest path from the initial node to node l of length k , we necessarily obtain a longest path of length $k+1$ from the initial node to node $n+1$. As there are exactly $f_k(n)$ nodes to which the initial node has a longest path of length k , the probability of obtaining a longest path of length $k+1$ to node $n+1$ using one of the links in the in-degree of $n+1$ is $f_k(n)/n$. The probability to create a path with

⁴The approach in (P. Persoon et al., 2021) is slightly different: in that contribution we count each path to an increasing number of initial nodes. Yet it can be demonstrated (see supplementary material) that this amounts to a simple change of initial conditions, the effect of which on the number of paths and expected path length is negligible for large n .

length $k + 1$ or less using one the links in the in-degree of $n + 1$ is thus $\sum_{t=0}^k f_t(n)/n$. Hence collectively considering all links in the in-degree of $n + 1$, the probability to obtain a longest path of length $k + 1$ is $(\sum_{t=0}^k f_t(n)/n)^m - (\sum_{t=0}^{k-1} f_t(n)/n)^m$, and the expected increase $\Delta_n f_{k+1}(n)$ is

$$\Delta_n f_{k+1}(n) = \left(\sum_{t=0}^k \frac{f_t(n)}{n} \right)^m - \left(\sum_{t=0}^{k-1} \frac{f_t(n)}{n} \right)^m \quad (4.17)$$

Introducing $H_k(n) = \sum_{t=0}^k f_t(n)$ i.e. the number of longest paths with length shorter than $k + 1$, and summing both the left and the right of Equation 4.17 over k , starting from $k = 0$, we obtain

$$\Delta_n H_{k+1}(n) = n^{-m} H_k(n)^m. \quad (4.18)$$

It is not straightforward to obtain an exact solution to this equation, we can however identify a number of characteristic properties and use these to derive a greater estimate of $f_k(n)$. First, we rewrite Equation 4.18 as

$$H_{k+1}(n) = 1 + \sum_{s=1}^{n-1} \frac{H_k(s)^m}{s^m} \quad (4.19)$$

From this form, knowing that $H_0(n) = 1$ for all n , it is clear that for $n \rightarrow \infty$ and $m > 1$, we have $H_1(n) \rightarrow \zeta(m) + 1$, where $\zeta(m)$ is the Riemann Zeta function. The number of longest paths of at most length 1 hence does not grow unbounded, but instead converges to some finite value. In turn we can use the fact that $H_1(n)$ converges to show that $H_2(n)$ converges, etc., concluding that each $H_k(n)$ ultimately converges to some limit for $n \rightarrow \infty$, which we will denote by H_k^∞ . Likewise, the $f_k(n)$ will converge to $f_k^\infty = H_k^\infty - H_{k-1}^\infty$. A main question about the distribution is therefore: how does H_k^∞ depend on k ? In the following we analyze this relation in more detail by simplifying the dependence of $H_k(n)$ on n . More precisely, in the next section we discuss a zeroth order approximation in n of $H_k(n)$ and in the section that follows a first order approximation in n .

4.3.1 Zeroth order approximation

We will investigate the dependence of H_k^∞ on k by maximally simplifying the dependence of $H_k(n)$ on n . We start by noting that, since $H_k(n)$ counts the number of longest paths of length k or less, we have that $H_k(n) = n$ for $n < k$. For $n < k$ therefore, $H_k(n)$ increases linear with 1 path per added node. At the same time, we see from Equation 4.18 that, for $n > k$, $\Delta_n H_k(n)$ monotonously decreases to zero, hence $H_k(n)$ slowly but gradually comes closer to H_k^∞ . We can roughly approximate this development by assuming $H_k(n)$ continues to grow linear in n until it reaches H_k^∞ at $n = H_k^\infty$, after which it no longer increases and takes the constant value H_k^∞ . As there is no real dependence on n in this approximation, we will refer to this as a zeroth order approximation in n . As $H_k(n)$ in fact already starts to be slightly smaller than n for $n > k$, we note that our approximation is generally equal or greater than the actual value, hence resulting in an overestimation of H_k^∞ .

Equation 4.19 in this approximation becomes

$$H_{k+1}^\infty \approx 1 + \sum_{n=1}^{H_k^\infty} 1 + \sum_{n=H_k^\infty}^{\infty} \frac{H_k^{\infty m}}{n^m} \quad (4.20)$$

$$\approx 1 + H_k^\infty + H_k^{\infty m} \left(\frac{H_k^{\infty(1-m)}}{m-1} + \mathcal{O}(H_k^{\infty-m}) \right). \quad (4.21)$$

From this last expression we see that $H_{k+1}^\infty \approx 1 + H_k^\infty \frac{m}{m-1}$, hence $H_k^\infty \propto \left(\frac{m}{m-1}\right)^k$. We therefore conclude that the upper bounds of the number of longest paths of length k depend exponentially on k , and a first approximation for the base of the exponent is $\beta_0 = \frac{m}{m-1}$. This base approaches 1 for larger values of m with a rate $1/(m-1)$. The upper bounds of $f_k(n)$ therefore increase more slowly in k when the average in-degrees are larger. This makes sense, as with larger in-degrees, the creation of longer paths is more likely, hence resulting in relatively less longest paths with short length. As will turn out later however, this approximation to the exponential base could use some improvement. We will lay out the main steps to arrive at this improvement, for the details we refer to the supplementary material.

4.3.2 First order approximation

In Equation 4.20 we approximate $H_k(n)$ by a linear and a constant part. This translates to an $f_k(n)$ which is zero for $n < H_k^\infty$, and then abruptly $f_k(n) = f_k^\infty$ for $n \geq H_k^\infty$, hence there is no real dynamic dependence on n in that approximation. As an improvement, we could therefore include the first order of n in our approximation for $H_k(n)$. As $H_k(n) = n$ for $n < k$ let us suppose, as before, that $H_k(n) = n$ up to some n_k which we specify later. This allows us to split the sum in Equation 4.19 in a part for $1 \leq n < n_k$ and a part for greater values of $n \geq n_k$, and we suppose that n_k is sufficiently large such that the latter sum can be well approximated by an integral, leaving us with, for an $n > n_k$

$$H_{k+1}(n) = n_k + \int_{n_k}^n \frac{H_k(n)^m}{n^m} dn. \quad (4.22)$$

This relation is satisfied to first order in n for

$$H_k(n) = \begin{cases} n & \text{if } n < n_k \\ a_k - \frac{a_{k-1}^m}{(m-1)n^{m-1}} & \text{if } n \geq n_k \end{cases} \quad (4.23)$$

where it counts for the parameters a_k that

$$a_{k+1} = n_k + \frac{a_k^{m+1}}{(m+1)a_{k-1}^m} - \frac{1}{(m+1)a_{k-1}^m} \left(a_k - \frac{a_{k-1}^m}{(m-1)n_k^{m-1}} \right)^{m+1}. \quad (4.24)$$

Note therefore that for $n \rightarrow \infty$, we have $H_k(n) \rightarrow a_k$ and therefore $a_k = H_k^\infty$. Next we specify n_k . From Equation 4.18 we know that $\Delta_n H_{k+1}(n) = 1$ for $n < k$ and for greater values of n it (slowly) decreases. While we would like to therefore choose n_k as close to k as possible, it should also satisfy $\Delta_n H_{k+1}(n) \leq 1$. If we take the solution in Equation 4.23 for $k+1$, differentiate with respect to n and substitute n_k , we obtain an expression for the slope of $H_{k+1}(n)$ at $n = n_k$, which is a_k^m/n_k^m .

The least value for n_k we can thus choose while keeping the slope at n_k smaller or equal than 1 is $n_k = a_k$, which we shall henceforth implement. Note that this value implies that slope of $H_{k+1}(n)$ equals 1, thus ensuring a smooth transition between the part $n < n_k$ and the part $n \geq n_k$. Similar to the zeroth order approximation, we can show that this first order approximation to $H_k(n)$ is generally equal or above its actual value, and will therefore result in an overestimation of H_k^∞ .

While obtaining an exact solution for the relation in Equation 4.24 after substituting $n_k = a_k$ remains challenging, we note all the terms on the right-hand side of the equation are of net order 1 in a_k and/or a_{k-1} , indicating that, at least for large k , a_k and thus H_k^∞ increases exponentially in k . Let us suppose for large k we can write $a_{k-1}\beta_1 = a_k$, Equation 4.24 then reduces to

$$\beta_1 = 1 + \frac{\beta_1^m}{m+1} - \frac{\beta_1^m}{m+1} \left(1 - \frac{1}{(m-1)\beta_1^m}\right)^{m+1} \quad (4.25)$$

While this equation does not allow us to write β_1 in terms of elementary functions of m , expanding the part in brackets to second order in $1/\beta_1^m$ gives

$$\beta_1 \approx 1 + \frac{1}{m-1} - \frac{m}{2(m-1)^2\beta_1^m} + \dots \quad (4.26)$$

This shows that $1 < \beta_1 < \beta_0$ for all m , thus confirming this approximation is an improvement to the zeroth order approximation. Also, the last term on the right-hand side is of net order $1/(m-1)$, which implies that the term of order $1/(m-1)$ in the expansion of β_1 cannot simply be taken to equal $1/(m-1)$ (as it is for β_0). This indicates that this first order approximation to $H_k(n)$ is not just an improvement in orders greater than $1/(m-1)$. In the supplementary material it is demonstrated how may use Equation 4.25 to derive the greater estimate

$$\beta_1 = 1 + \frac{e - e^{1-\frac{1}{e}}}{m-1} + \frac{e^{-1-\frac{1}{e}}(1 + 3e + 3e^2) - 3e}{2(m-1)^2} + \dots \quad (4.27)$$

This therefore shows that, similar to β_0 , the exponential base β_1 approaches 1 for larger values of m , yet where β_0 does so by a rate $1/(m-1)$, β_1 does so by a rate which is an approximate factor $e - e^{1-\frac{1}{e}} \approx 0.84$ smaller.

4.3.3 Expected path length

Theoretically, as long as we can find solutions for $H_k(n)$ to second, third etc order in n we can continue to derive better approximations β_2, β_3 etc for the exponential base of H_k^∞ . In this contribution we however stop here and instead derive what the exponential dependence of H_k^∞ implies for the expected path length. Even though we only demonstrated that H_k^∞ approaches a exponential function for larger values of k , let us approximate $a_k \propto \beta_1^k$ for all k . Using the condition that for all n we have a single path of length 0, thus $H_0^\infty = f_0^\infty = 1$, we can approximate

$$H_k^\infty = \frac{\beta_1^{k+1} - 1}{\beta_1 - 1} \quad \text{and} \quad f_k^\infty = \beta_1^k. \quad (4.28)$$

Let us define k_n as the largest k for which $f_k(n)$ is non-zero. As we require that the sum $f_k(n)$ over all k to equal n , (as there is one longest path for each node), this

allows us to write $n = \sum_{s=0}^{k_n} f_s(n) = H_{k_n}(n)$, or

$$n = \frac{\beta_1^{k_n+1} - 1}{\beta_1 - 1} - \left(\frac{\beta_1^{k_n} - 1}{\beta_1 - 1} \right)^m \frac{1}{(m-1)n^{m-1}}. \quad (4.29)$$

When n gets large the second term on the right-hand side goes to zero (and note this term is absent in the zeroth order approximation of Section 4.3.1). For both the zeroth and first order approximation we can therefore deduce that, when n is large $k_n \approx \frac{\log(n(\beta_1-1)+1)}{\log(\beta_1)} - 1$, allowing us to compactly write for the distribution

$$f_k(n) = \begin{cases} \beta_1^k & \text{if } k \leq \frac{\log(n(\beta_1-1)+1)}{\log(\beta_1)} - 1 \\ 0 & \text{if } k > \frac{\log(n(\beta_1-1)+1)}{\log(\beta_1)} - 1. \end{cases}$$

We can use this to calculate the expected longest path length $\ell(n) = \sum_k k f_k(n)/n$. For large n this expression can be shown to reduce to, up to constant terms, $\ell(n) \approx k_n$. We therefore conclude that the expected longest path length increases logarithmically in n , with a coefficient $d_m = \log(\beta_1)^{-1}$ (see also Equation 4.8), which implies that a greater estimate of β_1 results in a lower estimate of d_m . Before we consider the value of d_m for β_1 in more detail, let us first consider it for β_0 instead. We can for $m > 2$ approximate $\log(\beta_0)^{-1} \approx m - \frac{1}{2}$. Recall that we derived the exact value $d_m = m$ when we considered all unique paths in Section 4.2.1. While the value for d_m based on β_0 is thus of the same proportion, the small shift of $1/2$ in fact makes it somewhat smaller. β_0 Should therefore not be considered an accurate approximation: the longest paths are expected to be at least as long, yet probably longer on average, than the collection of all unique paths. Hence let us finally approximate d_m based on β_1 . We thereby use the greater estimate for β_1 from Equation 4.27. This gives $\log(\beta_1)^{-1} \approx 1.2m - 0.6$. Note that, up to a minor shift, this is a factor 1.2 greater than the d_m found for all unique paths, which makes more sense than results based on β_0 . It suggests that, regardless of the number of nodes and average in-degree in the network, the longest paths are larger than the rest of the paths at least by a fixed proportion. We derived a lower estimate for this of 1 : 1.2, yet with an improved approximation of the exponent base β_1 we are likely to find a greater value for this proportion.

4.4 Conclusions

Studying cumulative structure in knowledge networks is key to understanding the advancement of science and technology, and has besides theoretical implications also relevance for science and technology policies. Approaching a body of knowledge as a network of discrete findings connected through knowledge flows, the notion of network paths and path length can be used to study to what extent sequences of findings appear, which form a key element of cumulative knowledge structures. It is in that context key to study (all) intermediate steps of development, hence not to limit the analysis to the shortest paths. In this contribution, we have therefore studied the path length distribution of (i) all unique paths from a given initial node to each node in the network and (ii) the longest paths from the initial node to each node in the network.

In the part of this work where we considered all unique paths, we derived an exact solution for the path length distribution and expected path length in the

particular context of the commonly used Price model. In this model, two main properties play a role: the average in-degree (AID) and the 'Cumulative Advantage Effect' (CAE). We find that, for large networks, the path length distributions can be characterized as Poisson-like, and are more skewed to lower path length values when the AID is smaller and the CAE is stronger. Similarly, we find that the expected path length grows logarithmically with the number of nodes and that the coefficient of this growth is smaller when the AID is smaller and the CAE stronger. In fact, the CAE puts an upper limit to this coefficient, and this upper limit is lower when the CAE is stronger. The upper limit disappears when there is no CAE. These results are more nuanced when the AID is less than 1 (which, though possible, may be rather uncommon in knowledge networks). In that case, a stronger CAE may slightly accelerate the growth of the number of paths, yet still has a tempering effect on the path length growth.

These results may be generalized by allowing the AID to increase with the number of nodes in any power relation. As it turns out, the CAE then plays a crucial role in keeping Poisson-like path length distributions and logarithmic expected path length growth. Only when the AID increases very fast, to be precise exponentially with the number of nodes, then we obtain binomial-like path length distributions and linear path length growth. Without the CAE, these types of path length distribution and expected path length growth would already be obtained for an AID that increases linearly with the number of nodes. The CAE therefore categorically tempers path length growth.

In the part of this work where we consider only the longest paths from an initial node to each node in the network, we only approximate the path length distribution and expected path length, as deriving exact solutions is in this case analytically more challenging. For simplicity, we also focus on the neutral case where the CAE is absent. Notwithstanding our analysis indicates key differences with the case where we consider all unique paths. Where for the latter, the number of paths of a given length grows unbounded, the number of longest paths of a given length is bound to an upper limit. Our approximation suggests that these upper limits increase exponentially with associated lengths and that the base of this exponent is a number slightly larger than 1, and approaches 1 for a greater AID. This makes sense as with a greater AID, we obtain longer paths at a rather earlier stage of the network development than for lower AID. While the distributions over the path lengths thus appear to be rather different, the expected path length appears to develop in fact rather similar. First-order estimates indicate that the expected path length of the longest paths increases at least logarithmically with the number of nodes, with a coefficient proportional to the AID, and an additional constant factor of at least 1.2. This is similar to the case of all unique paths without the CAE, except for the constant factor of 1.2. This is however a first theoretical approximation of this factor, and more elaborate approximations are likely to correct this to a greater value.

To conclude, we have shown that fundamental network properties and dynamics characteristically shape elements of knowledge networks that we can associate with cumulative structures, such as the notion of path length. In particular, the (development of the) AID and the strength of the CAE are relevant properties to consider in this context, as they can be meaningfully interpreted to determine variations in cumulative structures across different knowledge networks.

4.5 Discussion

Finally, we discuss some deeper implications and shortcomings of our analysis. First, our results have a number of deeper implications in particular for the study of cumulative knowledge structures. While researchers aiming for a quantitative approach benefit from a network approach to knowledge structures, they should be aware of the various choices that network analysis allows to identify knowledge flow, in particular the differences between using the shortest, longest, or all unique paths in the network. Where the average distance based on the length of the shortest paths in a scale-free network (of which the Price network is a special case) is known to increase with the log log number of nodes (R. Cohen & Havlin, 2003), we have shown that the average path length based on the length of all unique paths from an initial node to each node in a Price network increases with the log number of nodes. Furthermore, we have shown that there are fundamentally different properties of the path length distributions of all unique paths and the subset of longest paths, even without including sophisticated dynamic principles such as the cumulative advantage effect.

Additionally, before a certain path length metric is applied to study the cumulative structure of a particular field of knowledge or discipline, the researcher is advised to investigate a number of characteristics of the network, such as a possible development of in-degree as the network grows as well as the presence of the cumulative advantage effect. Our work indicates that the presence of either (and especially the presence of both) greatly affects cumulative structures in those networks. Our work allows the researcher to then formulate a number of specific expectations, especially for the path length distribution and expected path length of all unique paths. Our contribution thus provides a first step towards a framework in which cumulative structures can generally be studied and in which variations between fields or disciplines can meaningfully be interpreted.

A second deeper implication of our results is of more theoretical nature. In this contribution, we have shown that the cumulative advantage effect explicitly prohibits path lengths to grow faster than logarithmically as long as the average in-degree does not increase exponentially. In another contribution (P. Persoon et al., 2021), where we include an empirical analysis of technological knowledge using patent and patent citation data, we actually find that the average path length (based on counting all unique paths) increases linearly, even though the in-degrees do not increase exponentially (but linearly instead). This may imply that the cumulative advantage effect plays no role in these networks, yet, interestingly, other contributions have suggested that the cumulative advantage effect does play a role in these networks (Érdi et al., 2013; Valverde et al., 2007). Another explanation may be that this differs per technology, or that there may be other effects at work, which were not included in this analysis.

One of those excluded effects, which brings us to the first shortcoming of this analysis, is the time dependence of knowledge dynamics. As other contributions have indicated, these effects may play a rather substantial role (Garavaglia et al., 2017; Golosovsky, 2017). Indeed one of the criticisms of the Price model is that the oldest nodes in the networks effectively gain the greatest out-degree. In real-life situations, the fact that a finding is old need not automatically imply it is more relevant than any new finding. The model discussed in this work would therefore benefit from an extension that takes into account time effects, such as the fading of relevance. While

a number of such models can be found in the literature (Golosovsky, 2017; D. Wang et al., 2013; Wu et al., 2014), it is however not directly clear how to analytically calculate the path length distributions in these models.

A second shortcoming is our focus on (only) counting the paths from a single given initial node. While this focus may be perfect for studies interested in the particular impact or role of a single finding, for a general understanding of cumulative structures, depending too much on a particular choice for a single node might appear arbitrary and may even be misleading. A simple way to generalize this would be to allow for the possibility of multiple initial nodes, or for the number of initial nodes to increase as the network grows. We explain in more detail in the supplementary material and in (P. Persoon et al., 2021), how these choices could be implemented by slightly changing the initial conditions for Equations 4.3 and 4.15. While these changes introduce an extra parameter, they are found not to lead to fundamentally different results when we consider networks with a substantial number of nodes.

4.6 Acknowledgements

The author is grateful to Floor Alkemade, Rudi Bekkers and Elena Mas Tur for helpful comments on the script. This work was supported by NWO (Dutch Research Council) grant nr. 452-13-010.

Appendix

4.A Derivations with all unique paths

In this section of the supplementary material we consider for all unique paths in a Price network first the path length distribution (first excluding, then including the cumulative advantage effect), then calculate the total number of paths, the expected path length and finally we consider the generalization for increasing in-degree.

4.A.1 Excluding the cumulative advantage effect

When we exclude the cumulative advantage effect (i.e. $c = 0$) the solution for $f_k(n)$ involves the unsigned Stirling numbers of the first kind $S(n, k)$, which are defined as the coefficients in the rising factorial, $x^{\bar{n}} = x(x+1)\dots(x+n-1) = \sum_{k=0}^n S(n, k)x^k$, which satisfy the recurrence relation $S(n+1, k+1) = nS(n, k+1) + S(n, k)$ and for which it counts in particular that $S(n, k) = 0$ for $n < k$ and $S(n, 1) = \Gamma(n)$ for all n , where $\Gamma(n)$ is the gamma function, which satisfies $\Gamma(n+1) = n\Gamma(n)$. We need to show that

$$f_k(n) = \frac{m^k}{\Gamma(n)} S(n, k+1) \quad (4.30)$$

i.e. Equation 4.4, satisfies the relation $\Delta_n f_{k+1}(n) = m f_k(n)/n$. Substituting Equation 4.30 in $\Delta_n f_{k+1}(n) = f_{k+1}(n+1) - f_{k+1}(n)$ gives

$$\Delta_n f_{k+1}(n) = m^{k+1} \left(\frac{S(n+1, k+2)}{\Gamma(n+1)} - \frac{S(n, k+2)}{\Gamma(n)} \right) \quad (4.31)$$

$$= \frac{m^{k+1}}{\Gamma(n+1)} (S(n+1, k+2) - nS(n, k+2)). \quad (4.32)$$

using the recurrence relation, we end up with

$$\Delta_n f_{k+1}(n) = \frac{mm^k}{n\Gamma(n)} S(n, k+1) = \frac{m}{n} f_k(n). \quad (4.33)$$

Furthermore, because $S(n, 1) = \Gamma(n)$, $f_0(n) = 1$ for all n . Finally, noting that $S(n, k+1) = 0$ for $n < k$, we note that $f_k(1) = 0$ for all $k > 0$, as required. Finally, we derive Equation 4.6. We use that

$$x^{\bar{n}} = x(x+1)\dots(x+n-1) = \frac{\Gamma(x+n)}{\Gamma(x)} \quad (4.34)$$

and that $S(n, 0) = 0$ for all $n \neq 0$. Summing over all $k < n$ (hence up to $n - 1$), we can write

$$K(n) = \sum_{k=0}^{n-1} f_k(n) \quad (4.35)$$

$$= \sum_{k=0}^{n-1} \frac{m^k}{\Gamma(n)} S(n, k+1) \quad (4.36)$$

$$= \frac{1}{m\Gamma(n)} \sum_{k=1}^n m^k S(n, k) \quad (4.37)$$

$$= \frac{1}{m\Gamma(n)} \frac{\Gamma(m+n)}{\Gamma(m)} \quad (4.38)$$

$$= \frac{\Gamma(m+n)}{\Gamma(n)\Gamma(m+1)}, \quad (4.39)$$

which is Equation 4.6 in the paper.

4.A.2 Including the cumulative advantage effect

The general solution for $f_k(n)$ involves unsigned non-central Stirling numbers of the first kind $S_y(n, k)$ (Koutras, 1982; M. D. Schmidt, 2016), which are defined as $x^{\overline{n}} = \sum_{t=0}^n S_y(n, t)(x-y)^t$, satisfy the recurrence relation $S_y(n+1, k+1) = (n+y)S_y(n, k+1) + S_y(n, k)$ and for which it counts in particular that $S_y(n, k) = 0$ for $n < k$ and $S_y(n, n) = 1$ for all y and n . We need to show that the general solution for $f_k(n)$ (i.e. Equation 4.9),

$$f_k(n) = \frac{1}{\Gamma(n)(-c)^k} \sum_{t=k+1}^n (t-k)S_p(n, t)(-p)^{t-1}, \quad (4.40)$$

which introduces the parameter $p = mc/(1+mc)$, satisfies the relation (i.e. Equation 4.3)

$$\Delta_n f_{k+1}(n) = m \frac{f_k(n) + c f_{k+1}(n)}{(1+cm)n}. \quad (4.41)$$

We start by rewriting this expression as

$$\Delta_n f_{k+1}(n) = \frac{p}{cn} f_k(n) + \frac{p}{n} f_{k+1}(n) \quad (4.42)$$

$$f_{k+1}(n+1) - f_{k+1}(n) = \frac{p}{cn} f_k(n) + \frac{p}{n} f_{k+1}(n) \quad (4.43)$$

$$f_{k+1}(n+1) - \frac{n+p}{n} f_{k+1}(n) = \frac{p}{cn} f_k(n). \quad (4.44)$$

Next we use Equation 4.40 to substitute the expressions for $f_{k+1}(n+1)$, $f_{k+1}(n)$ and $f_k(n)$, which gives

$$\frac{1}{\Gamma(n+1)(-c)^{k+1}} \sum_{t=k+2}^{n+1} (t-k-1)S_p(n+1, t)(-p)^{t-1} - \frac{n+p}{n\Gamma(n)(-c)^{k+1}} \sum_{t=k+2}^n (t-k-1)S_p(n, t)(-p)^{t-1} = \frac{p}{cn\Gamma(n)(-c)^k} \sum_{t=k+1}^n (t-k)S_p(n, t)(-p)^{t-1} \quad (4.45)$$

$$\frac{1}{\Gamma(n+1)(-c)^{k+1}} \sum_{t=k+2}^{n+1} (t-k-1)S_p(n+1, t)(-p)^{t-1} - \frac{n+p}{\Gamma(n+1)(-c)^{k+1}} \sum_{t=k+2}^n (t-k-1)S_p(n, t)(-p)^{t-1} = \frac{-p}{\Gamma(n+1)(-c)^{k+1}} \sum_{t=k+1}^n (t-k)S_p(n, t)(-p)^{t-1} \quad (4.46)$$

$$\sum_{t=k+2}^{n+1} (t-k-1)S_p(n+1, t)(-p)^{t-1} - (n+p) \sum_{t=k+2}^n (t-k-1)S_p(n, t)(-p)^{t-1} = -p \sum_{t=k+1}^n (t-k)S_p(n, t)(-p)^{t-1} \quad (4.47)$$

$$\sum_{t=k+2}^{n+1} (t-k-1)S_p(n+1, t)(-p)^{t-1} - (n+p) \sum_{t=k+2}^n (t-k-1)S_p(n, t)(-p)^{t-1} = \sum_{t=k+1}^n (t-k)S_p(n, t)(-p)^t \quad (4.48)$$

$$(n-k)S_p(n+1, n+1)(-p)^n + \sum_{t=k+2}^n (t-k-1)S_p(n+1, t)(-p)^{t-1} - (n+p) \sum_{t=k+2}^n (t-k-1)S_p(n, t)(-p)^{t-1} = \sum_{t=k+1}^{n-1} (t-k)S_p(n, t)(-p)^t + (n-k)S_p(n, n)(-p)^n \quad (4.49)$$

$$\sum_{t=k+2}^n (t-k-1)S_p(n+1, t)(-p)^{t-1} - (n+p) \sum_{t=k+2}^n (t-k-1)S_p(n, t)(-p)^{t-1} = \sum_{t=k+1}^{n-1} (t-k)S_p(n, t)(-p)^t \quad (4.50)$$

$$\sum_{t=k+2}^n (t-k-1)(S_p(n+1, t) - (n+p)S_p(n, t))(-p)^{t-1} = \sum_{t=k+1}^{n-1} (t-k)S_p(n, t)(-p)^t \quad (4.51)$$

$$\sum_{t=k+1}^{n-1} (t-k)(S_p(n+1, t+1) - (n+p)S_p(n, t+1))(-p)^t = \sum_{t=k+1}^{n-1} (t-k)S_p(n, t)(-p)^t \quad (4.52)$$

where we used that $S_p(n, n) = 1$ for all n . We see that, the last expression is just the recurrence relation for the unsigned non-central Stirling numbers of the first kind added $n-k$ times. This relation is therefore satisfied for all k and n . We therefore conclude that Equation 4.40 satisfies Equation 4.41.

Before we show that Equation 4.40 satisfies $f_0(n) = 1$ for all n , we first need to derive a more general result. We will need that the derivative of the Gamma function $\Gamma(x)' = \psi(x)\Gamma(x)$, where $\psi(x)$ is the digamma function, which satisfies the recurrence relation $\psi(x) + \frac{1}{x} = \psi(x+1)$. Using Equation 4.34, let us differentiate the definition of the non-central unsigned Stirling numbers $\sum_{t=0}^n S_y(n, t)(x-y)^t = \frac{\Gamma(n+x)}{\Gamma(x)}$ left and right with respect to x to obtain

$$\sum_{t=0}^n tS_y(n, t)(x-y)^{t-1} = \frac{\Gamma(n+x)}{\Gamma(x)}(\psi(n+x) - \psi(x)) \quad (4.53)$$

$$= \frac{\Gamma(n+x)}{\Gamma(x)}(\psi(n+x) - \psi(x+1) + \frac{1}{x}) \quad (4.54)$$

$$= \frac{\Gamma(n+x)}{\Gamma(x)}(\psi(n+x) - \psi(x+1)) + \frac{\Gamma(n+x)}{\Gamma(x+1)}. \quad (4.55)$$

If we take the limit of $x \rightarrow 0$, note that $\Gamma(x) \rightarrow \infty$, hence we obtain for $n \geq 1$ and $y \neq 0$,

$$\sum_{t=0}^n tS_y(n, t)(-y)^{t-1} = \sum_{t=1}^n tS_y(n, t)(-y)^{t-1} = \Gamma(n). \quad (4.56)$$

Using Equation 4.40 to write

$$f_0(n) = \frac{1}{\Gamma(n)} \sum_{t=1}^n tS_p(n, t)(-p)^{t-1}, \quad (4.57)$$

we can directly use Equation 4.56 to see that $f_0(n) = 1$ for all n . Finally, we note that from the sum appearing in Equation 4.40, it is obvious that $f_k(n) = 0$ for $k > n$.

Total number of paths

Next we will calculate the total number of paths $K(n) = \sum_{k=0}^{\infty} f_k(n)$, i.e. deriving Equation 4.12. Because $f_k(n)$ is zero for $k > n$ we only need to sum up to $k = n$. First, we use Equation 4.56 and the fact that $\sum_{t=0}^n S_y(n, t)(-y)^t = 0$ to rewrite the Equation 4.40 as

$$f_k(n) = \frac{1}{\Gamma(n)(-c)^k} \sum_{t=k+1}^n (t-k)S_p(n, t)(-p)^{t-1} \quad (4.58)$$

$$= \frac{1}{\Gamma(n)(-c)^k} \left(\sum_{t=0}^n - \sum_{t=0}^k \right) (t-k)S_p(n, t)(-p)^{t-1} \quad (4.59)$$

$$= \frac{1}{\Gamma(n)(-c)^k} \left(\sum_{t=0}^n tS_p(n, t)(-p)^{t-1} - \sum_{t=0}^n kS_p(n, t)(-p)^{t-1} - \sum_{t=0}^k (t-k)S_p(n, t)(-p)^{t-1} \right) \quad (4.60)$$

$$= \frac{1}{\Gamma(n)(-c)^k} \left(\sum_{t=0}^n tS_p(n, t)(-p)^{t-1} - \sum_{t=0}^k (t-k)S_p(n, t)(-p)^{t-1} \right) \quad (4.61)$$

$$= \frac{1}{\Gamma(n)(-c)^k} \left(\Gamma(n) - \sum_{t=0}^{k-1} (t-k)S_p(n, t)(-p)^{t-1} \right) \quad (4.62)$$

$$= \frac{1}{(-c)^k} \left(1 - \frac{1}{\Gamma(n)} \sum_{t=0}^{k-1} (t-k)S_p(n, t)(-p)^{t-1} \right) \quad (4.63)$$

Summing this expression for all $k \leq n$ gives

$$K(n) = \sum_{k=0}^n \frac{1}{(-c)^k} \left(1 - \frac{1}{\Gamma(n)} \sum_{t=0}^{k-1} (t-k) S_p(n, t) (-p)^{t-1} \right) \quad (4.64)$$

$$= \sum_{k=0}^n \frac{1}{(-c)^k} + \sum_{k=0}^n \frac{(-1/c)^k}{p\Gamma(n)} \left(\sum_{t=0}^{k-1} t S_p(n, t) (-p)^t - k \sum_{t=0}^{k-1} S_p(n, t) (-p)^t \right) \quad (4.65)$$

$$= \frac{(-1/c)^n + c}{1+c} + \sum_{k=0}^n \frac{(-1/c)^k}{p\Gamma(n)} \left(\sum_{t=0}^{k-1} t S_p(n, t) (-p)^t - k \sum_{t=0}^{k-1} S_p(n, t) (-p)^t \right) \quad (4.66)$$

$$= \frac{(-1/c)^n + c}{1+c} + \sum_{k=0}^n \frac{(-1/c)^k}{p\Gamma(n)} \sum_{t=0}^{k-1} t S_p(n, t) (-p)^t - \sum_{k=0}^n \frac{(-1/c)^k}{p\Gamma(n)} k \sum_{t=0}^{k-1} S_p(n, t) (-p)^t \quad (4.67)$$

To proceed, we for clarity separately consider the two 'sum of a sum' terms. For the first term we have

$$\sum_{k=0}^n \frac{(-1/c)^k}{p\Gamma(n)} \sum_{t=0}^{k-1} t S_p(n, t) (-p)^t = \sum_{k=2}^n \frac{(-1/c)^k}{p\Gamma(n)} \sum_{t=1}^{k-1} t S_p(n, t) (-p)^t \quad (4.68)$$

Explicitly writing out the right-hand side of this equation, choosing a separate line for each k , gives us (without the prefactor $\frac{1}{\Gamma(n)p}$)

$$(-1/c)^2 S(n, 1) (-p) + \quad (4.69)$$

$$(-1/c)^3 S(n, 1) (-p) + (-1/c)^3 2 S(n, 2) (-p)^2 +$$

$$(-1/c)^4 S(n, 1) (-p) + (-1/c)^4 2 S(n, 2) (-p)^2 + (-1/c)^4 3 S(n, 3) (-p)^3 +$$

⋮

$$(-1/c)^n S(n, 1) (-p) + (-1/c)^n 2 S(n, 2) (-p)^2 + \dots + (-1/c)^n (n-1) S(n, n-1) (-p)^{n-1}.$$

We see that we can alternatively sum the geometric progressions in the vertical direction. Using that $\sum_{k=r}^n x^k = (x^{n+1} - x^r)/(x-1)$, or that

$$\sum_{k=r}^n (-1/c)^k = \frac{(-1/c)^{n+1} + (-1/c)^r}{-1/c - 1} \quad (4.70)$$

$$= \frac{(-1/c)^n - (-1/c)^{r-1}}{1+c}, \quad (4.71)$$

we can rewrite Expression 4.69 as

$$\begin{aligned} & \frac{(-1/c)^n - (-1/c)}{1+c} S(n, 1) (-p) + 2 \frac{(-1/c)^n - (-1/c)^2}{1+c} S(n, 2) (-p)^2 \\ & + 3 \frac{(-1/c)^n - (-1/c)^3}{1+c} S(n, 3) (-p)^3 + \dots + (n-1) \frac{(-1/c)^n - (-1/c)^{n-1}}{1+c} S(n, n-1) (-p)^{n-1}. \end{aligned} \quad (4.72)$$

Writing this in a summation form, again including the prefactor $\frac{1}{\Gamma(n)p}$, this becomes

$$\sum_{t=1}^{n-1} \frac{t(-p)^t S_p(n, t)}{p\Gamma(n)(1+c)} \left((-1/c)^n - (-1/c)^t \right). \quad (4.73)$$

Next we consider the second 'sum of sum' term in Equation 4.67, for which we have

$$\sum_{k=0}^n \frac{(-1/c)^k}{p\Gamma(n)} k \sum_{t=0}^{k-1} S_p(n, t) (-p)^t = \sum_{k=2}^n \frac{(-1/c)^k}{p\Gamma(n)} k \sum_{t=1}^{k-1} S_p(n, t) (-p)^t \quad (4.74)$$

Analogous to the first sum of sum term, we write this out without the prefactor $\frac{1}{\Gamma(n)_p}$, giving

$$\begin{aligned}
& 2(-1/c)^2 S(n, 1)(-p) + & (4.75) \\
& 3(-1/c)^3 S(n, 1)(-p) + 3(-1/c)^3 S(n, 2)(-p)^2 + \\
& 4(-1/c)^4 S(n, 1)(-p) + 4(-1/c)^4 S(n, 2)(-p)^2 + 4(-1/c)^4 S(n, 3)(-p)^3 + \\
& \vdots \\
& n(-1/c)^n S(n, 1)(-p) + n(-1/c)^n S(n, 2)(-p)^2 + \dots + n(-1/c)^n S(n, n-1)(-p)^{n-1}.
\end{aligned}$$

Using that, for any real number x

$$\sum_{k=r}^n kx^k = \frac{x^r(r(1-x) + x) - x^{n+1}(n(1-x) + 1)}{(x-1)^2}, \quad (4.76)$$

so that

$$\sum_{k=r}^n k(-1/c)^k = \frac{(-1/c)^r(r(1+1/c) + (-1/c)) - (-1/c)^{n+1}(n(1+1/c) + 1)}{((-1/c) - 1)^2} \quad (4.77)$$

$$= \frac{c(-1/c)^r((c+1)r - 1) + (-1/c)^n((c+1)n + c)}{(1+c)^2} \quad (4.78)$$

$$= \frac{c(-1/c)^r((c+1)(r-1) + c) + (-1/c)^n((c+1)n + c)}{(1+c)^2} \quad (4.79)$$

$$= \frac{-(-1/c)^{r-1}((c+1)(r-1) + c) + (-1/c)^n((c+1)n + c)}{(1+c)^2}, \quad (4.80)$$

we can rewrite Expression 4.75 as

$$\begin{aligned}
& \frac{(-1/c)^n((c+1)n + c) - (-1/c)((c+1) + c)}{(1+c)^2} S(n, 1)(-p) + & (4.81) \\
& \frac{(-1/c)^n((c+1)n + c) - (-1/c)^2((c+1)2 + c)}{(1+c)^2} S(n, 2)(-p)^2 + \\
& \frac{(-1/c)^n((c+1)n + c) - (-1/c)^3((c+1)3 + c)}{(1+c)^2} S(n, 3)(-p)^3 + \\
& \dots + \frac{(-1/c)^n((c+1)n + c) - (-1/c)^{n-1}((c+1)(n-1) + c)}{(1+c)^2} S(n, n-1)(-p)^{n-1},
\end{aligned}$$

or, more compactly written in summation form (bringing back the prefactor $\frac{1}{\Gamma(n)_p}$),

$$\sum_{t=1}^{n-1} \frac{(-p)^t S_p(n, t)}{p\Gamma(n)(1+c)^2} \left((-1/c)^n((c+1)n + c) - (-1/c)^t((c+1)t + c) \right). \quad (4.82)$$

We note that we may equally well choose to do this sum from $t = 0$ to n , as this amounts to adding terms that are zero (note that $S_y(n, 0) = 0$ for $n > 0$), thus writing

$$\sum_{t=0}^n \frac{S_p(n, t)(-p)^t}{p\Gamma(n)(1+c)^2} \left((-1/c)^n((c+1)n + c) - (-1/c)^t((c+1)t + c) \right). \quad (4.83)$$

Using that $\sum_{t=0}^n S_y(n, t)(-y)^t = 0$, this simplifies to

$$-\sum_{t=0}^n \frac{S_p(n, t)(-p)^t}{p\Gamma(n)(1+c)^2} (-1/c)^t ((c+1)t+c). \quad (4.84)$$

For Expression 4.73 we can likewise adjust the summation limits to $t = 0$ and n . Doing this, substituting the Expressions 4.73 and 4.84 back into Equation 4.67, gives

$$K(n) = \frac{(-1/c)^n + c}{1+c} + \sum_{t=0}^n \frac{t(-p)^t S_p(n, t)}{p\Gamma(n)(1+c)} ((-1/c)^n - (-1/c)^t) \quad (4.85)$$

$$\begin{aligned} &+ \sum_{t=0}^n \frac{S_p(n, t)(-p)^t}{p\Gamma(n)(1+c)^2} (-1/c)^t ((c+1)t+c) \\ &= \frac{(-1/c)^n + c}{1+c} + \sum_{t=0}^n \frac{tS_p(n, t)(-p)^t}{p\Gamma(n)(1+c)} ((-1/c)^n - (-1/c)^t) \end{aligned} \quad (4.86)$$

$$\begin{aligned} &+ \sum_{t=0}^n \frac{tS_p(n, t)(-p)^t}{p\Gamma(n)(1+c)} (-1/c)^t + \sum_{t=0}^n \frac{cS_p(n, t)(-p)^t}{p\Gamma(n)(1+c)^2} (-1/c)^t \\ &= \frac{(-1/c)^n + c}{1+c} + \sum_{t=0}^n \frac{tS_p(n, t)(-p)^t}{p\Gamma(n)(1+c)} (-1/c)^n + \sum_{t=0}^n \frac{cS_p(n, t)(-p)^t (-1/c)^t}{p\Gamma(n)(1+c)^2} \end{aligned} \quad (4.87)$$

$$= \frac{(-1/c)^n + c}{1+c} - \frac{(-1/c)^n}{\Gamma(n)(1+c)} \sum_{t=0}^n tS_p(n, t)(-p)^{t-1} + \frac{c}{p\Gamma(n)(1+c)^2} \sum_{t=0}^n S_p(n, t)(p/c)^t \quad (4.88)$$

Using Equation 4.56 to do the sum in the second term and the definition of the non-central unsigned Stirling numbers, $\sum_{t=0}^n S_y(n, t)(x-y)^t = \frac{\Gamma(n+x)}{\Gamma(x)}$ to do the sum in the third term, we obtain

$$= \frac{(-1/c)^n + c}{1+c} - \frac{(-1/c)^n}{1+c} + \frac{c\Gamma(n+p/c+p)}{p(1+c)^2\Gamma(n)\Gamma(p/c+p)} \quad (4.89)$$

$$= \frac{c}{1+c} + \frac{c\Gamma(n+p/c+p)}{c(1+c)(p/c+p)\Gamma(n)\Gamma(p/c+p)} \quad (4.90)$$

$$= \frac{c}{1+c} + \frac{\Gamma(n+\frac{p}{c}+p)}{(1+c)\Gamma(n)\Gamma(\frac{p}{c}+p+1)} \quad (4.91)$$

$$= \frac{c}{1+c} + \frac{\Gamma(n+m_c)}{(1+c)\Gamma(n)\Gamma(m_c+1)}, \quad (4.92)$$

where we used on the last step that $m_c = m(1+c)/(1+mc) = p/c+p$. Note that this reproduces Equation 4.12 in the paper.

To obtain the generating function $K_z(n)$ of the series $f_0(n), f_1(n), f_2(n), \dots$, we need to sum $K_z(n) = \sum_{k=0}^{\infty} f_k(n)z^k$. Using Equation 4.58, note this effectively amounts to doing the same calculation as in Equation 4.64 except that instead of c , we take c/z . Hence doing this substituting in Equation 4.91, we get the generating function

$$K_z(n) = \frac{c/z}{1+c/z} + \frac{\Gamma(n+z\frac{p}{c}+p)}{(1+c/z)\Gamma(n)\Gamma(z\frac{p}{c}+p+1)} \quad (4.93)$$

$$= \frac{c}{z+c} + \frac{\Gamma(n+z\frac{p}{c}+p)}{(1+c/z)\Gamma(n)\Gamma(z\frac{p}{c}+p+1)}. \quad (4.94)$$

Note that k times differentiating $K_z(n) = \sum_{k=0}^{\infty} f_k(n)z^k$ with respect to z takes away all $f_k(n)$ with $k < k$ and subsequently setting $z = 0$ takes away all $f_k(n)$ with $k > k$, thus leaving us with $k!f_k(n)$. An alternative way to obtain $f_k(n)$ is to

therefore differentiate $K_z(n)$ k times with respect to z , divide by $k!$ and set $z = 0$. For example for $f_1(n)$ we get

$$f_1(n) = \frac{\partial K_z(n)}{\partial z} \Big|_{z=0} \quad (4.95)$$

$$= \frac{\partial}{\partial z} \left(\frac{c}{z+c} + \frac{\Gamma(n+z\frac{p}{c}+p)}{(1+c/z)\Gamma(n)\Gamma(z\frac{p}{c}+p+1)} \right) \Big|_{z=0} \quad (4.96)$$

$$= \left(-\frac{c}{(c+z)^2} + \frac{c}{z^2(1+c/z)^2} \frac{\Gamma(n+z\frac{p}{c}+p)}{\Gamma(n)\Gamma(1+z\frac{p}{c}+p)} + \frac{\frac{p}{c}\Gamma(n+z\frac{p}{c}+p)\psi(n+z\frac{p}{c}+p)}{(1+c/z)\Gamma(n)\Gamma(1+z\frac{p}{c}+p)} \right. \\ \left. - \frac{\Gamma(1+z\frac{p}{c}+p)\frac{p}{c}\Gamma(n+z\frac{p}{c}+p)\psi(1+z\frac{p}{c}+p)}{(1+c/z)\Gamma(n)\Gamma(1+z\frac{p}{c}+p)^2} \right) \Big|_{z=0} \quad (4.97)$$

$$= \left(-\frac{c}{(c+z)^2} + \frac{c}{(z+c)^2} \frac{\Gamma(n+z\frac{p}{c}+p)}{\Gamma(n)\Gamma(1+z\frac{p}{c}+p)} + \frac{z\frac{p}{c}\Gamma(n+z\frac{p}{c}+p)\psi(n+z\frac{p}{c}+p)}{(z+c)\Gamma(n)\Gamma(1+z\frac{p}{c}+p)} \right. \\ \left. - \frac{z\frac{p}{c}\Gamma(n+z\frac{p}{c}+p)\psi(1+z\frac{p}{c}+p)}{(z+c)\Gamma(n)\Gamma(1+z\frac{p}{c}+p)} \right) \Big|_{z=0} \quad (4.98)$$

$$= -\frac{c}{c^2} + \frac{c\Gamma(n+p)}{c^2\Gamma(n)\Gamma(1+p)} \quad (4.99)$$

$$= \frac{1}{c} \left(-1 + \frac{\Gamma(n+p)}{\Gamma(n)\Gamma(1+p)} \right) \quad (4.100)$$

Differentiating again to obtain $f_2(n)$ and again to obtain $f_3(n)$, the expressions quickly become very large. Already for $f_2(n)$ there will be 21 distinct terms, which makes doing these calculations by hand labor-intensive and prone to mistakes. It is advisable to use a computer program to do these calculations. The following steps leading to $f_2(n)$ are derived using the program Mathematica to differentiate Equation 4.94 two times with respect to z , setting $z = 0$ and dividing by 2,

$$f_2(n) = \frac{1}{2} \frac{\partial^2 K_z(n)}{\partial z^2} \Big|_{z=0} \quad (4.101)$$

$$= \frac{1}{c^2} \left(1 - \frac{\Gamma(n+p)}{\Gamma(n)\Gamma(1+p)} + \frac{p\Gamma(n+p)(\psi(n+p) - \psi(1+p))}{\Gamma(n)\Gamma(1+p)} \right). \quad (4.102)$$

Similarly, we let the program differentiate Equation 4.94 three times with respect to z , set $z = 0$ and divide by 3!, to obtain $f_3(n)$

$$f_3(n) = \frac{1}{3!} \frac{\partial^3 K_z(n)}{\partial z^3} \Big|_{z=0} \quad (4.103)$$

$$= \frac{1}{c^3} \left(-1 + \frac{\Gamma(n+p)}{\Gamma(n)\Gamma(1+p)} - \frac{p\Gamma(n+p)(\psi(n+p) - \psi(1+p))}{\Gamma(n)\Gamma(1+p)} \right) \\ + \frac{p^2\Gamma(n+p) \left((\psi(n+p) - \psi(1+p))^2 + \psi_1(n+p) - \psi_1(1+p) \right)}{2\Gamma(n)\Gamma(1+p)} \quad (4.104)$$

The latter shows that, even after setting $z = 0$, the expressions quickly become very large for larger k . It therefore makes more sense to focus on the highest order in n instead, which we will explore in the following. We note that differentiating the Gamma function $\Gamma(x)$ l times results in terms proportional to $\Gamma(x)$ and different (powers of) polygamma functions $\psi_0(x), \psi_1(x), \dots, \psi_l(x)$. The polygamma functions converge to constant values for large x , with the exception of the digamma function $\psi_0(x) = \psi(x)$, which grows unbounded and can be approximated by $\psi(x) \approx \log(x)$. To select the highest order in n , we therefore focus on the terms with (powers of) $\psi(n)$. Using Equation 4.100 we observe no power of $\psi(n)$ for $f_1(n)$, yet in Equation 4.102 we observe a first power of $\psi(n)$ in $f_2(n)$ and in Equation 4.102 we observe a second power of $\psi(n)$ in $f_3(n)$. Furthermore, using that $\Gamma(x+a)/\Gamma(x) \approx x^a$ for large

x , we note that the prefactor in Equation 4.100, 4.102 and 4.104 $\frac{\Gamma(n+p)}{\Gamma(n)\Gamma(1+p)}$ increases approximately as $n^p/\Gamma(1+p)$. Considering only the terms proportional to $\frac{\Gamma(n+p)}{\Gamma(n)\Gamma(1+p)}$ and selecting from Equation 4.102 only the term linear in $\psi(n)$ and from Equation 4.104 only the term quadratic in $\psi(n)$, we obtain

$$f_1(n) \approx \frac{\Gamma(n+p)}{c\Gamma(n)\Gamma(1+p)} \quad (4.105)$$

$$f_2(n) \approx \frac{p\Gamma(n+p)(\psi(n+p) - \psi(1+p))}{c^2\Gamma(n)\Gamma(1+p)} \quad (4.106)$$

$$f_3(n) \approx \frac{p^2\Gamma(n+p)(\psi(n+p) - \psi(1+p))^2}{2c^3\Gamma(n)\Gamma(1+p)} \quad (4.107)$$

Continuing for $f_4(n), f_5(n), \dots$ etc we see that

$$f_k(n) \approx \left(\frac{p}{c}\right)^k \frac{\Gamma(n+p)}{p\Gamma(1+p)\Gamma(n)\Gamma(k)} (\psi(n+p) - \psi(1+p))^{k-1} \quad (4.108)$$

$$\approx \left(\frac{p}{c}\right)^k \frac{\Gamma(n+p)}{p\Gamma(1+p)\Gamma(n)\Gamma(k)} (\log(n+p) - \log(1+p))^{k-1} \quad (4.109)$$

$$\approx \left(\frac{p}{c}\right)^k \frac{\Gamma(n+p)}{p\Gamma(1+p)\Gamma(n)\Gamma(k)} \log^{k-1} \left(\frac{n+p}{1+p}\right), \quad (4.110)$$

which corresponds to Equation 4.11. Using that the prefactor in Equation 4.108 $\frac{\Gamma(n+p)}{\Gamma(n)\Gamma(1+p)} \approx n^p/\Gamma(1+p)$, we note that all $f_k(n)$ in the limit of large n grow as n^p , which is faster than the case where $c = 0$, when $f_k(n)$ grows as $\log(n)^k$. To show that there is an n below which $f_k(n)$ is greater for the $c = 0$ case than for general $c > 0$, we note that for $n = k + 1$, we obtain from Equation 4.40

$$f_k(k+1) = \frac{1}{\Gamma(n)(-c)^k} S_p(k+1, k+1)(-p)^k \quad (4.111)$$

$$= \frac{1}{\Gamma(n)(-c)^k} (-p)^k \quad (4.112)$$

$$= \frac{1}{\Gamma(n)} \left(\frac{m}{1+mc}\right)^k \quad (4.113)$$

Note that this expression for $c > 0$ is always smaller than the same expression for $c = 0$. We therefore conclude, for $n = k + 1$, which is the first n for which $f_k(n)$ is non-zero, that $f_k(n)$ for general $c > 0$ is smaller than $f_k(n)$ for $c = 0$. As it is opposite for greater n , there will be some n for which the $f_k(n)$ for $c > 0$ overtakes the value of $f_k(n)$ for $c = 0$.

Expected path length

The most straightforward way of calculating the expected path length $\ell(n) = \sum_k k f_k(n)/K(n)$ is by using the generating function $K_z(n) = \sum_{k=0}^{\infty} f_k(n) z^k$. We first obtain $\sum_k k f_k(n)$ by differentiating $K_z(n)$ with respect to z and then taking

$z = 1$. Using the expression in Equation 4.98, we obtain

$$\frac{\partial K_z(n)}{\partial z} \Big|_{z=1} = \sum_k k f_k(n) = \left(-\frac{c}{(c+z)^2} + \frac{c}{(z+c)^2} \frac{\Gamma(n+z\frac{p}{c}+p)}{\Gamma(n)\Gamma(1+z\frac{p}{c}+p)} + \right. \quad (4.114)$$

$$\left. \frac{z\frac{p}{c}\Gamma(n+z\frac{p}{c}+p)\psi(n+z\frac{p}{c}+p)}{(z+c)\Gamma(n)\Gamma(1+z\frac{p}{c}+p)} - \frac{z\frac{p}{c}\Gamma(n+z\frac{p}{c}+p)\psi(1+z\frac{p}{c}+p)}{(z+c)\Gamma(n)\Gamma(1+z\frac{p}{c}+p)} \right) \Big|_{z=1} \\ = \frac{\Gamma(n+p+z\frac{p}{c})(c^2+zp(c+z)\psi(n+p+z\frac{p}{c})-zp(c+z)\psi(z\frac{p}{c}+p+1))}{\Gamma(n)\Gamma(z\frac{p}{c}+p+1)c(c+z)^2} \quad (4.115)$$

$$- \frac{c}{(c+1)^2} \Big|_{z=1} \\ = \frac{\Gamma(n+p+\frac{p}{c})(c^2+p(c+1)\psi(n+p+\frac{p}{c})-p(c+1)\psi(\frac{p}{c}+p+1))}{\Gamma(n)\Gamma(\frac{p}{c}+p+1)c(c+1)^2} - \frac{c}{(c+1)^2} \quad (4.116)$$

$$= \frac{\Gamma(n+m_c)(c+m_c(\psi(n+m_c)-\psi(m_c+1)))}{(c+1)^2\Gamma(n)\Gamma(m_c+1)} - \frac{c}{(c+1)^2}, \quad (4.117)$$

where we used on the last step that $m_c = p + p/c$. Then, using Equation 4.92 to divide by $K(n)$, we obtain

$$\ell(n) = \frac{\sum_k k f_k(n)}{K(n)} = \frac{\frac{\Gamma(n+m_c)(c+m_c(\psi(n+m_c)-\psi(m_c+1)))}{(c+1)^2\Gamma(n)\Gamma(m_c+1)} - \frac{c}{(c+1)^2}}{\frac{c}{1+c} + \frac{\Gamma(n+m_c)}{(1+c)\Gamma(n)\Gamma(m_c+1)}} \quad (4.118)$$

$$= \frac{\frac{\Gamma(n+m_c)(c+1+m_c(\psi(n+m_c)-\psi(m_c+1)))}{(c+1)^2\Gamma(n)\Gamma(m_c+1)} - \frac{\Gamma(n+m_c)}{(c+1)^2\Gamma(n)\Gamma(m_c+1)} - \frac{c}{(c+1)^2}}{\frac{c}{1+c} + \frac{\Gamma(n+m_c)}{(1+c)\Gamma(n)\Gamma(m_c+1)}} \quad (4.119)$$

$$= \frac{\frac{\Gamma(n+m_c)(c+1+m_c(\psi(n+m_c)-\psi(m_c+1)))}{(c+1)^2\Gamma(n)\Gamma(m_c+1)}}{\frac{c}{1+c} + \frac{\Gamma(n+m_c)}{(1+c)\Gamma(n)\Gamma(m_c+1)}} - \frac{1}{1+c} \quad (4.120)$$

$$= \frac{\Gamma(n+m_c)(c+1+m_c\psi(n+m_c)-m_c\psi(m_c+1))}{(1+c)(c\Gamma(n)\Gamma(m_c+1)+\Gamma(n+m_c))} - \frac{1}{1+c} \quad (4.121)$$

$$= \frac{c+1+m_c\psi(n+m_c)-m_c\psi(m_c)+1}{(1+c)c\Gamma(n)\Gamma(m_c+1)/\Gamma(n+m_c)+1+c} - \frac{1}{1+c} \quad (4.122)$$

$$= \frac{c+1+m_c\psi(n+m_c)-m_c\psi(m_c+1)}{cr(n)+1+c} - \frac{1}{1+c}, \quad (4.123)$$

where we introduced

$$r(n) = \frac{(1+c)\Gamma(n)\Gamma(m_c+1)}{\Gamma(n+m_c)}, \quad (4.124)$$

which, using Equation 4.92, can be observed to equal $\frac{1}{K(n)-c/(1+c)}$. Equation 4.123 corresponds to Equation 4.13.

4.A.3 Generalization for increasing in-degree

Finally we consider the generalization where we allow the in-degree $\mu(n)$ of node n to depend on n (note that $\mu(n) \neq m(n)$, the latter being the *average* in-degree when then total number of nodes is n). More specifically, we will consider the case where $\mu(n)$ is any linear combination of integer or non-integer powers of n , which is finite and positive for all n and in which the largest power of n has an exponent $\alpha > 0$. We can write this as

$$\mu(n) = w_0 n^{\alpha_0} + w_1 n^{\alpha_1} + \dots + w_n n^\alpha, \quad (4.125)$$

for any set of constants w_0, w_1, \dots, w where $w > 0$ and where the constants $\alpha_0, \alpha_1, \dots$ are smaller than α . To explore the dynamics for large n , let us calculate the total number of links $M(n) = \sum_{l=1}^n \mu(l)$, which appears in the denominator of Equation 4.15, i.e.

$$\Delta_n f_{k+1}(n) = \mu(n) \frac{f_k(n) + c f_{k+1}(n)}{n + c \sum_{l=1}^n \mu(l)} \quad (4.126)$$

$$= \mu(n) \frac{f_k(n) + c f_{k+1}(n)}{n + c M(n)}, \quad (4.127)$$

Let us assume we can approximate $M(n)$ by an integral, i.e.

$$\int_1^n \mu(l) dl = \frac{w_0}{1 + \alpha_0} n^{\alpha_0+1} + \frac{w_1}{1 + \alpha_1} n^{\alpha_1+1} + \dots + \frac{w}{1 + \alpha} n^{\alpha+1} + M_0, \quad (4.128)$$

where M_0 is some constant. Using this approximation, we can write for $\mu(n)/M(n)$

$$\frac{\mu(n)}{M(n)} = \frac{w_0 n^{\alpha_0} + w_1 n^{\alpha_1} + \dots + w n^{\alpha}}{\frac{w_0}{1 + \alpha_0} n^{\alpha_0+1} + \frac{w_1}{1 + \alpha_1} n^{\alpha_1+1} + \dots + \frac{w}{1 + \alpha} n^{\alpha+1} + M_0} \quad (4.129)$$

$$= \frac{w_0 n^{\alpha_0 - \alpha - 1} + w_1 n^{\alpha_1 - \alpha - 1} + \dots + w n^{-1}}{\frac{w_0}{1 + \alpha_0} n^{\alpha_0 - \alpha} + \frac{w_1}{1 + \alpha_1} n^{\alpha_1 - \alpha} + \dots + \frac{w}{1 + \alpha} + M_0 n^{-\alpha - 1}} \quad (4.130)$$

$$= \frac{1}{n} \cdot \frac{w_0 n^{\alpha_0 - \alpha} + w_1 n^{\alpha_1 - \alpha} + \dots + w}{\frac{w_0}{1 + \alpha_0} n^{\alpha_0 - \alpha} + \frac{w_1}{1 + \alpha_1} n^{\alpha_1 - \alpha} + \dots + \frac{w}{1 + \alpha} + M_0 n^{-\alpha - 1}} \quad (4.131)$$

In the limit where n gets very large, this expression reduces to $\mu(n)/M(n) \rightarrow (1 + \alpha)/n$. Furthermore, as $M(n)$ increases with n by a power $\alpha + 1 > 1$, note also that in the same limit, $n/M(n) \rightarrow 0$. Hence dividing the numerator and denominator in Equation 4.126 by $M(n)$ gives, in the limit of large n

$$\Delta_n f_{k+1}(n) \approx \frac{(1 + \alpha) f_k(n) + c f_{k+1}(n)}{n}. \quad (4.132)$$

If instead we would have chosen $\mu(n) \propto u^n$, note that we would then approximate $M(n) \propto u^n / \log(u)$, so in the limit where n is large, in that case $m_n/M(n) \rightarrow \log(u)$ and $n/M(n) \rightarrow 0$. We would then obtain, in the limit of large n ,

$$\Delta_n f_{k+1}(n) \approx \log(u) \frac{f_k(n) + c f_{k+1}(n)}{c}. \quad (4.133)$$

In P. Persoon et al., 2021 we solve a similar equation, which arises when there is no cumulative advantage effect and when $\mu(n)$ increases linear in n . There the solution allow us to show that expected path length growth linear in n . Equation 4.133 however has an additional term with $f_{k+1}(n)$, yet we will briefly demonstrate this only affects the coefficient of linear expected path length. Let us first rewrite Equation 4.133 as

$$f_{k+1}(n+1) - f_{k+1}(n) \approx \log(u) \frac{f_k(n) + c f_{k+1}(n)}{c} \quad (4.134)$$

$$f_{k+1}(n+1) \approx \log(u) f_k(n)/c + \log(u) f_{k+1}(n) + f_{k+1}(n) \quad (4.135)$$

$$f_{k+1}(n+1) \approx \log(u) f_k(n)/c + (1 + \log(u)) f_{k+1}(n) \quad (4.136)$$

$$\frac{1}{1 + \log(u)} f_{k+1}(n+1) \approx f_{k+1}(n) + \frac{\log(u)}{c(1 + \log(u))} f_k(n). \quad (4.137)$$

In the following we will show that the leading term of the solution to this equation is given by

$$f_k(n) = \left(\frac{\log(u)}{c}\right)^k (1 + \log(u))^{n-k} \binom{n}{k} \quad (4.138)$$

$$= u_1^k u_2^{n-k} \binom{n}{k}, \quad (4.139)$$

where we introduced the short-hand notation $u_1 = \frac{\log(u)}{c}$ and $u_2 = 1 + \log(u)$. Rewriting Equation 4.137 in short-hand notation gives

$$\frac{1}{u_2} f_{k+1}(n+1) \approx f_{k+1}(n) + \frac{u_1}{u_2} f_k(n). \quad (4.140)$$

Let us substitute in the equation the expressions for $f_k(n)$, $f_{k+1}(n)$ and $f_{k+1}(n+1)$ using Equation 4.139. We then obtain

$$\frac{1}{u_2} u_1^{k+1} u_2^{n-k} \binom{n+1}{k+1} \approx u_1^{k+1} u_2^{n-k-1} \binom{n}{k+1} + \frac{u_1}{u_2} u_1^k u_2^{n-k} \binom{n}{k} \quad (4.141)$$

$$\binom{n+1}{k+1} \approx \binom{n}{k+1} + \binom{n}{k} \quad (4.142)$$

which is immediately satisfied given the recurrence relation for the binomial coefficient $\binom{n+1}{k+1} = \binom{n}{k+1} + \binom{n}{k}$. Note that, using that $\binom{n}{k} = 0$ for $k > n$, $f_k(n)$ also satisfies the initial condition that $f_k(n) = 0$ for $k > n$. Using the binomial theorem $\sum_{k=0}^n \binom{n}{k} x^k y^{n-k} = (x+y)^n$ and Equation 4.138, we can directly sum $f_k(n)$ over all $k \leq n$, thus obtaining for large n that

$$K(n) \approx \sum_{k=0}^n u_1^k u_2^{n-k} \binom{n}{k} \quad (4.143)$$

$$\approx (u_1 + u_2)^n \quad (4.144)$$

$$\approx \left(\frac{\log(u)}{c} + 1 + \log(u)\right)^n \quad (4.145)$$

$$\approx \left(1 + \frac{1+c}{c} \log u\right)^n. \quad (4.146)$$

Furthermore, the expressions in Equation 4.138 and 4.146 we can directly calculate the expected path length $\ell(n) = \sum_k k f_k(n) / K(n)$, which is

$$\ell(n) \approx \frac{1}{(u_1 + u_2)^n} \sum_{k=0}^n k u_1^k u_2^{n-k} \binom{n}{k} \quad (4.147)$$

$$\approx \frac{1}{(u_1 + u_2)^n} u_1 (u_1 + u_2)^{n-1} n \quad (4.148)$$

$$\approx \frac{u_1 n}{u_1 + u_2} \quad (4.149)$$

$$\approx \frac{n \log(u)}{c + (1+c) \log u}. \quad (4.150)$$

We therefore obtain, for a large number of nodes, when the cumulative advantage effect is nonzero and the in-degree depends exponentially on the number of nodes, that the expected path length then grows linear with the number of nodes.

4.B Derivations with longest paths

In this section we will focus on the case where we consider the longest paths as a subset of all unique paths, which is discussed in Section 4.3. We will mostly focus on the steps in 'first order n ' derivation of $H_k(n)$, yet also briefly discuss how these derivations may be extended to include higher orders.

4.B.1 First order approximation

We start this analysis from Equation 4.22, which says for an $n > n_k$

$$H_{k+1}(n) = n_k + \int_{n_k}^n \frac{H_k(n)^m}{n^m} dn. \quad (4.151)$$

Substituting the expression $H_k(n) = a_k - \frac{a_{k-1}^m}{(m-1)n^{m-1}}$ (appearing in Equation 4.23) gives

$$n_k + \int_{n_k}^n \frac{1}{n^m} \left(a_k - \frac{a_{k-1}^m}{(m-1)n^{m-1}} \right)^m dn = n_k + \left[\frac{1}{(m+1)a_{k-1}^m} \left(a_k - \frac{a_{k-1}^m}{(m-1)n^{m-1}} \right)^{m+1} \right]_{n_k}^n, \quad (4.152)$$

and substituting the integration limits becomes

$$n_k + \frac{1}{(m+1)a_{k-1}^m} \left(a_k - \frac{a_{k-1}^m}{(m-1)n^{m-1}} \right)^{m+1} - \frac{1}{(m+1)a_{k-1}^m} \left(a_k - \frac{a_{k-1}^m}{(m-1)n_k^{m-1}} \right)^{m+1}. \quad (4.153)$$

Expanding the $m+1$ power, selecting the constant terms and the largest term involving n , i.e. the term of $\mathcal{O}\left(\frac{1}{n^{m-1}}\right)$, gives

$$n_k + \frac{a_k^{m+1}}{(m+1)a_{k-1}^m} - \frac{a_k^m}{(m-1)n^{m-1}} - \frac{1}{(m+1)a_{k-1}^m} \left(a_k - \frac{a_{k-1}^m}{(m-1)n_k^{m-1}} \right)^{m+1}. \quad (4.154)$$

As this expression should equal $H_k(n) = a_{k+1} - \frac{a_k^m}{(m-1)n^{m-1}}$, we directly conclude this is satisfied for the term proportional to $1/n^{m-1}$ and for the constant term a_{k+1} we obtain

$$a_{k+1} = n_k + \frac{a_k^{m+1}}{(m+1)a_{k-1}^m} - \frac{1}{(m+1)a_{k-1}^m} \left(a_k - \frac{a_{k-1}^m}{(m-1)n_k^{m-1}} \right)^{m+1}. \quad (4.155)$$

This is Equation 4.24. As explained earlier, we approximate n_k by a_k . Hence substituting $n_k = a_k$,

$$a_{k+1} = a_k + \frac{a_k^{m+1}}{(m+1)a_{k-1}^m} - \frac{1}{(m+1)a_{k-1}^m} \left(a_k - \frac{a_{k-1}^m}{(m-1)a_k^{m-1}} \right)^{m+1}, \quad (4.156)$$

and dividing by a_k gives

$$\frac{a_{k+1}}{a_k} = 1 + \frac{a_k^m}{(m+1)a_{k-1}^m} - \frac{1}{(m+1)a_{k-1}^m a_k} \left(a_k - \frac{a_{k-1}^m}{(m-1)a_k^{m-1}} \right)^{m+1} \quad (4.157)$$

$$= 1 + \frac{a_k^m}{(m+1)a_{k-1}^m} - \frac{a_k^m}{(m+1)a_{k-1}^m} \left(1 - \frac{a_{k-1}^m}{(m-1)a_k^m} \right)^{m+1}. \quad (4.158)$$

Using that $a_k \propto \beta_1^k$, we rewriting this expression in terms of the $\beta_1 = \frac{a_{k+1}}{a_k} = \frac{a_k}{a_{k-1}}$,

$$\beta_1 = 1 + \frac{\beta_1^m}{m+1} - \frac{\beta_1^m}{m+1} \left(1 - \frac{1}{(m-1)\beta_1^m}\right)^{m+1}. \quad (4.159)$$

This is Equation 4.25 in the paper. Expanding the part in brackets to third order gives

$$\beta_1 = 1 + \frac{\binom{m+1}{1}}{(m+1)(m-1)} - \frac{\binom{m+1}{2}}{(m+1)(m-1)^2\beta_1^m} + \frac{\binom{m+1}{3}}{(m+1)(m-1)^3\beta_1^{2m}} \dots \quad (4.160)$$

$$= 1 + \frac{1}{m-1} - \frac{m}{2(m-1)^2\beta_1^m} + \frac{m}{6(m-1)^2\beta_1^{2m}} + \dots \quad (4.161)$$

Equation 4.161 (note this corresponds to Equation 4.26) shows more clearly that, choosing a greater estimate for β_1 on the right-hand side will lead to an underestimate of the second order term, which is negative. Choosing a greater estimate for β_1 on the right-hand side will therefore lead to an overestimate of the total of the right-hand side of the equation. Let us suppose we substitute a greater estimate for β_1 only on the right-hand side. As the left hand side is simply β_1 , we conclude that we obtain an overestimate for this β_1 when the total of the right-hand side is overestimated. Substituting the greater estimate $\beta_1 = 1 + \frac{1}{m-1}$ on the right hand side gives

$$\beta_1 \approx 1 + \frac{(1 + \frac{1}{m-1})^m}{m+1} - \frac{(1 + \frac{1}{m-1})^m}{m+1} \left(1 - \frac{1}{(m-1)(1 + \frac{1}{m-1})^m}\right)^{m+1}. \quad (4.162)$$

We will expand these expressions for large $m-1$ to specifically consider the coefficients of the first and second order in $1/(m-1)$. We will first expand the term

$$y_0(m) = \frac{(1 + \frac{1}{m-1})^m}{m+1} \quad (4.163)$$

$$= \frac{(1 + \frac{1}{m-1})^{m-1+1}}{m-1+2} \quad (4.164)$$

$$= \frac{1}{m-1} \frac{(1 + \frac{1}{m-1})^{m-1}}{1 + \frac{2}{m-1}} \left(1 + \frac{1}{m-1}\right) \quad (4.165)$$

Using that $(1+1/x)^x \rightarrow e$ for large x , we conclude that for large $m-1$, $y_0(m) \rightarrow \frac{e}{m-1}$ to first order $1/(m-1)$. Next we consider the term

$$y_1(m) = \frac{(1 + \frac{1}{m-1})^m}{m+1} \left(1 - \frac{1}{(m-1)(1 + \frac{1}{m-1})^m}\right)^{m+1}. \quad (4.166)$$

Introducing the shorthand notation $z_m = (1 + \frac{1}{m-1})^m$, we can rewrite this as

$$y_1(m) = \frac{z_m}{m+1} \left(1 - \frac{1}{(m-1)z_m}\right)^{m+1} \quad (4.167)$$

$$= \frac{z_m}{m+1} \left(1 - \frac{1}{(m-1)z_m}\right)^{m-1+2} \quad (4.168)$$

$$= \frac{z_m}{m+1} \left(1 - \frac{1}{(m-1)z_m}\right)^{(m-1) - \frac{z_m}{z_m} + 2} \quad (4.169)$$

$$= \frac{z_m}{m+1} \left(1 + \frac{1}{-(m-1)z_m}\right)^{-\frac{(m-1)z_m}{-z_m}} \left(1 - \frac{1}{(m-1)z_m}\right)^2 \quad (4.170)$$

We see the first part of this multiplication, $z_m/(m+1)$, is just $y_0(m)$. As we showed earlier, for large $m-1$ this is approximated by $\frac{e}{m-1}$. For the second term of the multiplication, we again use that $(1+1/x)^x \rightarrow e$ for large x , (yet this time choosing $x = z_m(m-1)$). Note that, while taking this limit, in the exponent of this term we are left with $-1/z_m$, which goes for large $m-1$ to $-1/e$. Concluding, for second part of the multiplication we obtain $e^{-e^{-1}}$. Finally, the third part, $(1 - \frac{1}{(m-1)z_m})^2$, simply goes to 1. In total, we therefore obtain that $y_1(m) \rightarrow \frac{e^{1-e^{-1}}}{m-1}$ to first order in $1/(m-1)$. The expressions that allow us to obtain the terms of order $1/(m-1)^2$ for $y_0(m)$ and $y_1(m)$ quickly become very lengthy. We therefore calculate these using the computer program Mathematica. We then obtain for the expressions up to order $1/(m-1)^2$

$$y_0(m) = \frac{e}{m-1} - \frac{3e}{2(m-1)^2} + \mathcal{O}\left(\frac{1}{(m-1)^3}\right) \quad (4.171)$$

$$y_1(m) = \frac{e^{1-\frac{1}{e}}}{m-1} - \frac{e^{-1-\frac{1}{e}}(1+3e+3e^2)}{2(m-1)^2} + \mathcal{O}\left(\frac{1}{(m-1)^3}\right). \quad (4.172)$$

Noting that, using Equation 4.162, we can write $\beta_1 \approx 1 + y_0(m) - y_1(m)$, we may summarize our results as

$$\beta_1 \approx 1 + \frac{e - e^{1-\frac{1}{e}}}{m-1} + \frac{e^{-1-\frac{1}{e}}(1+3e+3e^2) - 3e}{2(m-1)^2} \quad (4.173)$$

Noting that $e - e^{1-\frac{1}{e}} \approx 0.84$, we conclude that the first order term is a factor 0.84 smaller than the first order term in β_0 . Earlier we explained that the expected path length of the longest paths $\ell(n) \approx k_n$, where $k_n \approx \frac{\log(n(\beta_1-1)+1)}{\log(\beta_1)} - 1$. This indicates that, to calculate the coefficient d_m of the log number of nodes (see Equation 4.8), we need to calculate $1/\log(\beta_1)$. To obtain a simple expression for d_m , we therefore expand $1/\log(\beta_1)$ for large $m-1$. We note first that for large $m-1$ we can approximate, for any parameters γ_0 and γ_1 ,

$$\frac{1}{\log\left(1 + \frac{\gamma_0}{m-1} + \frac{\gamma_1}{(m-1)^2}\right)} \approx \frac{m-1}{\gamma_0} + \frac{\gamma_0^2 - 2\gamma_1}{2\gamma_0^2} \quad (4.174)$$

$$\approx \frac{m}{\gamma_0} - \frac{1}{\gamma_0} + \frac{1}{2} - \frac{\gamma_1}{\gamma_0^2}. \quad (4.175)$$

Hence choosing $\gamma_0 = e - e^{1-\frac{1}{e}}$ and $\gamma_1 = \frac{1}{2}\left(e^{-1-\frac{1}{e}}(1+3e+3e^2) - 3e\right)$ using Equation 4.173, we obtain

$$d_m = \frac{1}{\log(\beta_1)} \quad (4.176)$$

$$\approx \frac{m}{e - e^{1-\frac{1}{e}}} - \frac{1}{e - e^{1-\frac{1}{e}}} + \frac{1}{2} - \frac{e^{-1-\frac{1}{e}}(1+3e+3e^2) - 3e}{2(e - e^{1-\frac{1}{e}})^2} \quad (4.177)$$

$$\approx 1.2m - 0.6, \quad (4.178)$$

which is the approximation used earlier.

Chapter 5

The knowledge mobility of renewable energy technology

Peter Persoon, Rudi Bekkers and Floor Alkemade

This chapter is accepted and being prepared for publication in Energy Policy

Abstract

In the race to achieve climate goals, many governments and organizations are encouraging the local development of Renewable Energy Technology (RET). The spatial innovation dynamics of development of a technology partly depends on the characteristics of the knowledge base on which this technology builds, in particular the analyticity and cumulateness of knowledge. Theoretically, greater analyticity and lesser cumulateness are positively associated with more widespread development. In this study, we first empirically evaluate these relations for general technology and then systematically determine the knowledge base characteristics for a set of 14 different RETs. We find that, while several RETs (photovoltaics, fuel-cells, energy storage) have a highly analytic knowledge base and develop more widespread, there are also important RETs (wind turbines, solar thermal, geothermal, and hydro energy) for which the knowledge base is less analytic and which develop less widespread. Likewise, the technological cumulateness tends to be lower for the former than for the latter group. This calls for regional and country-level policies to be specific for different RETs, taking for a given RET into account both the type of knowledge it builds on as well as the local presence of this knowledge.

Keywords: Renewable Energy Technology, Knowledge Base, Geography, Patents

5.1 Introduction

The widespread development and use of Renewable Energy Technologies (RETs) is an essential part of the transition towards a carbon-free society (IPCC, 2014). The ability of a country or region to participate in the development of a technology not only depends on the locally available knowledge and capabilities (Li et al., 2020), but also on the characteristics of *the knowledge base of that technology*. More specifically, Binz and Truffer (Binz & Truffer, 2017) argue that it is typically easier to enter in knowledge fields with a more global (and 'footloose') knowledge base

when compared to knowledge bases that are more local (and 'sticky'). These characteristics of the knowledge base have been linked to different modes of knowledge production; global, footloose knowledge to a 'Science-Technology and Innovation (STI) mode' observed in science-based industries that lean very much on analytical knowledge, and local, sticky knowledge bases to a 'Doing, Using and Interacting (DUI) mode' observed in engineering-based industries that lean very much on synthetic knowledge (Asheim et al., 2016; Jensen et al., 2007). RETs may thus differ in the extent to which their development spreads over countries or regions, i.e., the mobility of their knowledge base. Where the development of some RETs may take place in STI-mode, widespread and expanding, the development of other RETs may take place in DUI-mode, concentrated and difficult to relocate. This has implications for countries that seek to move closer to the knowledge frontier through technology and R&D investments as this may be easier for more footloose technologies (Keller, 2004). Understanding the knowledge base characteristics of renewable energy technologies—in particular the knowledge dimensions relating to the spatial dynamics of innovation—is thus pivotal input for targeted and evidence-based renewable energy policies.

Earlier studies analyzing RETs as a single technology class find that RETs on average build more on analytical and geographically distant knowledge than other technologies (Ocampo-Corrales et al., 2020), and that they benefit greatly from knowledge flows transcending national borders (Garrone et al., 2014; J. Noailly & Ryfisch, 2015). However, recent studies at the more detailed level of individual technologies find considerable heterogeneity in the extent to which RETs build on analytical knowledge (Hötte et al., 2020; P. G. J. Persoon et al., 2020). For example, the science dependence of some RETs, such as wind turbines, is relatively low, and closer to fossil fuel based energy technologies, whereas photovoltaics and non-fossil fuels are characterized by a high science dependence. Similar variations have been observed in other dimensions of the knowledge base that may affect the place-dependence of RETs such as the cumulateness (P. Persoon et al., 2021), which is associated with greater geographical concentration of innovative activities (Breschi et al., 2000; Malerba et al., 1997). Building on the framework outlined by Binz and Truffer (Binz & Truffer, 2017), we systematically investigate these different characteristics of the knowledge base of RETs in order to assess whether these technological are more local or global in nature. More specifically, we map analyticity, cumulateness, and knowledge mobility for the knowledge base of 14 different RETs.

The remainder of this paper is structured as follows. In Section 5.2 we discuss the theoretical background of the mentioned knowledge dimensions and our expectations for the different RETs. Then in Section 5.3 we explain how we measure the knowledge dimensions and distinguish the different RETs. Subsequently, we report our main observations in Section 5.4, discuss some deeper implications and shortcomings in Section 5.5 and end with a number of conclusions and policy recommendations in Section 5.6.

5.2 Theory

The process of knowledge creation and innovation in a certain technology depends for an important part on characteristics of the body of knowledge on which this

technology builds (Asheim & Coenen, 2005; Breschi et al., 2000), henceforth referred to as the 'knowledge base' of a technology. In the following, we will discuss three dimensions of the knowledge base which can theoretically be linked to spatial dynamics of innovation: the analyticity (Section 5.2.1) the cumulateness (Section 5.2.2), and the knowledge mobility (Section 5.2.3). We then discuss our expectations for these dimensions for the different RETs (Section 5.2.4).

5.2.1 Analyticity of knowledge

Knowledge bases are described to consist of three types of knowledge: analytic, synthetic, and symbolic knowledge (Asheim & Coenen, 2005; Moodysson et al., 2008). In this context, analytic knowledge is understood to be science-based, created in deductive processes based on formal models. Synthetic knowledge is understood as engineering-based, created through the application of existing knowledge or through inductively combining existing knowledge. Finally, symbolic knowledge is characterized as cultural or artistic knowledge.¹ As the knowledge base of technologies often contains multiple types of knowledge, it is more instructive to think about the extent to which it is analytic as a spectrum or a scale. In this line of thinking, we define the analyticity of a knowledge base as the extent to which it consists of analytic knowledge.

The analyticity of knowledge has been associated with several other theoretical dimensions of knowledge. First, where analytical knowledge is associated naturally with basic research, i.e. research aimed at truth-finding, synthetic knowledge is associated with applied research, i.e. research aimed at solving practical problems (Bentley et al., 2015; OECD, 2015). Technologies that strongly depend on analytic knowledge are therefore understood to have stronger ties with the natural sciences. While closely related, basic and analytic (or applied and synthetic) cannot be considered synonyms: basic research may occasionally produce synthetic knowledge, and vice versa. Second, and closely related, where analytic knowledge is universal and theoretical, synthetic knowledge is context-specific and practice related (Moodysson et al., 2008). It is therefore expected that it is more difficult to work with synthetic knowledge outside the context in which it was developed, that is synthetic knowledge is stickier and place dependent. Third, analytic knowledge is often associated with codified knowledge and synthetic knowledge with tacit knowledge. However, here too, there are certainly exceptions. Not all published work is easy to fully understand or reproduce without the aid of those that produced the work. Authors have therefore argued that there sometimes is a tacit element to analytic knowledge as well (Moodysson et al., 2008).

5.2.2 Technological cumulateness

Knowledge bases can also be characterized by their 'technological cumulateness', the idea that today's technologies are developed by building on the insights from yesterday's technologies and will themselves be used to develop the technologies of tomorrow (Breschi et al., 2000; Trajtenberg et al., 1997). Perspectives on the exact meaning of technological cumulateness however vary, for an overview of this

¹In this research we will mostly focus on the distinction between analytic and synthetic knowledge.

discussion we refer to (P. Persoon et al., 2021). In this work, we understand a technological development to be cumulative when a later technological result depends on an earlier technological result. In the context of technological knowledge, we broadly interpret this dependency as the usage, modification or improvement of earlier ideas. In this line of thinking, we understand the knowledge base of a technology to be more cumulative when the developments in this technology are more cumulative. This allows us to define the cumulateness of a technology as the extent to which developments within this technology are cumulative. In other words, the more a technology builds on its earlier developments, the greater its cumulateness.

Technological cumulateness is often mentioned as a defining characteristic of a *technological regime*, which is a description of the relevant environment or circumstances for companies and organizations to innovate (Breschi et al., 2000; Nelson & Winter, 1977). When a technological regime is characterized by high cumulateness, established parties largely dominate innovative activities and it is relatively hard for new parties to enter. Highly cumulative technologies allow firms or organisations to gain absorptive capacity through learning and specialization (W. M. Cohen & Levinthal, 1990). Within a technological regime therefore, greater cumulateness is associated with a greater appropriability of innovation and a greater geographical concentration of innovative activities (Malerba & Orsenigo, 1996; Malerba et al., 1997).

In an earlier contribution, where the cumulateness was explicitly approached as the extent to which a technology builds on its earlier developments, we established that the cumulateness of a technology increases approximately linearly with the size of its knowledge base, at a technology-specific rate (P. Persoon et al., 2021). Especially when cumulateness is compared across technologies, this suggests that next to considering the cumulateness of a technology, it will be useful to consider the rate at which the cumulateness increases, i.e., the cumulateness relative to the size of the knowledge base.

When cumulative developments stretch over longer periods of time, the products associated with a technology tend to become 'more complex', meaning that the number of interrelated (functional) parts of a product architecture increases. Technological complexity is therefore often associated with greater cumulateness. The complexity of technologies is in the literature however mostly approached anecdotally or on a case-to-case basis, as there is no general agreement on a single objective measure for complexity (Vaesen & Houkes, 2017).

5.2.3 Knowledge mobility

Where some types of knowledge travel easily from one place to another, other types are bound to a certain location. In order to investigate this dimension of knowledge, we define the knowledge mobility as the extent to which knowledge travels geographically. By geographical traveling, we mean that knowledge developed in one location is subsequently used or applied in another location, where the two locations are separated by a geographical distance. High knowledge mobility then corresponds to knowledge that travels with ease to more distant locations, i.e., 'footloose knowledge'. Low knowledge mobility corresponds to Knowledge that travels less easily, i.e. 'sticky knowledge'. A highly mobile body of knowledge is thus expected to travel farther, in other words, we expect mobile knowledge to be more widespread (or less

concentrated) than sticky knowledge.

Knowledge bases characterized by greater analyticity are expected to be more mobile (or 'footloose') (Asheim et al., 2011; Herstad et al., 2014). A motivation for this first expectation is the universality and theoretical nature of analytic knowledge, which almost per definition implies time, location, and application independence. The context specificity and practical nature of synthetic knowledge on the contrary make it more time, location, and application bound. Another motivation is the supposed association with codified knowledge: what is written down travels easier than know-how fixed in the minds of experts (Gertler, 2003; Lundvall & Johnson, 1994). As mentioned earlier though, this association is also criticized. These motivations also count when the causality is reversed: when innovative activities are fixed and concentrated geographically, there may be less need to formalize or rationalize findings because knowledge is communicated orally, developed during collaboration and hence may remain largely tacit and fragmented. A gradual shift towards knowledge more synthetic in nature is thus expected when engineers work close together. Likewise, when collaborators are forced to communicate their results at a distance, it may stimulate them to formalize or rationalize their implicit ideas or intuitions.

Knowledge bases characterized by higher cumulateness are expected to be stickier (Herstad et al., 2014). A motivation for this second expectation is the expected greater geographical concentration of innovative activities in technological regimes characterized by high cumulateness. With greater geographical concentration, we expect the development to be less widespread and hence to travel shorter distances. Note that this relation too can be reversed, namely that the knowledge is concentrated because it is sticky. Another motivation for this second expectation comes from the association between cumulateness and technological complexity: technologically complex knowledge does not travel well (Balland & Rigby, 2017). Working with or improving a complex system from a distance is challenging, because it becomes more difficult to experiment or interact with the system.

5.2.4 Knowledge dimensions of RETs

In this research, we aim to investigate how the knowledge dimensions vary for different Renewable Energy Technologies (RETs). While a 'technology' can be approached or characterized from many different angles, we will in this contribution largely focus on the knowledge properties of technologies, hence approaching the different RETs as distinct bodies of knowledge. The knowledge properties may cover various aspects of the technology, for example, how the technology operates or how it is constructed. While the purpose of the various RETs largely coincides (enable the generation of renewable energy), the renewable energy sources that the various RETs exploit (and thus their working principles) fundamentally differ. Following the International Renewable Energy Agency (IRENA), we distinguish between geothermal, hydropower, ocean, wind, solar thermal, solar photovoltaic, and bio-energy (IRENA, 2018). In addition, we include a number of enabling technologies allowing for the storage of energy such as hydrogen technology, and three energy-related technologies that are not entirely renewable yet may help reduce CO₂ emissions: nuclear energy, carbon capture & storage (ccs) and clean combustion. We will provide a more precise list of the individual RETs in the next section.

Earlier contributions have indicated that RETs generally build strongly on sci-

entific knowledge, suggesting a highly analytic knowledge base (Ocampo-Corrales et al., 2020). In agreement with this finding, innovative activities related to RETs are observed to take place on ever-larger geographic scales (Garrone et al., 2014; J. Noailly & Ryfisch, 2015). At same time the knowledge bases are known to vary greatly across different RETs (Barbieri et al., 2020) and across energy technology in general (Nemet, 2012). More specifically, we know there is a large variation across RETs in the extent to which the knowledge base is science-based (Hötte et al., 2020; P. G. J. Persoon et al., 2020). Where photovoltaics, non-fossil fuels and to some extent fuel-cells and hydrogen technology were found to be more science-based, wind turbines, hydroelectric and geothermal energy were found to be less science-based. The more a RET depends on science, the more analytic its knowledge base, the greater a knowledge mobility we expect for these technologies.

While the development of different RETs has been studied in numerous contributions, it appears that the current literature lacks a systematic comparison of the cumulateness across different RETs. Even though the size of the knowledge base varies across RETs, this does not automatically translate to a similar variation in cumulateness (P. Persoon et al., 2021). The closely related technological complexity however does appear to vary largely across RETs. Interpreting a larger technological complexity for systems with many interdependent parts, RETs such as wind turbines, geothermal energy, nuclear fission, and energy from sea are identified as rather complex (Huenteler, Ossenbrink, et al., 2016), more complex than photovoltaics and non-fossil fuels.² The variation in technological complexity suggests there may be large variation across RETs in cumulateness too (though this needs empirical validation). As the knowledge bases characterized by high cumulateness tend to be stickier, we expect the higher cumulateness RETs to show a lower knowledge mobility. Taking a slightly different perspective, Binz, Tang and Huenteler distinguish between 'complex engineered systems for specialized users' and 'standardized mass-manufactured goods', wind-turbines again being an example of the former and household energy storage systems, stationary fuel-cells and photovoltaics an example of the latter (Binz et al., 2017). Based on their findings about photovoltaics, they expect the life-cycle dynamics of the latter group to be more 'spatially fluid'.

Summarizing, we expect to observe greater knowledge mobility for RETs with a stronger dependence on science *and* RETs characterized by lower cumulateness.

5.3 Methodology

This section presents the methods used to measure analyticity, cumulateness and knowledge mobility for RETs. First, we discuss our data and present indicators for the knowledge dimensions. Subsequently, we discuss our selection of various RETs and present some descriptive statistics.

²In some cases the technological complexity varies with different applications of a technology. For example for solar thermal energy, the systems in domestic use are limited to elements that efficiently capture and store heat, and are therefore relatively simple, whereas the systems used in power plants are typically larger, contain more different elements and have the additional features of concentrating the heat and transforming it to electric power, making these systems far more complex.

5.3.1 Patents

Earlier approaches to measuring the analytic-synthetic knowledge distinction were often based on data from questionnaires or professional occupations (Martin, 2012; Moodysson et al., 2008; Plum & Hassink, 2012). While useful, these data are largely an *indirect* measure of knowledge characteristics, because they are based on the characteristics of the people that use or produce the knowledge, instead of the knowledge itself. In this contribution, we aim to *directly* measure the knowledge characteristics by studying codified forms of knowledge, more specifically, patent data.

Patent data directly represent technological knowledge, containing a wealth of detailed information about both the technological content as well as the inventor or applicant. Furthermore, the citations in patents, both to other patents and scientific literature, to some extent allow us to proxy knowledge connections and flow. While patent data offer a unique opportunity to quantitatively study novel and relevant technological knowledge development, there are also some limitations. Not all technology is patented and not all patents represent relevant technological developments. While acknowledging these disadvantages, we believe that for the purpose of understanding RET development there is a great potential for patent data.

A possible criticism of the usage of patent data to proxy the analytic-synthetic distinction is the supposed association with the codified-tacit distinction: as patents are codified knowledge, we risk observing analytic knowledge only. However, as mentioned earlier, the association with codified-tacit distinction is also criticized, and we strongly believe that patents, a key element of engineering practices, may equally well contain a large degree of synthetic knowledge. Our approach is, therefore, that within codified knowledge, there may be different degrees of analytic knowledge. More specifically in the context of technological knowledge, the more a body of codified knowledge can be associated with scientific activity, the greater we will interpret its degree of analytic knowledge.

Finally, we shortly comment on the geography of patents. In this research we will do a separate analysis for patents from the EPO (European Patent Office) and the USPTO (United States Patent and Trademark Office), henceforth 'EP patents' and 'US patents' respectively. There are two reasons for this choice. First, different patent offices, but in particular EPO and USPTO, have institutionalized different rules for citation, hence limiting the analysis of knowledge connections to one patent office may give biased results. Second, an applicant files a patent with an office if there is market potential in the geographical jurisdiction of that office. As we are interested in the worldwide geography of innovation, we do not want to limit the analysis to a single geographical jurisdiction.

5.3.2 Indicators

For the analysis of analyticity, we will mostly use the scientific character of this type of knowledge. To proxy for a given technology the dependence on science and the scientific content of the knowledge base we define the following indicators:

- The *science dependence*(sd) of a technology is defined as the average number of references to scientific literature per patent. A reference in a patent to

a scientific source can be interpreted as a dependency link, suggesting that scientific knowledge was somehow relevant in the content of the patent. The more scientific sources a patent therefore refers to, the more we expect it to be science-based. We therefore take the science dependence as an indicator of analytic knowledge.

- The *science dependence fraction*(sdf) of a patent is defined as its number of references to scientific literature divided by its total number of references. To obtain the sdf of a technology, we take the sdf of each patent in that technology and take the average. Hence where the sd is based on the absolute number of references, the sdf is based on the relative number of references, thus taking into account variation across patents and technologies in the number of references. A similar indicator was earlier used in (Hötte et al., 2021; Hötte et al., 2020).
- The *university fraction*(uf) of a technology is defined as the number patents in that technology for which the inventor or applicant is university³ affiliated divided by the total number of patents in that technology. When the inventor is affiliated with a university, we expect the patent to be based more on scientific knowledge than the average patent from non-scientific organizations. We therefore take the university fraction to be an indicator of analytic knowledge.

To proxy the cumulateness we will use

- The *internal dependence*(id) of technology is defined as the average number of internal references per patent. An internal reference is a reference in a patent to a patent within the same technology, which can be interpreted as a dependency link from the technology to itself. Cumulateness can be interpreted as the extent to which a technology builds on itself. This indicator was earlier used in (P. Persoon et al., 2021). For an approach based on general references, we refer to (Apa et al., 2018).
- The *internal dependence fraction*(idf) of a patent is defined as its number of internal references divided by its total number of patent references⁴. To obtain the idf of a technology, we take the idf of each patent in that technology and take the average. Hence where the id is based on the absolute number of references, the idf is based on the relative number of references, thus taking into account variation across patents and technologies in the number of references.
- The *relative internal dependence*(rid) of a technology is defined as the internal dependence of that technology relative to its total number of patents, or equivalently, the number of internal references per *patent squared*. As explained earlier, the internal dependence tends to increase linearly with the number of patents. When we compare technologies with a different number of patents or when we are interested in the rate at which the cumulateness increases, it is therefore useful to additionally consider the cumulateness per patent.

³As we will see later it more correct to speak of university-related organizations

⁴Alternatively, we can also include the references to scientific and/or other sources in this total. However, in this contribution we choose to define it as a fraction of patent references only, so that we can consider it to be independent of the sdf

For the knowledge mobility we define the following indicators:

- The *inter-patent distance*(ipd) of a technology is defined as the average geographic distance between each pair of patents within that technology. From the inventor or applicant addresses in patents we can create an overview of the approximate⁵ locations of inventing. The mutual distances between patents can thus be used to proxy the geographical spread of inventing in a certain technology.
- The *reference distance*(rd) of a patent is defined as the average geographic distance between that patent and the (set of) patent(s) it refers to⁶. The reference distance of a technology is defined as the average reference distance per patent. Where the inter-patent distance proxies the geographical spread, it does not directly proxy the possible knowledge flow between distant places. With the reference distance we therefore additionally consider the kilometers covered by references to obtain a better estimate of the actual movement of knowledge. Note however that the reference distance also includes references to other technologies, thus to some extent also measuring the knowledge flow of other technologies.⁷

Note that all of these indicators can be determined for technologies (i.e. groups of patents) and a selection of these indicators can also be determined on the level of individual patents. In the first part of our analysis, we will use the indicators acting on the level of individual patents to establish a baseline and demonstrate more general relations between analyticity, cumulateness and knowledge mobility (where the patents are not necessarily confined to the considered technologies).

This analysis is mainly based on data from Patstat (spring 2020 edition) focusing on European and US patents. As a consequence, there are a number of subtleties involved with the actual measurement of the indicators:

1. We count as a 'patent' each unique DOCDB patent family, where an 'EP patent' represents each unique family with an EPO patent application and likewise for 'US patent' but then for USPTO applications⁸.
2. To identify the references to scientific literature we use the type 's' classification of the cited non-patent-literature (NPL), which signals articles in journals and periodicals. Where in Patstat the NPL appears to be classified rather well for EP patents, for the US patents the large majority of NPL, probably due to a lacking of rich structure in references, is classified in the general category 'a' (abstract of no specific kind). To obtain a better indication of which fraction of the cited NPL is actually scientific, we use the database by Marx and Fuegi (Marx & Fuegi, 2020) which links the references in patent applications to scientific publications and is accurate for US patents.

⁵That is approximate, as there is no guarantee the actual process of inventing took place at the mentioned address

⁶We take the reference distance of a patent which does not refer to any other patent to be undefined

⁷Excluding the references to other technologies can be demonstrated to result in an indicator very closely related to the inter-patent distance

⁸The Patstat records of USPTO applications are biased to granted patents before the year 2000. As our focus is not on the time development we expect this to be a minor issue for our purpose

3. To identify the inventors or applicants affiliated with a university we use the automatized sector allocation in Patstat of persons (Magerman et al., 2006; Van Looy et al., 2006). This classification however allows an applicant to be allocated to multiple sectors. For the university fraction, we include each patent where at least of one the allocations is the 'UNIVERSITY' sector. We therefore also include organizations closely related to the university, making it more correct to speak of 'university-related organizations'.
4. To link the patents to geographical coordinates we use the 'Geocoding of worldwide patent data' database (shortly 'Geocoding') constructed by Rassenfosse, Kozak, and Seliger (de Rassenfosse et al., 2019) based on the applicant or inventor addresses. The Geocoding database is limited to first filed patent applications, which we linked back to patent families using Patstat. This research is based on the Geocoding table with inventor addresses. Yet, as the makers of the database acknowledge, disambiguation of inventors and applicants is generally challenging and a research task on its own. Indeed a quick comparison with the table bases on applicant addresses does not seem to amount to substantially different results.
5. The addresses of inventors of EP patents are not necessarily confined to Europe, and likewise for US patents and the US. The typical reference distance of a EP patent with an inventor from the US however structurally differs from that with an inventor from Europe: the reference distance is location-specific. While these variations are expected to average out when the number of patents in a technology is large, this effect may be disproportional for technologies with a smaller number of patents. To demonstrate this effect, we determine the reference distance from EP patents with US inventors and vice versa and compare these to the reference distances of EP (US) patents with European (US) inventors in Appendix 5.B. To account for this effect, when we determine the reference distance of EP patents we sub-select the patents with an inventor in Europe. Likewise, when we determine the reference distance of US patents we sub-select the patents with an inventor from the US. These sub-selections contain for most of the technologies considered the majority of patents. Note that, while we sub-select patents based on the location of the inventor, the references in these may still be to patents from inventors located anywhere in the world.

5.3.3 Technology selection and descriptive statistics

We base our selection of energy generating technologies on the set of renewable energy sources identified by the International Renewable Energy Agency IRENA (IRENA, 2018), including geothermal, hydropower, ocean, wind, solar, and bioenergy. As mentioned earlier, these energy-generating technologies are complemented with a set of technologies relating to energy storage and a set of technologies that may not be considered fully renewable but nonetheless help reducing greenhouse gas emissions, such as nuclear energy, clean combustion, and carbon capture and storage (ccs). For an overview see Table 5.3.1. To identify the patents associated with these (partial) RETs we use the Cooperative Patent Classification (CPC) used by both EPO and USPTO, or more specifically the CPC tagging scheme 'Y02' which

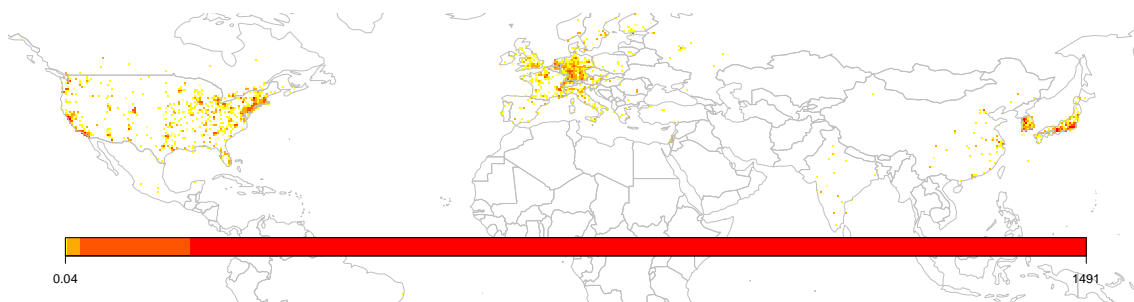


Figure 5.3.1: *Worldwide distribution of photovoltaics US patents based on inventor or applicant address* We plot the number of US patents per grid-cell for a grid defined for each longitudinal and latitudinal half degree. The scale chosen for the color coding of the cells is logarithmic (see scale-legend).



Figure 5.3.2: *Worldwide distribution of wind turbine US patents based on inventor or applicant address* Similar to Figure 5.3.1, except for adjusting the color scale (the maximum is here 241 patents).

identifies technologies with the potential to mitigate climate change (Veeffkind et al., 2012). Each of these RETs corresponds then to a collection of patents classified on the group or subgroup level in CPC. The various technologies and corresponding CPC descriptions are shown in Table 5.3.1, including the symbols which represent them in later figures. Note that a substantial number of EP and US patents are members of the same patent family, hence there is a substantial overlap between both data sets.

In Table 5.3.2 we include the descriptive statistics for a number of indicators discussed in the previous section. All of these indicators are only positive and characterized by distributions skewed towards the value zero, which is in line with the observation that the standard deviations are of the same order as the averages. The variation across technologies is substantial, especially across the analyticity and cumulativeness indicators. The science dependence of non-fossil fuels is much higher than that of the other RETs. This is in line with earlier findings (Hötte et al., 2020; P. G. J. Persoon et al., 2020) and may be related to the strong link of non-fossil fuels to fields such as (Applied) Microbiology, Biochemistry, and Molecular Biology. While we will measure and plot the indicator values for non-fossil fuels, we will not include it in data fits and statistical analysis.

To explore the geographical distributions of inventive activity in RETs worldwide, we use the geographical coordinates to plot the number of US patents (using the inventor or applicant address in the patents) in a grid defined for each longitudinal and latitudinal half degree. We do this for photovoltaics and wind-turbines respectively in Figures 5.3.1 and 5.3.2. Because the patenting activity is distributed









Technology	CPC description	CPC code	EP patents	US patents	in common
	Geothermal Energy	Y02E 10/1	495	1088	240
	Hydro Energy	Y02E 10/2	1865	6223	1159
	Energy from the sea, e.g. using wave energy or salinity gradient	Y02E 10/3	1228	2624	902
	Solar thermal energy, e.g. solar towers	Y02E 10/4	5425	11247	3034
	Photovoltaic energy (photovoltaics)	Y02E 10/5	14947	31490	12492
	Wind energy (wind turbines)	Y02E 10/7	10112	16454	7471
	Combustion technologies with mitigation potential (clean combustion)	Y02E 20	4956	7646	3575
	Nuclear fission reactors	Y02E 30/3	1337	4325	1038
	Technologies for an efficient electrical power generation, transmission or distribution (electric grids)	Y02E 40	2171	4031	1718
	Technologies for the production of fuel of nonfossil origin (non-fossil fuels)	Y02E 50	6310	9625	4548
	Energy storage using batteries, capacitors, thermal or mechanical systems.	Y02E 60/1	8858	17502	7166
	Hydrogen Technology	Y02E 60/3	4029	7307	3220
	Fuel cells	Y02E 60/5	3501	7254	3152
	Carbon capture and storage (ccs)	Y02C	3791	6297	3091

Table 5.3.1: *Symbols, CPC codes and total number of patents of selected RETs* To the CPC descriptions we added for some technologies a shortname (in brackets). The final column indicates the number of EP and US patents of which the family is the same.














RET	science dependence		internal dependence		inter-patent distance in km		reference distance in km	
	EP	US	EP	US	EP	US	EP	US
	0.11(0.54)	0.63(3.57)	3.34(3.04)	5.63(8.61)	3284(3546)	4844(3785)	3803(2558)	3117(1714)
	0.10(0.60)	0.11(2.66)	3.19(2.73)	3.14(6.94)	3755(3730)	5654(3732)	4388(2858)	3442(1786)
	0.19(1.93)	0.52(3.21)	3.96(3.54)	6.86(10.51)	3937(3689)	5689(3677)	4375(2529)	3502(1674)
	0.19(1.18)	0.53(3.20)	4.74(3.75)	8.41(18.54)	3953(4046)	5500(3854)	4085(2850)	3427(1808)
	1.85(5.90)	3.12(13.51)	3.88(7.35)	7.09(19.53)	6023(3992)	6285(3984)	5332(2791)	4185(2029)
	0.29(1.29)	0.37(3.49)	3.95(3.14)	6.89(10.30)	4308(3796)	5439(3672)	3451(2495)	3923(1774)
	0.24(1.25)	0.96(6.37)	1.85(1.99)	4.97(17.12)	5392(3895)	5957(3919)	3867(2628)	3683(1927)
	0.26(1.18)	0.66(2.18)	3.06(2.40)	4.18(7.70)	5346(3677)	5814(3774)	4553(3201)	3685(2471)
	0.63(1.63)	1.38(5.47)	2.07(2.05)	3.65(5.02)	5089(3941)	5958(3906)	4526(2831)	3678(2009)
	9.83(65.76)	11.69(47.30)	3.25(4.89)	4.84(9.36)	4527(3837)	5529(3756)	3671(2659)	3475(1837)
	0.71(3.22)	2.03(9.59)	2.11(2.35)	3.53(7.30)	5501(4216)	5579(4333)	4846(2903)	4257(2109)
H ₂	1.22(4.52)	2.38(11.80)	2.08(2.41)	3.59(6.08)	5564(3920)	6227(3887)	4735(2684)	3659(1909)
	1.11(3.81)	2.72(9.74)	1.94(2.40)	3.14(6.85)	6176(3926)	5945(4206)	5587(2536)	4130(1958)
	1.01(4.75)	3.44(13.05)	2.52(2.71)	5.75(10.97)	5578(3790)	5732(3812)	4265(2403)	3610(1889)

Table 5.3.2: **Descriptive statistics of main indicators** Note that all presented indicators are averages, the standard deviations are included in brackets. The units of the science and internal dependence are in reference/patent. All of the considered indicators are positive values only and highly skewed to zero. As explained earlier in Section 5.3.2, the reference distance is determined for a sub-selection of the patents.

highly unevenly (a small number of areas producing the majority of patents), we chose a coloring following a logarithmic scale. We observe some variation between the figures, Germany and France innovating strongly in photovoltaics, while Denmark focusing more on wind turbines. The main observation, however, at least on a global scale, is that the geographical distributions of innovative activities are fairly similar, even for rather different technologies such as photovoltaics and wind turbines. In fact in a ranking of countries by the total number of US patents, see appendix 5.A, the US, Japan and Germany are consistently in the top 5 for each RET considered in this research (and France for all but 3 RETs). For EP patents, these countries likewise dominate each top 5. Together these four countries account for 76 and 58 respective percentages of the US and EP patents (for RETs). This is in line with the findings of earlier literature considering energy technology in general (Bointner, 2014). The uneven distribution is not due to our choice for counting at the country level. When instead consider spatial the level below countries (corresponding to the 'name_1' level in the Geocoding database), we again see the same regions or locations recurring: California, New York, Tokyo, Bayern and Baden-Württemberg rather consistently dominate in the top 10 locations with most patents for each considered RET. An important part of the knowledge base development of RETs therefore appears to take place in a small number of dominant areas. Together with the similarity of the worldwide geographical distributions, these are relevant descriptive statistics: it indicates that despite the obvious location-boundness of *the application* of specific RETs (hydro energy near rivers, photovoltaics in sunny locations, wind turbines near windy locations, etc.), the development of the knowledge base of these RETs still largely occurs in dominant areas which work on the development of all RETs at the same time.

5.4 Results

We will start this section by exploring the general relations between the analyticity, cumulativeness and the knowledge mobility, where we consider a general data set of patents. We then focus the analysis on the considered RETs, thereby discussing the various relations between indicators both qualitatively and quantitatively. Following the first expectation in Section 5.2, we expect to observe a positive relation between analyticity and knowledge mobility. Following the second expectation in Section 5.2, we expect to see a negative relation between the cumulativeness and the knowledge mobility.

5.4.1 General relations between knowledge dimensions

We will first explore some general relations between on the one hand the knowledge mobility and on the other hand analyticity or cumulativeness of knowledge. We explore these relations using the indicators that are defined on the level of individual patents: the science dependence fraction (sdf), the internal dependence fraction (idf) and the reference distance (rd). In the following analysis we include all EP and US patents for which a reference distance could be determined. One exception is the analysis of the US sdf: there we include, due to calculation challenges, a random selection of 20 percent of all such patents. As explained in Section 5.3.2, we sub-select those EP patents for which the inventors are from Europe and those US

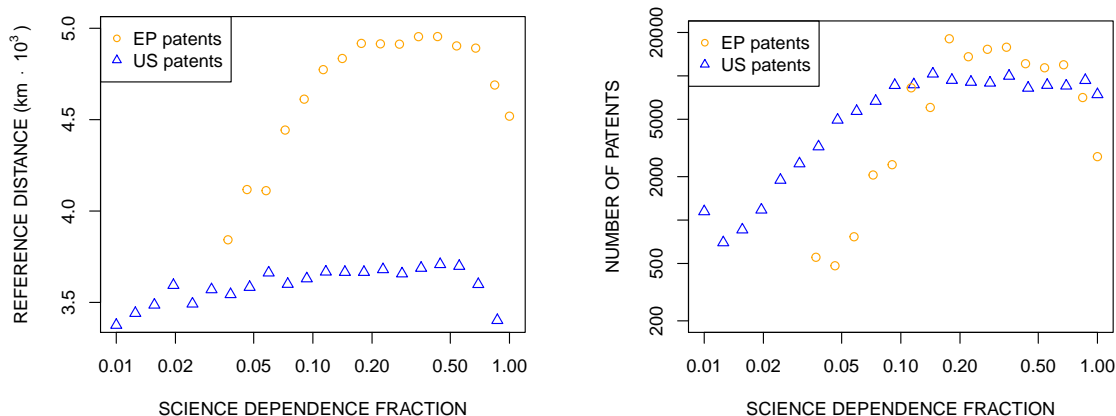


Figure 5.4.1: *Science dependence fraction, reference distance and number of patents* We divide the sdf into exponential bins (base 1.25) and determine for each bin the average rd (left panel) and the normalized cumulative number of patents (right panel). Note the sdf axes are logarithmic, hence the exponential bin sizes are constant in this plot. For sdf bins lower than 0.5 we observe a positive relation between the sdf and rd.

patents for which the inventor is from the US. To calculate the idf, the internal references are determined using within CPC-group references.

In Figure 5.4.1 we divide the sdf in exponential bins and plot the average rd (left panel) and number patents (right panel) for each bin. We observe in the left panel that the rd increases with increasing sdf for both EP and US patents (though more strongly for EP patents). Not included in these figures are the many patents for which the sdf is zero ($7.0 \cdot 10^5$ EP patents and $2.8 \cdot 10^5$ US patents, respectively 6.4 and 2.0 times the total number of EP and US patents in Figure 5.4.1). The average rd of these zero sdf patents are 4114 km for EP patents and 3380 km for US patents, which are similar values to those in the lowest sdf bins in Figure 5.4.1 and therefore in accordance with the observed relation. Even though the reference distance appears to go down for large sdf for both the EP and US patents, we note from the right panel that there are relatively few patents with an $\text{sdf} > 0.5$ (to be precise respectively 4 and 8 percent of the total EP and US patents). For the majority of the patents it therefore counts, in line with expectation, that the greater the sdf, the greater the rd. In other words, greater analyticity can be associated with greater knowledge mobility.

In Figure 5.4.2 we divide the idf in bins of constant size and plot the average rd (left panel) and cumulative number of patents (right panel) for each bin. We clearly observe that the rd decreases when the idf decreases (illustrated also by the linear fits). This is the case almost over the entire range of the idf. A minor exception are the US idf values < 0.15 . As is clear from the right panel, however, there are relatively little patents in this range. As the right panel in Figure 5.4.2 illustrates, most of the patents have mid-range idf values, although there are relatively many patents with idf equal to one (counted in the bin with the largest idf value). As the left panel illustrates, the average rd of the patents in this bin is however in line with the observed pattern. We therefore conclude, in line with expectation, that the greater the idf, the smaller the rd. In other words, greater technological cumulativeness can be associated with lesser knowledge mobility

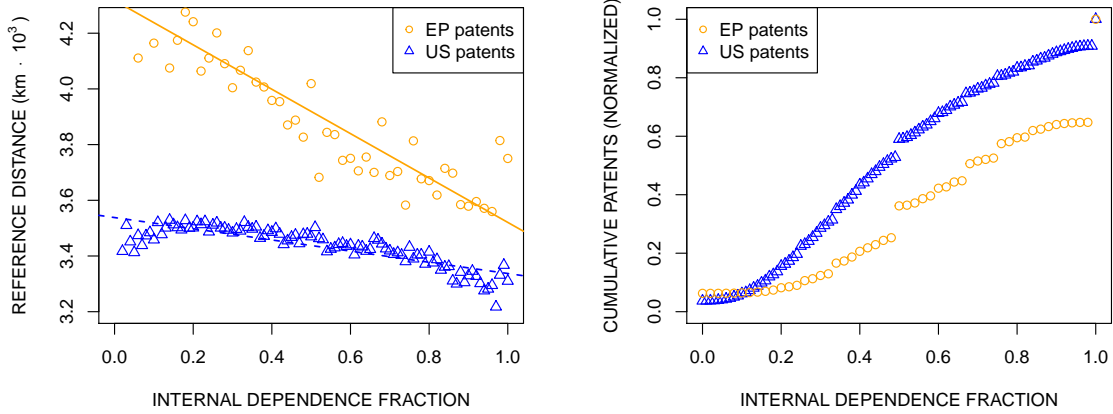


Figure 5.4.2: **Internal dependence fraction, reference distance and number of patents** We divide the *idf* into bins of constant size (0.01 for US and 0.02 for EP patents) and determine for each bin the average *rd* (left panel) and the normalized cumulative number of patents (right panel). For the *rd* linear fits are included, which indicate a negative relation between the *idf* and *rd*. Because the *idf* is a fraction, we observe small breaks in the right panel for highly frequent values such as $1/2$ and $2/3$.

5.4.2 Knowledge relations of RETs

Where in the previous section we discussed the general relations between knowledge dimensions based on a general data set of patents, we will in the following analyze these relations specifically considering the RETs. We will first qualitatively discuss the relation between on the one hand a knowledge mobility indicator and on the other hand either an analyticity or cumulativeness indicator. Subsequently, we will consider these relations more quantitatively, where we determine the correlations and estimate some models.

In Figure 5.4.3 we plot the *rd* for the *sdf* for both the EP patents (left) and the US patents (right). The main observation for both graphs is a positive relationship between both quantities which is well fitted by a linear relation. We refer to Table 5.3.1 for a legend of the icons and the short-names of the technologies. The *sdf* of non-fossil fuels can be observed to be exceptionally large, which is, as discussed earlier, not included in these and later fits. It is therefore also challenging to compare the *rd* of this technology to the rest. It appears the values of the other technologies do allow for comparison however, and in line with expectation, technologies such as wind turbines, geothermal, hydro, solar thermal and energy from sea show relatively low *rd*, whereas photovoltaics, fuel-cells, energy storage and hydrogen technology show relatively large *rd*. Nuclear fission, ccs, clean combustion and electric grid technology are somewhat in between these two groups. Using Table 5.3.1, we note that technologies with a large number of patents (wind turbines, solar thermal, photovoltaics, fuel-cells) occur on both ends of the spectrum. It seems therefore that sheer numbers of patents, often a proxy for the size of the knowledge base, are not sufficient to explain the observed relation. In Figure 5.4.4 we instead plot the inter-patent distance (*ipd*) for the university fraction (*uf*). The positive relation from Figure 5.4.3 remains largely unchanged, the technologies we find upper right (down left) in Figure 5.4.3 also tend to be in the upper right (down left) of Figure 5.4.4. This indicates that the variation across RETs in the considered knowledge dimensions is consistent for the different indicators for these dimensions. A closer

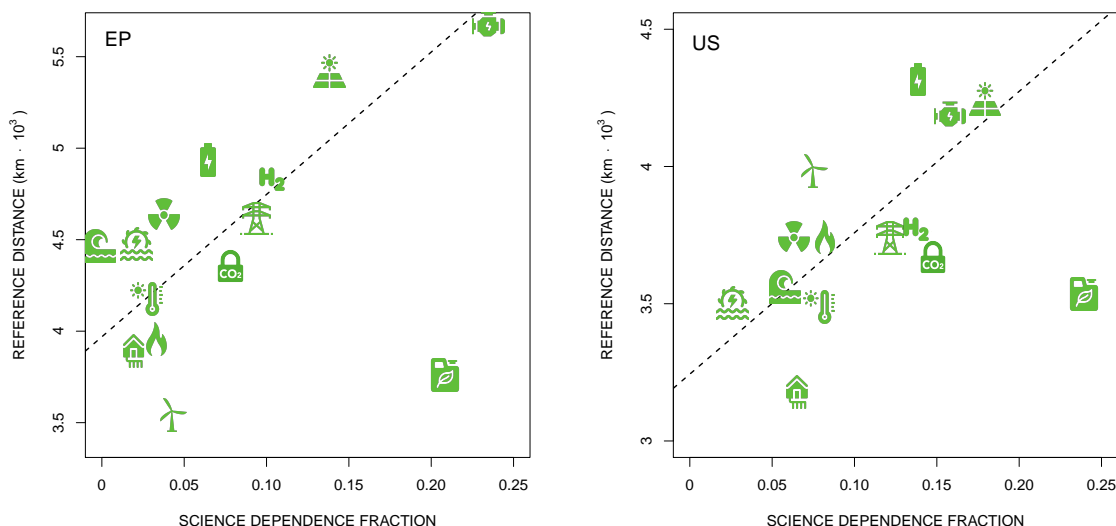


Figure 5.4.3: *Reference distance for science dependence fraction* On the left we display EP patents and on the right US patents. See Table 5.3.1 for a legend of the icons. Non-fossil fuels is excluded in the linear fit.

look reveals some minor variations. The differences between the EP and US patents in Figure 5.4.3 are relatively large in particular for wind turbines and fuel cells. In Figure 5.4.4, these differences are relatively less. This suggests that the ipd indicator may be more uniform between EP and US patents. We discuss the rd variations (and in particular those of wind turbines and fuel cells) in more detail in a part of Appendix 5.B. Especially for the US patents, the uf of nuclear fission and clean combustion is relatively low in Figure 5.4.4 as compared to their sdf in from Figure 5.4.3. This suggests that the knowledge base of these technologies, while retaining a scientific component, is to a lesser extent developed in universities. While there are these minor differences, for most technologies the overall pattern is in agreement with Figure 5.4.3, again confirming that the technologies that build stronger on science also tend to show greater knowledge mobility.

Next, we plot the ipd for the internal dependence (id) in Figure 5.4.5. The main observation is a negative relationship between both quantities which is rather well fitted by a negative logarithmic relation (note the horizontal axis is logarithmic).⁹ This is in line with the expected negative relation between knowledge mobility and technological cumulateness. The only technology defying this pattern, both for EP and US patents but especially US patents, appears to be photovoltaics, which despite a relatively large id, shows great ipd. In an earlier contribution however, we already demonstrated that the internal dependence tends to increase linearly with the number of patents (P. Persoon et al., 2021). Photovoltaics consists of far more patents than the other RETs (especially for US patents), which possibly explains the exceptional value for the internal dependence in this context. Alternatively, we may therefore consider the cumulateness relative to the size of the knowledge base, which we measure by the relative internal dependence (rid) in Figure 5.4.6. In that figure the value of photovoltaics indeed shifts both for the EP and US patents to the left, in better agreement with its large value for the ipd. We observe a similar

⁹We may also take the log of the ipd and instead fit a power relation, the results will be largely comparable.

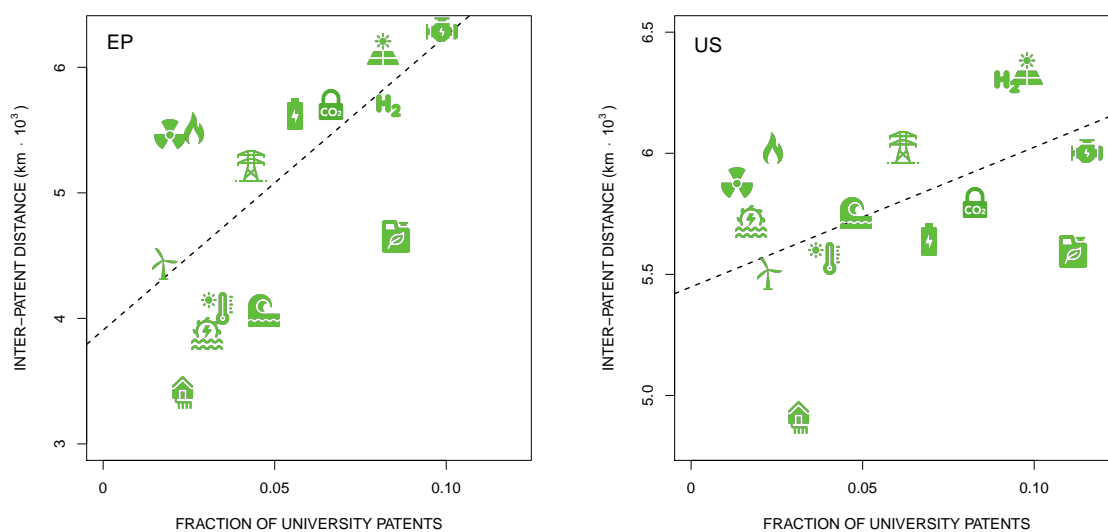


Figure 5.4.4: *Inter-patent distance for the fraction of university patents* On the left we display EP patents, on the right US patents. See Table 5.3.1 for a legend of the icons.

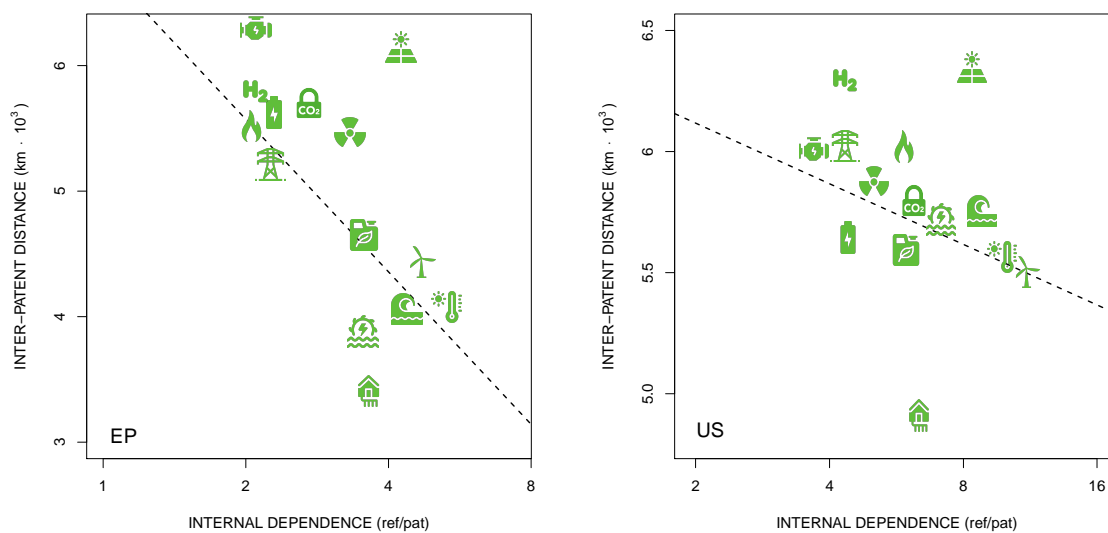


Figure 5.4.5: *Inter-patent distance for the internal dependence* On the left we display EP patents, on the right US patents. Note the horizontal axis is logarithmic. See Table 5.3.1 for a legend of the icons.

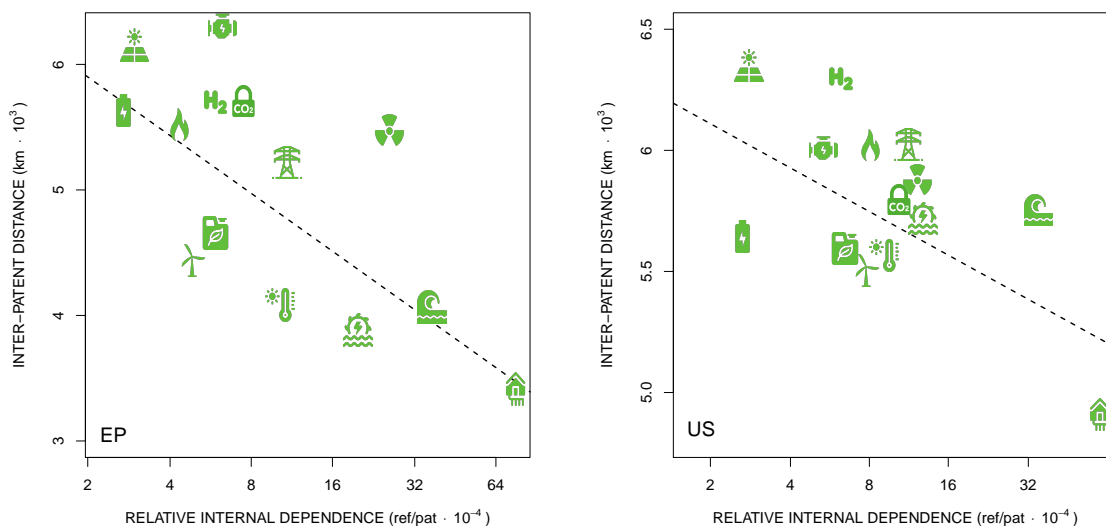


Figure 5.4.6: *Inter-patent distance for the relative internal dependence* On the left we display EP patents, on the right US patents. Note the horizontal axis is logarithmic. See Table 5.3.1 for a legend of the icons.

shift for wind turbines, though to a lesser extent, which is in line with expectation given its shorter ipd. Other than these changes the pattern is largely similar to the one in Figure 5.4.5.

Finally, note in both Figure 5.4.5 and 5.4.6 that the values for non-fossil fuels are more or less in line with the rest of the technologies, where earlier for the science dependence its values were rather exceptional. This therefore presents an extra reason for considering both the science dependence and internal dependence: an exceptional value for the former need not automatically imply an exceptional value for the latter. We will not plot all possible combinations between the indicators we consider, yet for completeness, we include in Figure 5.4.7 the Pearson correlation coefficients of each combination and whether or not this combination is statistically significant. We conclude from Figure 5.4.7 that all correlation coefficients, most of which are substantial, have the expected sign: all analyticity indicators have positive signs with knowledge mobility indicators and all cumulativeness indicators have a negative sign with all knowledge mobility indicators. Especially the analyticity indicators show strong correlations with the knowledge mobility indicators. As expected the correlations between indicators of the same knowledge dimension are generally strong. One exception is the relative internal dependence (rid), which despite strong correlations with knowledge mobility indicators does not correlate strongly with other cumulativeness indicators. Interestingly, the rid does not (anti)correlate strongly with analyticity indicators either, which suggests its relation to the knowledge mobility is to some extent independent of the other indicators. We also observe this for internal dependence (id) of the European patents. To follow up on this suggestion, we will finally consider the possibility to model the knowledge mobility as a linear combination of an analyticity indicator and a cumulativeness indicator. Again we will not present here all such possible linear combinations here in detail, there are simply too many, but instead share our general conclusions and include two examples (Table 5.4.1). As we only consider 13 technologies, i.e. 13 data points, it does not make much sense go much further than combinations of 2 variables. We will

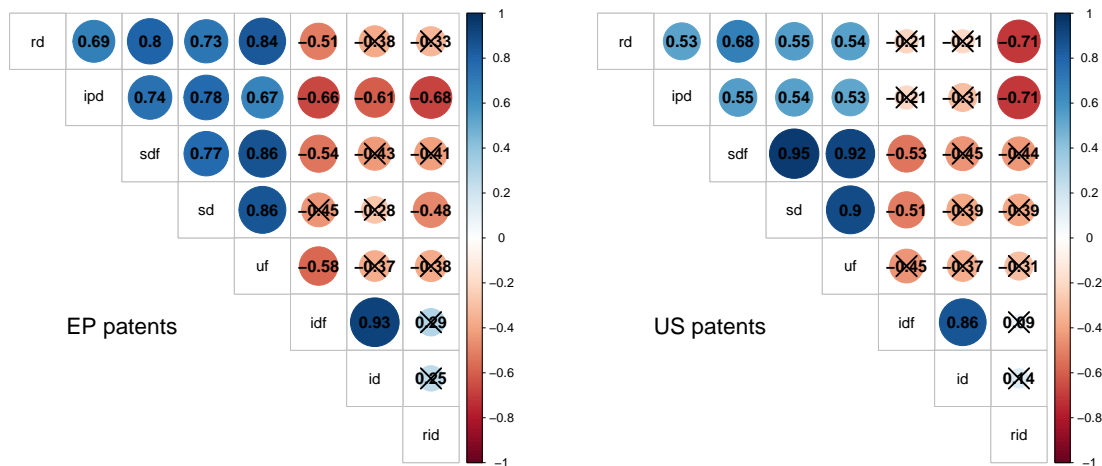


Figure 5.4.7: *Mutual correlations between indicators* On the left we display the correlations for EP patents, on the right for US patents. rd=reference distance, ipd=inter-patent distance, sdf=science dependence fraction, sd=science dependence, uf=university fraction, idf= internal dependence, id=internal dependence, rid=relative internal dependence. Each circle represents a mutual relation, the size and color of which represent the Pearson correlation coefficient. When a cross is included the relation is not significant on a 0.1 level. Non-fossil fuels are excluded while determining these correlations.

first discuss this for the EP patents and then for the US patents.

When we model for the EP patents the internal patent dependence (ipd) as a linear combination of any given analyticity and any given cumulateness indicator, the performance of the model in terms of minimizing the residual standard error and maximizing the Pearson correlation squared (R^2) is much better than for the case where all these indicators are individually considered as predictors. One specific example is given in Table 5.4.1 (left panel), where the EP ipd is modeled as a linear combination of the sdf and rid. The residual standard error (548.6) is much lower than that in a model with only the sdf (662) or the rid (724). The found $R^2 = 0.72$, corresponding to a Pearson coefficient of $R = 0.85$, is also greater than the R values in Figure 5.4.7 for ipd-sdf and ipd-rid. When we take linear combinations of only analyticity or only cumulateness indicators to model the ipd, this only results in a better model in half of the cases. This therefore indicates it makes sense to consider the analyticity and cumulateness as independent factors relating to the knowledge mobility. As Figure 5.4.7 already indicates, the EP patent reference distance very strongly correlates with most of the analyticity indicators, which is difficult to improve considering extra indicators. For the EP rd, we therefore only find very few combinations which present better models than the indicators considered individually.

When we model the knowledge mobility indicators for the US patents as a linear combination of indicators we reach similar conclusions. We find for both knowledge mobility indicators that any combination between the rid and any analyticity indicator result in a better model than when the indicators are considered individually (again judged on the basis of the residual standard error and R^2). We present one example in Table 5.4.1 (right panel), where we model the US rd as a linear combination of the US sdf and US rid. The found residual standard error (207.4) is much

	<i>Dependent variable:</i>		<i>Dependent variable:</i>
	ipd EP patents		rd US patents
sdf EP patents	8,548** (2,826)	sdf US patents	3,490** (1,521)
rid EP patents	-225,936** (91,980)	rid US patents	-112,047** (45,023)
Constant	4,726*** (302)	Constant	3,535*** (180)
Observations	13	Observations	13
R ²	0.72	R ²	0.67
Adjusted R ²	0.66	Adjusted R ²	0.61
Residual Std. Error	548.6 (df = 10)	Residual Std. Error	207.4 (df = 10)
F Statistic	12.8*** (df = 2; 10)	F Statistic	10.2*** (df = 2; 10)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01	<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Table 5.4.1: **Examples of regression outcomes for linear models** We present the results of two regressions, one for EP patents (left panel) and one for US patents (right panel). On the left, we choose the ipd as the dependent variable and the sdf and rid as independent variables, where we also allow a constant term. On the right, we choose the rd as the dependent variable instead. Note there are 13 data points as we exclude non-fossil fuels.

smaller than that in a model with only the sdf (251) or only the rid (244). Also the found $R^2 = 0.67$, corresponding to $R = 0.82$, is greater than the R values in Figure 5.4.7 for rd-sdf and rd-rid. We note that this is largely due to the success of the rid. It is not directly clear why this indicator, as compared to the other cumulativeness indicators, performs much better for the US patents. At least it underlines the need to consider multiple indicators to describe these knowledge dimensions. Only one combination (sd & sdf) of either considering only analyticity indicators or only cumulativeness indicators can be evaluated as a better model than considering the indicators individually. This again indicates that especially the combination of a cumulativeness indicator and an analyticity indicator results in a better model, thus confirming the earlier assertion that the science and internal dependence are complementary indicators, both relating to the knowledge mobility.

In sum, while there is considerable variation across different RETs in terms of knowledge mobility, this variation is to some extent explained by their variation in analyticity and cumulativeness, thus in line with the expectations of Section 5.2. We can distinguish rather consistently a collection of footloose RETs (photovoltaics, energy storage, and fuel cells) characterized by relatively high analyticity and low cumulativeness, from a collection of sticky RETs (energy from sea, wind-turbines, geothermal, hydro, and solar thermal energy) characterized by relatively low analyticity and high cumulativeness.

5.5 Discussion

In this research, we established a close relationship between the analyticity, cumulativeness and the knowledge mobility of technology in general and in particular for various RETs. In this section, we discuss several deeper theoretical aspects and limitations of our approach.

First, our results suggest that analyticity and cumulateness are two distinct characteristics of a knowledge base, in the same way that technological cumulateness and building on scientific knowledge are two distinct mechanisms for technological change. While both the science dependence and internal dependence strongly relate to the knowledge mobility, we find that they are largely independent indicators, i.e., that a high value for the one need not imply a low value for the other. A comprehensive approach to the mechanisms underlying knowledge mobility therefore should not be limited to analyticity or technological cumulateness but should instead treat these as complementary.

Second, we emphasize our focus is on *technological* cumulateness in this research, i.e. studying the relevance of technological knowledge to later technological knowledge. This is not to mean scientific knowledge is not cumulative: 'scientific cumulateness' should however be studied in the context of science building on science. Neither does it imply technology does not influence science: where technology provides science with the necessary instruments, science provides technology with the necessary analytical knowledge.

We can also identify a number of limitations to our research. Our focus in this contribution is on the knowledge aspects of technology, though we acknowledge that there may be many more factors determining the geographical development of a technology, perhaps most importantly the (prospective) market valuation of that technology. For a more inclusive perspective we refer to (Binz & Truffer, 2017). Furthermore, earlier contributions have argued that, while geographical distance remains an important or possibly the most important metric to measure knowledge mobility (Caragliu & Nijkamp, 2016), a more comprehensive approach additionally includes a number of other metrics, based on, for example, organizational, institutional or cognitive proximity (Boschma, 2005). Recognizing this criticism, we performed additional analyses with alternative distance measures, such as the fraction of references staying with a region or country and the Herfindahl index of the distribution of patents over regions and countries. In both cases, however, the results were challenging to interpret, especially since we only considered 13 technologies. Where the fraction of references within region or country suggested contrary results for regions and countries, the Herfindahl index showed contrary results for EP and US patents (and showed some scaling with the number of patents, which further complicated matters). To keep this contribution simple, we excluded a detailed discussion of these results.

5.6 Conclusions and policy implications

This paper contributes to the literature on local and global innovation systems through a systematic empirical analysis of the extent to which Renewable Energy Technologies (RETs) can be characterized as sticky or footloose (Binz & Truffer, 2017). It illustrates the relationship between the spatial innovation dynamics of technologies and characteristics of the knowledge base of these technologies, such as the extent to which this knowledge base is analytic (the 'analyticity') and the extent to which it is cumulative (the 'cumulateness'). The tendency of technology to be spatially sticky or footloose can be systematically approached using concept of knowledge mobility, that is, the extent to which knowledge travels geographically. After empirically confirming, for general technology, the positive relation

between analyticity and knowledge mobility and the negative relation between cumulateness and knowledge mobility, we investigate these relations in more detail for various RETs. We find, in line with theoretical expectations, that the RETs with high analyticity, low cumulateness knowledge bases (photovoltaics, fuel-cells, energy storage and hydrogen technology) show a greater knowledge mobility than those with low analyticity, high cumulateness knowledge bases (wind turbines, geothermal, solar thermal, hydro energy and energy from sea). We will refer to the former group with 'analytic RETs' and the latter group with 'cumulative RETs'. Comparing non-fossil fuels to the other RETs is challenging, as its dependence on analytic knowledge appears to be exceptionally strong.

Our findings lead to a number of recommendations for decarbonizing strategies and policies. For the transition from general R&D stimulating and technology-neutral subsidy schemes to more mission-oriented science and technology policies, a deep understanding of the knowledge characteristics of the considered technology is key. As RET in general depends strongly on analytic knowledge, stimulating scientific research appears to be an effective and targeted measure to stimulate RET development. However, in this work, we have demonstrated that there is also substantial variation across different RETs in various knowledge dimensions, and that this variation across RETs can be used to more effectively target the development of these RETs. More precisely, we have demonstrated that we can distinguish between analytic and cumulative RETs. Where the development of the former allows for easier entry and more flexibility in choosing locations, the development of the latter may be relatively harder to enter and is limited to locations providing the necessary synthetic knowledge. To encourage the development of analytic RETs in particular, policymakers may focus more on strengthening scientific activity. To encourage the development of cumulative RETs in particular, policy mixes focusing on system building are needed to stimulate the local presence of synthetic knowledge. In sum, our results call for policies that are more RET specific, taking into account the variation across RETs in various knowledge dimensions, which relate predictably to spatial dynamics of innovation.

5.7 Acknowledgements

This work was supported by NWO (Dutch Research Council) grant nr. 452-13-010. The icons in Tables 5.3.1 and 5.3.2 and Figures 2-7 are made by Freepik, Kiranshastry, Pixel perfect, and Pixelmeetup from www.flaticon.com.

Appendix

5.A Country ranking by number of patents














													
1	DE 100	DE 236	US 131	DE 1039	US 2129	DE 1965	US 1051	US 370	DE 386	US 998	JP 1221	US 729	US 623
2	US 45	FR 146	GB 98	US 525	JP 1857	US 1058	DE 886	FR 216	US 352	DE 794	DE 1009	DE 487	JP 497
3	CH 26	US 132	DE 82	FR 338	DE 1833	DK 1003	JP 485	DE 145	JP 243	FR 373	US 953	JP 383	DE 312
4	SE 19	GB 11	FR 69	IT 255	KR 848	JP 590	FR 315	JP 115	FR 111	JP 164	KR 508	FR 347	FR 139
5	FR 19	JP 72	NO 55	ES 198	FR 692	ES 451	CH 182	SE 62	SE 98	NL 163	FR 481	IT 101	KR 133

Table 5.A.1: *Country rankings of EP RET patents* We denote the countries by their alpha-2 letter codes and include the number of EP patents.














													
1	US 450	US 925	US 700	US 3921	US 8154	US 3759	US 2801	US 984	US 1064	US 3037	US 3450	US 2061	US 1890
2	JP 38	JP 159	GB 91	DE 488	JP 4155	DE 1501	JP 740	JP 295	JP 390	DE 395	JP 2894	JP 892	JP 1291
3	DE 34	DE 135	FR 63	JP 367	KR 2318	DK 858	DE 548	FR 196	DE 302	JP 326	KR 1233	DE 492	KR 591
4	CA 33	CA 97	JP 63	ES 167	DE 1679	JP 637	FR 243	DE 155	KR 111	FR 282	DE 788	FR 299	DE 312
5	IL 21	FR 95	DE 47	FR 166	FR 627	ES 392	CH 153	SE 71	SE 93	CA 200	FR 378	KR 200	FR 149

Table 5.A.2: *Country rankings of US RET patents* We denote the countries by their alpha-2 letter codes and include the number of US patents.

5.B Reference distances of inventors in and outside Europe and US

In this appendix, we discuss the effect of considering the reference distance of patents where the inventor is located (far) away from the jurisdiction of the patent office,

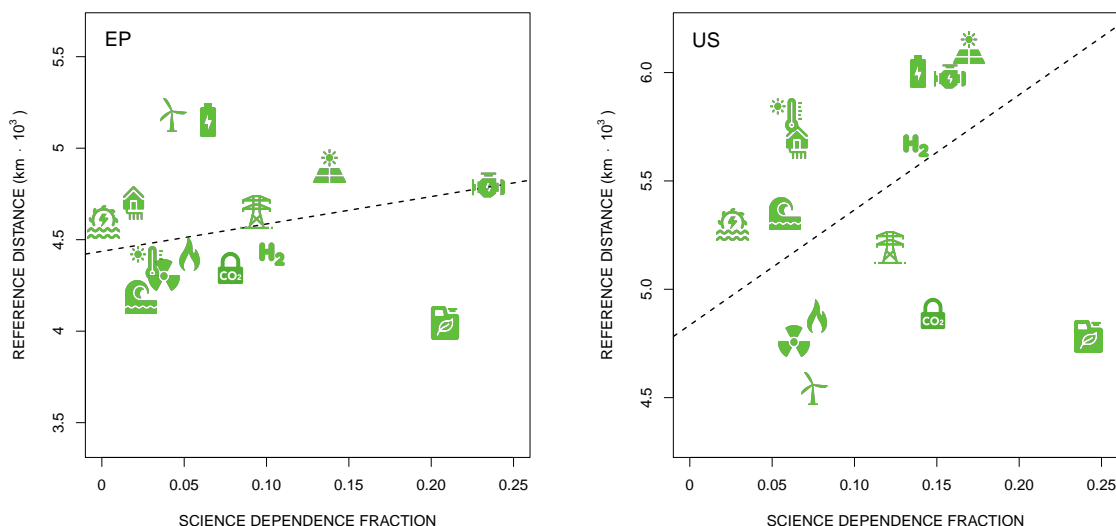


Figure 5.B.1: *Reference distance for science dependence fraction* On the left we display EP patents of which the inventor is from the US and on the right we display US patents of which the inventor is from EP. See Table 5.3.1 for a legend of the icons.

for example EP patents with a US inventor (Figure 5.B.1 left panel) or US patents with a EP inventor (Figure 5.B.1 right panel).

In Figure 5.B.1, the relation between the science dependence fraction and the reference distance again appears to be positive, be it more irregular than the earlier observed relations. However, there are also a number of differences with Figure 5.4.3. The most striking difference is that the reference distance of the US patents in 5.B.1 are much greater. The reference distances can therefore be concluded to partly depend on the location of the inventor. If we would have included the US patents from Figure 5.B.1 in Figure 5.4.3, this would especially affect the reference distance (which is an average over all patents) of technologies with lower numbers of patents. Another striking difference is that the positive relation between science dependence fraction and reference distance is for the EP patents a lot steeper in Figure 5.4.3 than in Figure 5.B.1. This might be a result of the following. There are generally more US patents and inventors than EP patents and as a consequence, the US patents are cited relatively often. The reference distances of an EP inventor referring to a US inventor are much larger than that of US inventor referring to a US inventor. There may therefore be less variation in the reference distance of US inventors, because even when they apply for a EP patent (i.e. Figure 5.B.1 left panel), they may still be citing US inventors relatively often. Finally, we note some typical differences on the level of individual technologies. Where wind turbines has a relatively high reference distance in the left panel of Figure 5.B.1 (as compared to both the right panel and Figure 5.4.3). This indicates that the EP wind turbine inventors refer to patents from inventors close to home (most likely within Europe) whereas the US wind turbine inventors tend to refer to patents from inventors far from home (most likely Europe). This may illustrate a European lead in the wind turbine innovative activities. We see the reverse relation with fuel cells, of which the reference distance in the left panel of Figure 5.B.1 is relatively low as compared to Figure 5.4.3. In the right panel, its the reference distance is actually relatively large, indicating a US innovative lead for this technology.

Chapter 6

Conclusions and research implications

Humanity has in many ways benefited from technological development and the functioning of society has come to strongly depend on various technological solutions. These technological solutions have however introduced problems of their own, of which the problem of global warming caused by the continuous emission of greenhouse gases is arguably most imminent and impacting. Ironically, technology is at the same time expected to partially solve this issue. This applies in particular to the replacement of Fossil Fuel based Energy Technologies (FFETs) by Renewable Energy Technologies (RETs). Most policies attempt to encourage the development of RETs through demand stimulation measures. While these policies work well to create a level-playing field for RETs that are (almost) market-ready, their influence on the development of novel and early-stage technologies is less clear. Coming up with policies that instead act on the level of knowledge development and specifically target RET development, hence including novel and early-stage RETs, is however not trivial, and in the first place requires a deep understanding of these technologies and the scientific and technological knowledge they build on, i.e. their 'knowledge base'. By studying various dimensions of the knowledge base of RETs, this research aims to generate insights that may form a useful evidence base for developing or modifying these policies.

In this research, I focused on three relevant knowledge base dimensions. The first dimension I focused on is the extent to which a technology depends on scientific knowledge, also referred to as the 'science dependence'. Technological development is intricately related to scientific development, yet the strength of the interaction varies across technologies. A concept closely related to the science dependence is the 'science base', which is a more detailed overview of the specific scientific disciplines on which a technology builds. Technologies however also build on earlier technological knowledge, their development is 'cumulative'. Therefore, the second dimension I focused on is the extent to which a technology depends on its earlier development, also referred to as 'technological cumulativeness'. Finally, the third dimension I focused on is the extent to which the knowledge a technology builds on travels geographically, also referred to as the 'knowledge mobility'. Other than the first two dimensions, which can be considered as knowledge intrinsic properties, the knowledge mobility can to a significant degree be considered a property of the users and producers of that knowledge. Therefore, I expected the knowledge mobility

to depend on the other two dimensions. In this research, I developed methodologies to study the science base and technological cumulativeness, and systematically compared these to the knowledge mobility of various RETs.

6.1 Main research conclusions

In the first part of this research, I focused on the question: what is the science base of RETs and how does it differ from that of FFETs? To answer this question, I developed a methodology to study the science base of a technology, which allowed me to identify characteristic differences between both types of energy technologies. I found that RETs generally have a more substantial science base and draw on a more diverse set of scientific disciplines. On average, the science on which RETs build is more recent, less applied, and is published in journals with a higher WOS Journal Impact Factor. However, while the previous findings hold for the average of all RETs, for different RETs (e.g., photovoltaics, wind turbines, and non-fossil fuels), I observed much more variation across these dimensions than for different FFETs (e.g., combustion and gas turbines). Furthermore, the broad spectrum of sciences on which RETs build largely includes the smaller spectrum on which FFETs build.

In the second part of this research, I focused on the question: how can technological cumulativeness be identified and measured? To answer this question I performed an in-depth theoretical and empirical analysis of cumulativeness in the context of technological knowledge. I interpreted a development to be cumulative when a later invention builds on an earlier invention. When this development takes place within a given technology, the technology is understood to develop cumulatively. The cumulativeness of a technology is therefore characterized by the structure of its knowledge base (how knowledge flow connects inventions), which is different from, but closely related to, the size of its knowledge base (the number of inventions). Based on various perspectives appearing in the literature, I further distinguished between cumulativeness in two dimensions. The transversal dimension measures to what extent a technology builds on earlier knowledge at a given step of the development. The longitudinal dimension measures to what extent the intermediate steps of development form sequences of developments, i.e. to what extent there is a continuous chain of developments. To develop a better understanding of how cumulativeness develops along these two dimensions I approached technology as a body of knowledge consisting of interlinked inventions, i.e. a network approach. Following a highly simplified model of the search process that inventing engineers undertake, I derived a set of elementary rules describing how inventions are connected by knowledge flow. Using these rules, I analytically derived that the cumulativeness along both dimensions approximately increases proportionally with the size of the knowledge base, at a rate that may vary across technologies.

Empirical tests of the above ideas, using patent data, confirmed this proportional relation and indicated that the rate varies considerably across technologies. At the same time, I found that across technologies, this rate is inversely related to the rate of invention over time. This suggests that the cumulativeness increases relatively slow in rapidly growing technologies. I also found the empirical distributions for both the transversal as well as the longitudinal dimension of cumulativeness are in good agreement with the predicted distributions for these dimensions, and that likewise,

the mutual relation between both dimensions agrees with analytical expectations. This is not a trivial result, as both dimensions are theoretically very different. There are some consequences to this finding concerning the practicality of the indicators, as the cumulateness along the transversal dimension is typically a lot easier and faster to calculate. The transversal dimension may hence provide a way to obtain a relatively quick estimate of a technology's cumulateness, though I emphasize again this conclusion is based on observing a limited number of technologies, and that more research is required before general validity can be safely assumed. Finally, I also found that my cumulateness measurements of a set of various technologies are largely consistent with other cumulateness measurements of the same technologies appearing elsewhere in the literature, which were determined using a different methodology.

In the third part of this research, I focused on the question: how can the metrics based on network paths be generalized to study cumulative knowledge structures? The commonly used metrics, such as network distance and diameter, are often based on the shortest paths in a network, yet for the study of cumulative knowledge structures, it makes sense to consider the longest, or all unique paths instead. To answer this research question, I studied the theoretical path length distributions of the longest and all unique paths, which could subsequently be used to calculate the metrics. Using the Price model as a starting point, in which the in-degree is constant as usual, I derived an exact solution for the path length distribution of all unique paths from a given initial node to each node in the network. This led me to the conclusion that, where a stronger cumulative advantage effect fundamentally slows down path length growth, a greater average in-degree of the network on the contrary accelerates path length growth. Using this distribution I calculated the expected path length, which as a metric can be considered analogous to the average network distance based on the shortest paths. Where, for a non-zero cumulative advantage effect, the average network distance based on the shortest paths is known to increase with the log log number of nodes, I found that the expected path length increases with the log number of nodes, with a pre-factor which is greater for larger in-degree, yet smaller for a stronger cumulative effect. Furthermore, in a generalization of the Price model, where I allowed the in-degree to increase with the number of nodes, the cumulative advantage effect was found to play a crucial role in maintaining logarithmic path length growth. I demonstrated that, without the cumulative advantage effect, the expected path length would already increase linearly with the number of nodes for a linearly increasing in-degree. For a non-zero cumulative advantage effect, linear expected path length growth is only attained when the number of nodes increases very fast, namely exponentially.

As the collection of all unique paths may contain many redundancies, I additionally considered the subset of the longest paths to each node in the network. As this case is more complicated, I only approximated the longest path length distribution in a context where the cumulative advantage effect is ignored. Where the number of all unique paths of a given length grows unbounded, the number of longest paths of a given length converges to a finite limit, which depends exponentially on the given path length. The distributions of all unique paths and the subset of longest paths are therefore rather different, and this distinction should be carefully taken into account in research approaches to cumulative structures. More generally, my research demonstrated that to meaningfully apply metrics based on the longest

and/or all unique paths in studies of cumulative knowledge structures, it is crucial to take into account network properties such as the cumulative advantage effect and in-degree.

In the fourth part of this research, I focused on the question: how do different RETs vary with respect to science dependence and cumulateness, and how does this relate to their knowledge mobility? To answer these questions I determined for an extensive group of RETs both the science dependence and technological cumulateness and systematically compared this to the knowledge mobility of their knowledge base. Theoretical considerations relate the knowledge mobility positively to the analyticity of the knowledge base (which closely relates to the science dependence) and negatively to the cumulateness of a knowledge base. After having confirmed these relations for general technology, I found that, while several RETs (photovoltaics, fuel cells, energy storage) indeed have a highly analytic knowledge base and indeed develop more widespread, there are also important RETs (wind turbines, solar thermal, geothermal, and hydro energy) for which the knowledge base is less analytic and which develop less widespread. Likewise, the technological cumulateness tends to be lower for the former than for the latter group.

More generally, these results demonstrate the use and relevance of studying knowledge bases to better understand how intrinsic properties of technological knowledge characterize technologies and affect the economics and geography of technological development. Where technological development is sometimes considered a 'black box' in economics, these results show that the way people invent and the underlying knowledge structures allow for systematic study, which may lead to useful insights for policies aiming to steer technology development. More precisely, I have demonstrated that there are important variations in the economics and geography of individual RET development, which arise as a result of characteristic variations in the knowledge base properties of these individual technologies. Knowing these variations is therefore of great relevance to policies aiming to specifically strengthen RET development.

6.2 Policy implications

Many countries and organizations have policies in place to increase the share of renewable energy sources in their total energy mix. Most of these policies make use of demand stimulating measures. Although these policies can be expected to create a level-playing field for RETs that are (almost) market-ready, their influence on the development of novel and early-stage RETs is less clear. The insights created in this study may be helpful for the design of policies that instead directly stimulate knowledge development specifically for RETs, thus including the development of novel or early-stage RETs. Note that such policies, while in the first place fostering more robust and long-term development of RETs, would still positively affect the diffusion of RETs.

For such policies, it is important to understand what distinguishes the knowledge development in RETs from that in other technologies. The answer is perhaps less exciting: that depends. It depends in the first place on which technologies the RETs are compared to and in the second place on the particular RET it is focused on. When I compared RETs overall to FFETs overall, I found that the former build relatively stronger on scientific knowledge than the latter. This suggests that

policies promoting scientific research in general (and basic, high-impact science in particular) are expected to lead to a strengthening of RETs. Yet when I considered RETs and FFETs individually, I found far more variation in the science dependence for the different RETs than for the different FFETs. In fact, I found that a number of important RETs, such as wind turbines and geothermal energy, have a science dependence that is on the same level as most FFETs. Therefore, instead of asking what characterizes the science dependence of RETs, it may make more sense to ask what characterizes the science dependence of FFETs (the answer being that is relatively low).

I reached a similar conclusion when I considered the spectra of scientific disciplines RETs and FFETs build on. Where the spectrum of scientific disciplines RETs build on is very broad, that of FFETs is rather thin, consisting only of a small number of disciplines. Again, therefore, it appears to be more difficult to characterize the science base of RETs than that of FFETs. For policymakers, it is however important to take into account that the broader spectrum of scientific disciplines RETs build on encompasses the smaller spectrum FFETs build on (and oftentimes RETs depend even stronger than FFETs on key disciplines for FFETs). Reducing support to specific scientific disciplines is therefore not likely to be a successful policy for promoting RETs as a replacement for FFETs or for accelerating the phasing out of FFETs.

While policies promoting scientific research in general may therefore lead to an overall strengthening of RETs, it should be taken into account that these policies may, intentionally or unintentionally, have a disproportionate effect on different RETs. As it is to some extent possible to identify scientific disciplines particularly relevant for specific RETs, it would make more sense for such policies to specify a particular set of RETs they aim to strengthen and subsequently stimulate research in the scientific disciplines relevant to that set of RETs.

When I instead considered the technological cumulateness of RETs I arrived at a similar conclusion. For this dimension too, it is challenging to characterize the knowledge base of RETs, as again, considerable heterogeneity was found across different RETs. Policies aiming to encourage entry or diversification into specific RETs (which is generally more difficult when their cumulateness is higher) are therefore advised to take into account the cumulateness of these specific RETs and should be aware that there may be alternative RETs for which the cumulateness is rather different. However, in case there is no real choice as to which RET to go with, such policies should take into account the specialized and often location-bound knowledge that the development of highly cumulative technologies requires.

Finally, when I considered the knowledge mobility of RETs, I observed, for the third time, considerable variation across different RETs. Moreover, this variation across RETs agreed rather well with theoretical expectations based on the variation in science dependence and cumulateness. I concluded, therefore, that the RET knowledge dimensions vary as they would for technology in general, which suggests that the variation in knowledge dimensions for different RETs is in fact not extraordinary or unusual. Altogether, therefore, it turns out to be challenging to give a typical characterization of the RET knowledge base, as the considerable heterogeneity across different RETs results in substantial variation across all considered knowledge base dimensions. If one should insist on a characterization, then perhaps it is exactly this heterogeneity that characterizes the knowledge base of RETs.

I therefore repeat the general advice to policymakers to make a well-considered choice for a particular set of RETs and accordingly design a specific science and technology policy to strengthen the development of this particular set of RETs.

To guide policymakers through the heterogeneity of RETs, I end this section with typification based on the variation across knowledge mobility, cumulateness and science dependence. On the one hand, there are type I RETs, characterized by high science dependence, low cumulateness, high knowledge mobility, and high rates of invention of time. Most of these characteristics count for photovoltaics, non-fossil fuels, general technologies for energy storage, fuel cells, and hydrogen technology. On the other hand, there are type II RETs, characterized by low science dependence, high cumulateness, low knowledge mobility, and low rates of invention over time. Most of these characteristics count for wind turbines, geothermal, solar thermal, hydro energy, and energy from sea. I stress that this typification is not exhaustive for all RETs and that there are exceptions (for example high cumulateness does not automatically imply low science dependence). This typification however may provide a starting point for policies aiming to stimulate a subgroup of RETs, as science-enhancing policies may be more effective for type I RETs and are less location bound, but stimulating type II RETs may require long term investments in developing local specialized technological knowledge. Assuming that novel or early-stage RETs can consistently be typified using intrinsic knowledge base properties, the typification may further be useful in quickly setting up targeted policies to accelerate the development of these technologies.

6.3 Implications to further research

In the following, I will first mention how a number of my findings can be useful for future knowledge bases research. Second, I discuss a number of remaining open problems that touch the nature of knowledge development on a deeper level.

This study required in the first place the development of methodologies to determine the science base and the cumulateness of a technology. Under the premise that similar data is available for other technologies, the approach I chose can be applied equally well to those other technologies and in that sense provides a general framework for future studies. It would for example be interesting to study the knowledge bases of other upcoming, potentially high-impact technologies such as artificial intelligence, robotics or genetic modification. Researchers aiming to use these methodologies are strongly recommended to approach the science base and cumulateness from various angles, using multiple indicators. The discussion about the diversity of the science base in Chapter 2 pointed out that, while some quantitative indicators, such as the Shannon entropy, provided relatively quick insights, a more qualitative analysis involving the various scientific disciplines substantially nuanced the insights provided by these indicators. Likewise, for the cumulateness dimension, I emphasized the need to complement the measurement of the absolute cumulateness of a technology with the cumulateness relative to the size of its knowledge base.

My research also demonstrated the importance of considering both the cumulateness and the science dependence when researching the knowledge base of a technology. The theoretical association between the science dependence and analytic knowledge might lead one to associate cumulateness with synthetic knowledge,

which would suggest that high analyticity implies low cumulativeness. My research however indicated that the science dependence and cumulativeness are largely independent (at least for RETs) thus representing rather distinct properties of the knowledge base. As both dimensions associate (the first positive, the second negative) with knowledge mobility, considering both of them substantially increases the explanatory potential. Especially for studies interested in the geography of a developing technology, it is therefore vital to analyze both the science dependence as well as the cumulativeness.

6.3.1 Open problems

Finally, I discuss three somewhat puzzling findings that would be worth further research. The first of these findings concerns the nature of the inventive process. Starting from a simple search model which is essentially a series of trial and error, I derived a geometric distribution for the number of backward links per invention, which I also observed rather convincingly in patent data. Because of the trial and error nature of the process, it does not matter how many errors the inventor has already undergone, the probability to succeed on the next trial stays the same, (a property sometimes referred to as 'memorylessness') and overall, the case of succeeding at once remains most probable. However, considering that inventions are created in process of (re)combining existing pieces of knowledge (Arthur, 2009; Fleming, 2001; Fleming & Sorenson, 2001), I would expect the number of possible combinations of existing inventions to play a role in these dynamics. If each combination is equally likely to crystallize into a real invention, this would suggest that the inventor is more likely to succeed for some ideal number of combined elements greater than 1.¹ This idea is attractive, as it would for an increasing number of inventions automatically lead to an expected number of combined elements which develops linear with the total number of elements, (something which I observed). However, the above essentially describes a binomially distributed number of backward links, which is something I did not observe. The fact that I obtained stronger evidence for a geometric distribution suggests therefore that the mechanism of combination plays a lesser role than it is often expected, or at least that there is a special type of combination at play, which for example only allows a small subset of the combinations. It remains an open problem how exactly the memoryless geometric distribution can be reconciled with the nature of invention as a basic recombinative process.

A second somewhat puzzling finding in my research concerns the negative relation I found between the cumulativeness rate (i.e. the pace at which the cumulativeness increases for the number of inventions) and the rate of inventions over time in Chapter 3. This seems to contradict the intuition that, when technological developments quickly succeed one another in time, the cumulativeness increases extra fast. The hypothesis in this work was that, when a technology develops rapidly, many different people work on it at the same time, causing them to specialize in different sub-fields of the technology. This fragments the technology and causes its development to be less linear (instead of a single line of development, there are multiple). This linearity may be greater when a small group of inventors continues to build on its own work over a long period of time, thus resulting in a relatively stronger cumulative devel-

¹For example, when there are a total of 6 elements to recombine, there are 15 ways to combine 2 elements, 15 ways to combine 4 elements, yet 20 ways to combine 3 elements.

opment. While this indeed predicts a greater cumulateness rate for a lower rate of invention over time, the empirical analysis I did does not allow me to verify this specific hypothesis. It therefore remains unclear whether this explanation indeed applies or whether there is perhaps some other mechanism at work.

A third somewhat puzzling finding in my research concerns the linear growth I observed both for the transversal as well as the longitudinal dimension of cumulateness in Chapter 3. As was rigorously demonstrated in Chapter 4, these dynamics can only take place in case there is no cumulative advantage effect (note the transversal dimension is then associated with the in-degree and the longitudinal dimension with the expected path length). However, other contributions have indicated that the cumulative advantage effect likely plays a role in networks of technological knowledge (Érdi et al., 2013; Valverde et al., 2007). This brings me to the conclusion that either it differs per technology whether the cumulative advantage effect plays a role, or there are yet other effects at play that accelerate path length growth. This remains an open question for now. One of such mechanisms may be an 'aging effect', in which inventions gradually lose relevance, thus making it more likely that new inventions build on recent inventions. For future path length research, it would therefore be interesting to explore if such effects can additionally be implemented, which may not be a straightforward task.

Author contributions table

	Chapter 2			Chapter 3			Chapter 4	Chapter 5		
	<i>P</i>	<i>B</i>	<i>A</i>	<i>P</i>	<i>B</i>	<i>A</i>	<i>P</i>	<i>P</i>	<i>B</i>	<i>A</i>
Conceptualization	✓	✓	✓	✓			✓	✓	✓	✓
Data curation	✓			✓				✓	✓	
Modeling				✓			✓	✓		
Methodology	✓	✓	✓	✓			✓	✓	✓	✓
Visualization	✓	✓	✓	✓				✓		
Writing - original draft	✓			✓			✓	✓		
Writing - review and editing	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Funding acquisition			✓			✓				✓

Table 6.3.1: **Author contributions** We specify for Chapter 2,3,4 and 5 the contributions of each author, where *P*=Peter Persoon, *B*=Rudi Bekkers and *A*=Floor Alkemade. We note that Floor Alkemade also provided the funding acquisition for Chapter 4.

Nederlandse samenvatting

Eén van de belangrijkste maatregelen om klimaatverandering wereldwijd tegen te gaan is de vervanging van Fossiele Brandstof gebaseerde Energie Technologie (FET) door Hernieuwbare Energie Technologie (HET). Beleid dat zich tot doel stelt de ontwikkeling van HET te bevorderen kan profiteren van een diepgaand begrip van de kennisbasis van HET, dat wil zeggen, de kennis die nodig is om HET te ontwikkelen. Binnen een kennisbasis kunnen meerdere kennisdimensies bestudeerd worden. Een eerste relevante dimensie is de mate waarin die kennis voortbouwt of afhankelijk is van wetenschappelijke kennis, ik noem dit de 'wetenschaps-afhankelijkheid'. Uiteraard bouwen technologieën ook op technologische kennis, de ontwikkeling van technologie is in essentie 'cumulatief'. Een tweede relevante dimensie van de kennisbasis van een technologie is daarom de 'technologische cumulativiteit', de mate waarin het voortbouwt op eerdere technologische kennis ontwikkeld binnen dezelfde technologie, dat wil zeggen, voortbouwt op zichzelf. In dit proefschrift ontwikkel ik methodologieën om de wetenschapsbasis en de cumulativiteit te meten en pas deze toe op de kennisbasis van HET.

Het eerste gedeelte van dit onderzoek is toegewijd aan een gedetailleerde analyse van de wetenschappelijke kennisbasis en wetenschaps-afhankelijkheid van beide FET en HET, waarbij speciale aandacht wordt besteed aan de verschillen tussen deze. De wetenschapsbasis van HET is substantieel groter dan die van FET, en bestaat uit een diversere groep wetenschapsgebieden. De wetenschap waar HET op bouwt is gemiddeld recenter, minder toegepast van aard en gepubliceerd in wetenschappelijke tijdschriften met een hogere impact factor. Wel merk ik op dat verschillende HETen (bijv. fotonvoltaïsche cellen, windturbines en niet-fossiele brandstoffen) sterk variëren in de genoemde wetenschapsdimensies, sterker dan de verschillende FETen (bijv. verbrandingstechnologie, gasturbines). Ik observeer dat de wetenschapsgebieden waar HET op bouwt de wetenschapsgebieden waar FET op bouwt grotendeels omvat.

Het tweede deel van dit onderzoek is toegewijd aan een diepgaande theoretische en empirische analyse van technologische cumulativiteit. Ondanks het feit dat dit concept over het algemeen van groot belang wordt geacht, verschillen de perspectieven op het concept, waardoor niet altijd duidelijk is welke rol cumulativiteit nu werkelijk speelt in kennisontwikkeling, en hoe dat verschilt per technologie. In dit onderzoek karakteriseer ik cumulativiteit door de structuur van de kennisbasis (bestaande uit de kennisverbindingen tussen uitvindingen). Dit is iets anders dan -maar is wel gerelateerd aan- de grootte van die kennisbasis (bestaande uit het aantal uitvindingen). Mijn benadering van de kennisbasis als een netwerk stelt me in staat om indicatoren te definiëren die de cumulativiteit meten. Aan de hand van een simpel model van kennis zoekende uitvinders, kan ik een recht-evenredig verband afleiden tussen de cumulativiteit en de grootte van de kennisbasis, waarbij de

sterkte van dit verband technologie afhankelijk is. Empirische testen van deze aanpak, gebruikmakende van octrooi-data, verifiëren deze recht-evenredigheid en wijzen uit dat de sterkte van het verband substantieel varieert voor verschillende technologieën. Tegelijkertijd vind ik dat deze variatie in sterkte omgekeerd gerelateerd is aan de hoeveelheid uitvindingen die gedaan worden per tijdseenheid voor die technologie. Dit suggereert dat cumulativiteit relatief langzaam groeit in technologieën die zich relatief snel ontwikkelen.

Het derde deel van dit onderzoek gaat dieper in op de vraag hoe netwerk paden en het concept padlengte gebruikt kunnen worden om cumulatieve kennis structuren te bestuderen. Uitgaande van het Price model van netwerkgroei, leid ik een exact oplossing af voor de padlengte distributie van alle unieke paden vanaf een oorspronkelijke node naar alle andere nodes in het netwerk. Daarbij bestudeer ik de relatieve invloeden van de gemiddelde netwerkgraad en het cumulatieve voordeel effect en introduceer ik een generalisering voor een toenemende netwerkgraad. Ik stel daarbij vast dat het cumulatieve voordeel effect de groei van padlengte substantieel afremt. Omdat we verwachten dat de verzameling van alle unieke paden een groot aantal redundanties bevat, beschouw ik daarnaast specifiek de subverzameling van alle langste paden van de oorspronkelijk node naar alle nodes in het netwerk. Aangezien het analytisch beschrijven van deze subverzameling een stuk uitdagender is, volsta ik in dit geval met een benadering van de padlengte distributie, waarbij de netwerkdynamiek tot een minimum is versimpeld. Ik toon aan dat, waar het aantal unieke paden van een gegeven lengte ongelimiteerd groeit, het aantal langste paden van een gegeven lengte een bovenste limiet heeft. Deze limiet hangt exponentieel af van de gegeven padlengte. Fundamentele netwerk eigenschappen en dynamische effecten bepalen daarom mede hoe cumulatieve structuren in kennisnetwerken tot stand komen, en moeten daarom binnen beschouwing genomen worden in studies naar die cumulatieve structuren.

In het vierde deel van dit onderzoek bepaal ik voor een uitgebreide groep HETen beide de wetenschaps-afhankelijkheid en de cumulativiteit. Verder vergelijk ik deze systematisch met de kennismobiliteit van de respectievelijke kennisbases. De kennismobiliteit meet hoe ver technologische kennis geografisch reist. De kennismobiliteit wordt positief geassocieerd met de analytische van de kennisbasis (een begrip nauw gerelateerd aan mate van afhankelijkheid van wetenschap) en negatief geassocieerd met de cumulativiteit van de kennisbasis. Ik identificeer aan de ene kant een belangrijke groep HETen (fotovoltaïsche cellen, brandstofcellen, technologie voor energieopslag) met een behoorlijk analytische kennisbasis en (inderdaad) een substantiële kennismobiliteit, en aan de andere kant een belangrijke groep HETen (windturbines, zonnearmte-, aardwarmte- en waterkracht technologie) met een minder analytische kennisbasis en (inderdaad) een kleinere kennismobiliteit. Verder is, in de lijn der verwachting, de cumulativiteit lager voor de eerste dan voor de tweede groep.

De eerdergenoemde karakteristieken van de HET kennisbasis hebben een aantal consequenties voor beleid dat zich tot doel stelt de ontwikkeling van HET te bevorderen. HET bouwt over het algemeen significant sterker op wetenschappelijke kennis dan FET. Om die reden verwacht ik dat beleid dat wetenschap in het algemeen stimuleert, en in het bijzonder wetenschap met een hoge impact-factor en van een meer fundamentele aard, leidt tot een versterking van HET ontwikkeling ten opzicht van FET ontwikkeling. Tegelijkertijd is er behoorlijke heterogeniteit tussen

verschillende HETen in de afhankelijkheid van wetenschap en cumulativiteit, welke op karakteristieke wijze relateren aan andere dimensies zoals de kennismobiliteit en het aantal uitvindingen gedaan per tijdseenheid. Dit pleit daarom voor regionaal beleid dat specifiek is voor een specifieke HET, waarbij er rekening gehouden wordt niet alleen met het type kennis waar de HET op bouwt, maar ook met de vraag of deze kennis lokaal aanwezig is en hoe moeilijk het is een eventuele achterstand in te lopen.

Acknowledgements

It feels the five years I spent working on this PhD have flown by in a moment, which is a good indication of how much I enjoy working in academia. Surely, as in any type of work, there are ups and downs, but overall I feel that the academic freedom, intellectual challenges and satisfaction of doing (small) discoveries make that this job never starts to bore. I am very grateful to Floor Alkemade for giving me the opportunity to do this work, for her trust and endless patience. Similarly, I am very grateful to Rudi Bekkers, whose detailed advice was always spot on. For both: I know working with me was not always easy, with me mainly following my own ideas. I therefore appreciate your continuous support even more.

My colleagues at the TIS group also deserve a big thanks, it was really fantastic to work in such a friendly and welcoming group. Perhaps sometimes, during seminars and talks, we were even somewhat too friendly and could have been more critical. It was a big honor (and a lot of fun) to serve as your captain during the yearly TUE Sportsdays and also the TIS camping trips are nothing short of legendary. I am grateful to Önder and Carolina for helping me especially in the early stage of my PhD. Finally, I would like to especially thank Aleid, Elena, Mart, Ben and Deyu for all the inspiring talks we had while sharing an office.

Recharging outside office hours is as important to stay motivated, so I am happy I can always count on my buddies at the USRS, the Italy Jings, the Ghasten, Wim & Stef, Hitch, DeVierGeneraties, and the Nerds. Most of you were genuinely interested in following my progress, which I very much appreciated. This counts also for my brothers Leon and Roel, my godparents Coby and Jos, and my parents-in-law Ans and Marius, who I know would have supported me regardless of the outcome.

Concluding, I would like to especially thank my friend and paranymph Georgios, who was already a colleague at theoretical physics and to my happiness joined me at TUE, making my time there even more enjoyable. Further, there is no one who more closely witnessed my research develop, who shared more in the ups and downs, who helped me out more with a thousand coding things and at the same time remained the most enjoyable company, no one other than my friend and other paranymph Aziiz. Thanks a million, buddy. Pap en mam, dankzij de liefde, support en richting die jullie gegeven hebben, ben ik gekomen tot waar ik nu sta en geworden tot wat ik nu ben. Ik heb nog een weg te gaan, maar hoop iets mee te krijgen van hoe jullie in het leven staan. Lieve Cherique, ik heb je begrip en geduld op de proef gesteld, en terwijl ik daarin wellicht faalde, slaagde jij met vlag en wimpel. Jij bent mijn inspiratie om alles uit het leven te halen wat erin zit, want een goed voorbeeld doet volgen.

Curriculum vitae

Peter Persoon grew up in the Westland, an area in Zuid Holland known for its many greenhouses. Although he did not aspire to become a greener himself, he always admired the clever solutions that greeners come up with to improve their production processes. An example is the 'opraapmachine' depicted on the cover, which was developed by the author's father Hennie Persoon. For his bachelor studies, Peter went to University College Utrecht. There he combined, amongst other subjects, physics, mathematics, and philosophy of science. After that he went on to obtain a master's degree in theoretical physics (Utrecht University) and worked as a trainee at TNO, the Dutch organisation for applied research. Even though his time at TNO sparked his interest in invention and innovation, he afterward preferred to return to more fundamental research. He was therefore happy to start a PhD in the direction of Innovation Sciences at the Eindhoven University of Technology, which led to this work. Peter will now work as a postdoc at the Oxford Martin School, where he will again focus his efforts to better understanding technological change.

Next to his work in academia, Peter likes to read, paint and play rugby. When the weather is good, chances are you find him next to his barbecue or on an adventurous hike together with his wife Cherique.

Bibliography

- Abernathy, W. J., & Utterback, J. M. (1978). Patterns of industrial innovation. *Technology review*, 80(7), 40–47.
- Acemoglu, D., Akcigit, U., & Kerr, W. R. (2016). Innovation network [Publisher: National Academy of Sciences Section: Social Sciences]. *Proceedings of the National Academy of Sciences*, 113(41), 11483–11488. <https://doi.org/10.1073/pnas.1613559113>
- Adamchik, V. S. (1997). On Stirling Numbers and Euler Sums [Publisher: Carnegie Mellon University]. <https://doi.org/10.1184/R1/6607910.v1>
- Albert, R., & Barabasi, A.-L. (2002). Statistical mechanics of complex networks [Publisher: American Physical Society]. *Reviews of Modern Physics*, 74(1), 47–97. <https://doi.org/10.1103/RevModPhys.74.47>
- Albert, R., & Barabási, A.-L. (2000). Topology of Evolving Networks: Local Events and Universality. *Physical Review Letters*, 85(24), 5234–5237. <https://doi.org/10.1103/PhysRevLett.85.5234>
- Albert, R., Jeong, H., & Barabasi, A.-L. (2000). Error and attack tolerance of complex networks [Number: 6794 Publisher: Nature Publishing Group]. *Nature*, 406(6794), 378–382. <https://doi.org/10.1038/35019019>
- Alcácer, J., & Gittelman, M. (2006). Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations. *Review of Economics and Statistics*, 88(4), 774–779. <https://doi.org/10.1162/rest.88.4.774>
- Anderson, P., & Tushman, M. L. (1990). Technological Discontinuities and Dominant Designs: A Cyclical Model of Technological Change. *Administrative Science Quarterly*, 35(4), 604–633. <https://doi.org/10.2307/2393511>
- Apa, R., De Noni, I., Orsi, L., & Sedita, S. R. (2018). Knowledge space oddity: How to increase the intensity and relevance of the technological progress of European regions. *Research Policy*, 47(9), 1700–1712. <https://doi.org/10.1016/j.respol.2018.06.002>
- Arthur, W. (2009). *The Nature of Technology* (1st). Free Press.
- Arts, S., & Fleming, L. (2018). Paradise of Novelty—Or Loss of Human Capital? Exploring New Fields and Inventive Output [Publisher: INFORMS]. *Organization Science*, 29(6), 1074–1092. <https://doi.org/10.1287/orsc.2018.1216>
- Asheim, B., Boschma, R., & Cooke, P. (2011). Constructing Regional Advantage: Platform Policies Based on Related Variety and Differentiated Knowledge Bases [Publisher: Routledge]. *Regional Studies*, 45(7), 893–904. <https://doi.org/10.1080/00343404.2010.543126>
- Asheim, B., & Coenen, L. (2005). Knowledge bases and regional innovation systems: Comparing Nordic clusters. *Research Policy*, 34(8), 1173–1190. <https://doi.org/10.1016/j.respol.2005.03.013>

- Asheim, B., Grillitsch, M., & Trippl, M. (2016). Regional innovation systems: Past - present - future. *Handbook on the Geographies of Innovation* (pp. 45–62). <https://doi.org/10.4337/9781784710774>
- Azagra-Caro, J. M., & Tur, E. M. (2018). Examiner trust in applicants to the European Patent Office: Country specificities. *Scientometrics*, *117*(3), 1319–1348. <https://doi.org/10.1007/s11192-018-2894-4>
- Bacchiocchi, E., & Montobbio, F. (2010). International Knowledge Diffusion and Home-bias Effect: Do USPTO and EPO Patent Citations Tell the Same Story? *The Scandinavian Journal of Economics*, *112*(3), 441–470. Retrieved November 5, 2019, from <https://www.jstor.org/stable/40783300>
- Balland, P.-A. (2016). Relatedness and the geography of innovation. *Chapters* (pp. 127–141). Edward Elgar Publishing. Retrieved June 23, 2020, from https://ideas.repec.org/h/elg/eechap/16055_6.html
- Balland, P.-A., & Rigby, D. (2017). The Geography of Complex Knowledge [Publisher: Routledge _eprint: <https://doi.org/10.1080/00130095.2016.1205947>]. *Economic Geography*, *93*(1), 1–23. <https://doi.org/10.1080/00130095.2016.1205947>
- Barabasi, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, *286*(5439), 509–512. <https://doi.org/10.1126/science.286.5439.509>
- Barbieri, N., Marzucchi, A., & Rizzo, U. (2020). Knowledge sources and impacts on subsequent inventions: Do green technologies differ from non-green ones? *Research Policy*, *49*(2), 103901. <https://doi.org/10.1016/j.respol.2019.103901>
- Bar-Ilan, J., & Halevi, G. (2017). Post retraction citations in context: A case study. *Scientometrics*, *113*(1), 547–565. <https://doi.org/10.1007/s11192-017-2242-0>
- Basalla, G. (1989). *The Evolution of Technology*. Cambridge University Press.
- Bentley, P. J., Gulbrandsen, M., & Kyvik, S. (2015). The relationship between basic and applied research in universities. *Higher Education*, *70*(4), 689–709. <https://doi.org/10.1007/s10734-015-9861-2>
- Binz, C., Tang, T., & Huenteler, J. (2017). Spatial lifecycles of cleantech industries – The global development history of solar photovoltaics. *Energy Policy*, *101*, 386–402. <https://doi.org/10.1016/j.enpol.2016.10.034>
- Binz, C., & Truffer, B. (2017). Global Innovation Systems—A conceptual framework for innovation dynamics in transnational contexts. *Research Policy*, *46*(7), 1284–1298. <https://doi.org/10.1016/j.respol.2017.05.012>
- Bointner, R. (2014). Innovation in the energy sector: Lessons learnt from R&D expenditures and patents in selected IEA countries. *Energy Policy*, *73*, 733–747. <https://doi.org/10.1016/j.enpol.2014.06.001>
- Boschma, R. (2005). Proximity and Innovation: A Critical Assessment [Publisher: Routledge _eprint: <https://doi.org/10.1080/0034340052000320887>]. *Regional Studies*, *39*(1), 61–74. <https://doi.org/10.1080/0034340052000320887>
- Boschma, R., Balland, P.-A., & Kogler, D. F. (2015). Relatedness and technological change in cities: The rise and fall of technological knowledge in US metropolitan areas from 1981 to 2010 [Publisher: Oxford Academic]. *Industrial and Corporate Change*, *24*(1), 223–250. <https://doi.org/10.1093/icc/dtu012>
- Breschi, S. (2000). The Geography of Innovation: A Cross-sector Analysis [Publisher: Routledge _eprint: <https://doi.org/10.1080/00343400050015069>]. *Regional Studies*, *34*(3), 213–229. <https://doi.org/10.1080/00343400050015069>

- Breschi, S., Malerba, F., & Orsenigo, L. (2000). Technological Regimes and Schumpeterian Patterns of Innovation [Publisher: [Royal Economic Society, Wiley]]. *The Economic Journal*, 110(463), 388–410. Retrieved April 9, 2020, from <https://www.jstor.org/stable/2566240>
- Butler, S. (2014). *The Notebooks of Samuel Butler*. The Floating Press.
- Callaert, J., Van Looy, B., Verbeek, A., Debackere, K., & Thijs, B. (2006). Traces of Prior Art: An analysis of non-patent references found in patent documents. *Scientometrics*, 69(1), 3–20. <https://doi.org/10.1007/s11192-006-0135-8>
- Calvert, J., & Martin, B. (2001). Changing conceptions of basic research? [Background Document for the Workshop on Policy Relevance and Measurement of Basic Research, Oslo, October 2001]. Retrieved July 19, 2018, from <http://www.oecd.org/sti/sci-tech/2674369.pdf>
- Caragliu, A., & Nijkamp, P. (2016). Space and knowledge spillovers in European regions: The impact of different forms of proximity on spatial knowledge diffusion. *Journal of Economic Geography*, 16(3), 749–774. <https://doi.org/10.1093/jeg/lbv042>
- Caravenna, F., Garavaglia, A., & Hofstad, R. v. d. (2019). Diameter in ultra-small scale-free random graphs. *Random Structures & Algorithms*, 54(3), 444–498. <https://doi.org/https://doi.org/10.1002/rsa.20798>
- Castaldi, C., Frenken, K., & Los, B. (2015). Related Variety, Unrelated Variety and Technological Breakthroughs: An analysis of US State-Level Patenting. *Regional Studies*, 49(5), 767–781. <https://doi.org/10.1080/00343404.2014.940305>
- Catalini, C., Lacetera, N., & Oettl, A. (2015). The incidence and role of negative citations in science [Publisher: National Academy of Sciences Section: Social Sciences]. *Proceedings of the National Academy of Sciences*, 112(45), 13823–13826. <https://doi.org/10.1073/pnas.1502280112>
- Cefis, E. (2003). Is there persistence in innovative activities? *International Journal of Industrial Organization*, 21(4), 489–515. [https://doi.org/10.1016/S0167-7187\(02\)00090-5](https://doi.org/10.1016/S0167-7187(02)00090-5)
- Clancy, M. S. (2018). Combinations of technology in US patents, 1926–2009: A weakening base for future innovation? [Publisher: Routledge]. *Economics of Innovation and New Technology*, 27(8), 770–785. <https://doi.org/10.1080/10438599.2017.1410007>
- ClarivateAnalytics. (n.d.-a). Journal Impact Factor. Retrieved November 30, 2018, from <http://ipscience-help.thomsonreuters.com/inCites2Live/indicatorsGroup/aboutHandbook/usingCitationIndicatorsWisely/jif.html>
- ClarivateAnalytics. (n.d.-b). Web of Science. Retrieved August 29, 2018, from <https://login.webofknowledge.com/>
- COHEN, A. (2016). Comparing Regression Coefficients Across Subsamples: A Study of the Statistical Test [Publisher: SAGE PUBLICATIONS]. *Sociological Methods & Research*. <https://doi.org/10.1177/0049124183012001003>
- Cohen, R., & Havlin, S. (2003). Scale-Free Networks are Ultrasmall [arXiv: cond-mat/0205476]. *Physical Review Letters*, 90(5), 058701. <https://doi.org/10.1103/PhysRevLett.90.058701>
- Cohen, W. M., & Levinthal, D. A. (1990). Absorptive Capacity: A New Perspective on Learning and Innovation. *Administrative Science Quarterly*, 35(1), 128–152. <https://doi.org/10.2307/2393553>

- CPC. (2018). Cooperative Patent Classification - Table. Retrieved July 20, 2018, from <https://www.cooperativepatentclassification.org>
- Criscuolo, P., & Verspagen, B. (2008). Does it matter where patent citations come from? Inventor vs. examiner citations in European patents. *Research Policy*, *37*(10), 1892–1908. <https://doi.org/10.1016/j.respol.2008.07.011>
- Dahlin, K. B., & Behrens, D. M. (2005). When is an invention really radical? *Research Policy*, *34*(5), 717–737. <https://doi.org/10.1016/j.respol.2005.03.009>
- Dean, L. G., Vale, G. L., Laland, K. N., Flynn, E., & Kendal, R. L. (2014). Human cumulative culture: A comparative perspective: Human cumulative culture. *Biological Reviews*, *89*(2), 284–301. <https://doi.org/10.1111/brv.12053>
- Dechezleprêtre, A., Martin, R., & Mohnen, M. (2014). *Knowledge Spillovers from Clean and Dirty Technologies* (CEP Discussion Paper). Centre for Economic Performance, LSE. Retrieved April 2, 2017, from <http://econpapers.repec.org/paper/cepecdps/dp1300.htm>
- de Rassenfosse, G., Griffiths, W. E., Jaffe, A. B., & Webster, E. (2016). *Low-quality Patents in the Eye of the Beholder: Evidence from Multiple Examiners* (Working Paper No. 22244) [Series: Working Paper Series]. National Bureau of Economic Research. <https://doi.org/10.3386/w22244>
- de Rassenfosse, G., Kozak, J., & Seliger, F. (2019). Geocoding of worldwide patent data [Number: 1 Publisher: Nature Publishing Group]. *Scientific Data*, *6*(1), 260. <https://doi.org/10.1038/s41597-019-0264-6>
- Dereich, S., Mönch, C., & Mörters, P. (2012). Typical Distances in Ultrasmall Random Networks [Publisher: Cambridge University Press]. *Advances in Applied Probability*, *44*(2), 583–601. <https://doi.org/10.1239/aap/1339878725>
- Dereich, S., Mönch, C., & Mörters, P. (2017). Distances in scale free networks at criticality [arXiv: 1604.00779]. *arXiv:1604.00779 [math]*. Retrieved June 1, 2021, from <http://arxiv.org/abs/1604.00779>
- Dommers, S., van der Hofstad, R., & Hooghiemstra, G. (2010). Diameters in Preferential Attachment Models. *Journal of Statistical Physics*, *139*(1), 72–107. <https://doi.org/10.1007/s10955-010-9921-z>
- Dosi, G. (1982). Technological paradigms and technological trajectories: A suggested interpretation of the determinants and directions of technical change. *Research Policy*, *11*(3), 147–162. [https://doi.org/10.1016/0048-7333\(82\)90016-6](https://doi.org/10.1016/0048-7333(82)90016-6)
- Duguet, E., & MacGarvie, M. (2005). How well do patent citations measure flows of technology? Evidence from French innovation surveys [Publisher: Routledge _eprint: <https://doi.org/10.1080/1043859042000307347>]. *Economics of Innovation and New Technology*, *14*(5), 375–393. <https://doi.org/10.1080/1043859042000307347>
- Enquist, M., Ghirlanda, S., & Eriksson, K. (2011). Modelling the evolution and diversity of cumulative culture. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *366*(1563), 412–423. <https://doi.org/10.1098/rstb.2010.0132>
- EPO. (n.d.). Espacenet coverage codes & statistics [Library Catalog: www.epo.org]. Retrieved March 9, 2020, from <https://www.epo.org/searching-for-patents/data/coverage/regular.html>
- Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P., & Zalányi, L. (2013). Prediction of emerging technologies based on analysis

- of the US patent citation network. *Scientometrics*, 95(1), 225–242. <https://doi.org/10.1007/s11192-012-0796-4>
- Evans, T. S., Calmon, L., & Vasiliauskaite, V. (2020). The longest path in the Price model [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, 10(1), 10503. <https://doi.org/10.1038/s41598-020-67421-8>
- Fleming, L. (2001). Recombinant Uncertainty in Technological Search. *Management Science*, 47(1), 117–132. <https://doi.org/10.1287/mnsc.47.1.117.10671>
- Fleming, L., & Sorenson, O. (2001). Technology as a complex adaptive system: Evidence from patent data. *Research Policy*, 30(7), 1019–1039. [https://doi.org/10.1016/S0048-7333\(00\)00135-9](https://doi.org/10.1016/S0048-7333(00)00135-9)
- Fleming, L., & Sorenson, O. (2004). Science as a map in technological search. *Strategic Management Journal*, 25(89), 909–928. <https://doi.org/10.1002/smj.384>
- Foray, D. (2014). *Smart Specialisation: Opportunities and Challenges for Regional Innovation Policy*. Taylor & Francis.
- Freeman, C., & Soete, L. (1997). *The Economics of Industrial Innovation*. Pinter.
- Frenken, K., Izquierdo, L. R., & Zeppini, P. (2012). Branching innovation, recombinant innovation, and endogenous technological transitions. *Environmental Innovation and Societal Transitions*, 4, 25–35. <https://doi.org/10.1016/j.eist.2012.06.001>
- Frenz, M., & Prevezer, M. (2012). What Can CIS Data Tell Us about Technological Regimes and Persistence of Innovation? [Publisher: Routledge _eprint: <https://doi.org/10.1080/13662716.2012.694676>]. *Industry and Innovation*, 19(4), 285–306. <https://doi.org/10.1080/13662716.2012.694676>
- Garavaglia, A., Hofstad, R., & Woeginger, G. (2017). The Dynamics of Power laws: Fitness and Aging in Preferential Attachment Trees. *Journal of Statistical Physics*. <https://doi.org/10.1007/s10955-017-1841-8>
- Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics*, 1(4), 359–375. <https://doi.org/10.1007/BF02019306>
- Garrone, P., Piscitello, L., & Wang, Y. (2014). Innovation Performance and International Knowledge Spillovers: Evidence from the Renewable Energy Sector in OECD Countries [Publisher: Routledge]. *Industry and Innovation*, 21(7-8), 574–598. <https://doi.org/10.1080/13662716.2015.1011913>
- Gertler, M. S. (2003). Tacit knowledge and the economic geography of context, or The undefinable tacitness of being (there) [Publisher: Oxford Academic]. *Journal of Economic Geography*, 3(1), 75–99. <https://doi.org/10.1093/jeg/3.1.75>
- Gilfillan, S. (1935a). *Inventing the ship: A study of the inventions made in her history between floating log and rotorship; a self-contained but companion volume to the author's "Sociology of invention"; with 80 illustrations, bibliographies, notes and index*. Follett publishing company.
- Gilfillan, S. (1935b). *The Sociology of Invention: An Essay in the Social Causes of Technic Invention and Some of Its Social Results; Especially as Demonstrated in the History of the Ship. A Companion Volume to the Same Author's Inventing the Ship*. Follett Publishing Company.
- Gilsing, V., Nooteboom, B., Vanhaverbeke, W., Duysters, G., & van den Oord, A. (2008). Network embeddedness and the exploration of novel technologies: Technological distance, betweenness centrality and density. *Research Policy*, 37(10), 1717–1731. <https://doi.org/10.1016/j.respol.2008.08.010>

- Golosovsky, M. (2017). Power-law citation distributions are not scale-free [arXiv: 1708.01859]. *Physical Review E*, 96(3), 032306. <https://doi.org/10.1103/PhysRevE.96.032306>
- Herstad, S., Aslesen, H., & Ebersberger, B. (2014). On industrial knowledge bases, commercial opportunities and global innovation network linkages. *Research Policy*, 43(3), 495–504. <https://doi.org/10.1016/j.respol.2013.08.003>
- Hölzl, W., & Janger, J. (2014). Distance to the frontier and the perception of innovation barriers across European countries. *Research Policy*, 43(4), 707–725. <https://doi.org/10.1016/j.respol.2013.10.001>
- Hötte, K., Jee, S. J., & Srivastav, S. (2021). Knowledge for a warmer world: A patent analysis of climate change adaptation technologies [arXiv: 2108.03722]. Retrieved August 29, 2021, from <http://arxiv.org/abs/2108.03722>
- Hötte, K., Pichler, A., & Lafond, F. (2020). The rise of science in low-carbon energy technologies [arXiv: 2004.09959]. *arXiv:2004.09959 [cs, econ, q-fin]*. Retrieved January 12, 2021, from <http://arxiv.org/abs/2004.09959>
- Hu, X., Rousseau, R., & Chen, J. (2011). On the definition of forward and backward citation generations. *Journal of Informetrics*, 5(1), 27–36. <https://doi.org/10.1016/j.joi.2010.07.004>
- Huenteler, J., Ossenbrink, J., Schmidt, T. S., & Hoffmann, V. H. (2016). How a product's design hierarchy shapes the evolution of technological knowledge—Evidence from patent-citation networks in wind power. *Research Policy*, 45(6), 1195–1217. <https://doi.org/10.1016/j.respol.2016.03.014>
- Huenteler, J., Schmidt, T. S., Ossenbrink, J., & Hoffmann, V. H. (2016). Technology life-cycles in the energy sector — Technological characteristics and the role of deployment for innovation. *Technological Forecasting and Social Change*, 104, 102–121. <https://doi.org/10.1016/j.techfore.2015.09.022>
- Hummon, N. P., & Dereian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks*, 11(1), 39–63. [https://doi.org/10.1016/0378-8733\(89\)90017-8](https://doi.org/10.1016/0378-8733(89)90017-8)
- IPCC. (2014). *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.
- IPCC. (2018). *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty* (tech. rep.).
- IRENA. (2018). *Global Energy Transformation: A Roadmap to 2050*. Retrieved July 18, 2018, from </publications/2018/Apr/Global-Energy-Transition-A-Roadmap-to-2050>
- ISSN international Centre. (n.d.). Retrieved August 29, 2018, from <https://portal.issn.org/>
- Jaffe, A. (1989). Characterizing the "technological position" of firms, with application to quantifying technological opportunity and research spillovers. *Research Policy*, 18(2), 87–97. [https://doi.org/10.1016/0048-7333\(89\)90007-3](https://doi.org/10.1016/0048-7333(89)90007-3)
- Jaffe, A., Trajtenberg, M., & Fogarty, M. (2000). The NBER/sloan project on industrial technology and productivity: Incorporating learning from plant visits

- and interviews into economic research - Knowledge spillovers and patent citations: Evidence from a survey of inventors. *American Economic Review*, *90*(2), 215–218.
- Jaffe, A., Trajtenberg, M., & Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics*, *108*(3), 577–598. <https://doi.org/10.2307/2118401>
- Jaffe, A. B., & Lerner, J. (2004). *Innovation and Its Discontents*. Princeton University Press. <https://doi.org/10.2307/j.ctt7t655>
- Jaffe Adam B., & de Rassenfosse Gaétan. (2017). Patent citation data in social science research: Overview and best practices. *Journal of the Association for Information Science and Technology*, *68*(6), 1360–1374. <https://doi.org/10.1002/asi.23731>
- Jensen, M. B., Johnson, B., Lorenz, E., & Lundvall, B. A. (2007). Forms of knowledge and modes of innovation. *Research Policy*, *36*(5), 680–693. <https://doi.org/10.1016/j.respol.2007.01.006>
- Kaplan, S., & Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal*, *36*(10), 1435–1457. <https://doi.org/10.1002/smj.2294>
- Katzav, E., Nitzan, M., ben-Avraham, D., Krapivsky, P. L., Kuhn, R., Ross, N., & Biham, O. (2015). Analytical results for the distribution of shortest path lengths in random networks [Publisher: IOP Publishing]. *EPL (Europhysics Letters)*, *111*(2), 26006. <https://doi.org/10.1209/0295-5075/111/26006>
- Keller, W. (2004). International Technology Diffusion. *Journal of Economic Literature*, *42*(3), 752–782. <https://doi.org/10.1257/0022051042177685>
- Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2018). *Measuring Technological Innovation over the Long Run* (tech. rep. No. 25266) [Publication Title: NBER Working Papers]. National Bureau of Economic Research, Inc. Retrieved January 25, 2021, from <https://ideas.repec.org/p/nbr/nberwo/25266.html>
- Koutras, M. (1982). Non-central stirling numbers and some applications. *Discrete Mathematics*, *42*(1), 73–89. [https://doi.org/10.1016/0012-365X\(82\)90056-5](https://doi.org/10.1016/0012-365X(82)90056-5)
- Lee, J.-S., Park, J.-H., & Bae, Z.-T. (2017). The effects of licensing-in on innovative performance in different technological regimes. *Research Policy*, *46*(2), 485–496. <https://doi.org/10.1016/j.respol.2016.12.002>
- Leydesdorff, L., & Zhou, P. (2007). Nanotechnology as a field of science: Its delineation in terms of journals and patents. *Scientometrics*, *70*(3), 693–713. <https://doi.org/10.1007/s11192-007-0308-0>
- Li, D., Heimeriks, G., & Alkemade, F. (2020). The emergence of renewable energy technologies at country level: Relatedness, international knowledge spillovers and domestic energy markets [Publisher: Routledge]. *Industry and Innovation*, *27*(9), 991–1013. <https://doi.org/10.1080/13662716.2020.1713734>
- Library, U. (n.d.). Science and Engineering Journal Abbreviations | Woodward Library. Retrieved August 29, 2018, from <https://woodward.library.ubc.ca/research-help/journal-abbreviations/>
- Lundvall, B.-A., & Johnson, B. (1994). The learning economy [Publisher: Taylor & Francis]. *Journal of industry studies*, *1*(2), 23–42.
- Magerman, T., Van Looy, B., Song, X., European Commission, & Eurostat. (2006). *Data production methods for harmonised patent statistics: Patentee name*

- harmonisation*. [OCLC: 904336460]. Publications Office. Retrieved January 5, 2021, from http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-AV-06-002/EN/KS-AV-06-002-EN.PDF
- Malerba, F. (2005). Sectoral systems of innovation: A framework for linking innovation to the knowledge base, structure and dynamics of sectors. *Economics of Innovation and New Technology*, 14(1-2), 63–82. <https://doi.org/10.1080/1043859042000228688>
- Malerba, F., & Orsenigo, L. (1993). Technological Regimes and Firm Behavior [Publisher: Oxford Academic]. *Industrial and Corporate Change*, 2(1), 45–71. <https://doi.org/10.1093/icc/2.1.45>
- Malerba, F., & Orsenigo, L. (1996). Schumpeterian patterns of innovation are technology-specific. *Research Policy*, 25(3), 451–478. [https://doi.org/10.1016/0048-7333\(95\)00840-3](https://doi.org/10.1016/0048-7333(95)00840-3)
- Malerba, F., Orsenigo, L., & Peretto, P. (1997). Persistence of innovative activities, sectoral patterns of innovation and international technological specialization. *International Journal of Industrial Organization*, 15(6), 801–826. [https://doi.org/10.1016/S0167-7187\(97\)00012-X](https://doi.org/10.1016/S0167-7187(97)00012-X)
- Mansfield, E. (1995). Academic Research Underlying Industrial Innovations: Sources, Characteristics, and Financing [Publisher: The MIT Press]. *The Review of Economics and Statistics*, 77(1), 55–65. <https://doi.org/10.2307/2109992>
- Markard, J., Raven, R., & Truffer, B. (2012). Sustainability transitions: An emerging field of research and its prospects. *Research Policy*, 41(6), 955–967. <https://doi.org/10.1016/j.respol.2012.02.013>
- Markard, J., & Truffer, B. (2006). Innovation processes in large technical systems: Market liberalization as a driver for radical change? *Research Policy*, 35(5), 609–625. <https://doi.org/10.1016/j.respol.2006.02.008>
- Martin, R. (2012). Measuring Knowledge Bases in Swedish Regions [Publisher: Routledge _eprint: <https://doi.org/10.1080/09654313.2012.708022>]. *European Planning Studies*, 20(9), 1569–1582. <https://doi.org/10.1080/09654313.2012.708022>
- Martinelli, A., & Nomaler, O. (2014). Measuring knowledge persistence: A genetic approach to patent citation networks. *Journal of Evolutionary Economics*, 24(3), 623–652. <https://doi.org/10.1007/s00191-014-0349-5>
- Marx, M., & Fuegi, A. (2019). *Reliance on Science: Worldwide Front-Page Patent Citations to Scientific Articles* (SSRN Scholarly Paper No. ID 3331686). Social Science Research Network. Rochester, NY. <https://doi.org/10.2139/ssrn.3331686>
- Marx, M., & Fuegi, A. (2020). Reliance on Science in Patenting [type: dataset]. <https://doi.org/10.5281/zenodo.4235193>
- Mazzucato, M. (2016). From market fixing to market-creating: A new framework for innovation policy. *Industry and Innovation*, 23(2), 140–156. <https://doi.org/10.1080/13662716.2016.1146124>
- McKinsey. (2013). Pathways to a low-carbon economy: Version 2 of the global greenhouse gas abatement cost curve | McKinsey. Retrieved November 13, 2018, from <https://www.mckinsey.com/business-functions/sustainability-and-resource-productivity/our-insights/pathways-to-a-low-carbon-economy>

- McMillan, G., Narin, F., & Deeds, D. L. (2000). An analysis of the critical role of public science in innovation: The case of biotechnology. *Research Policy*, 29(1), 1–8. [https://doi.org/10.1016/S0048-7333\(99\)00030-X](https://doi.org/10.1016/S0048-7333(99)00030-X)
- McShea, D. W. (1991). Complexity and evolution: What everybody knows. *Biology and Philosophy*, 6(3), 303–324. <https://doi.org/10.1007/BF00132234>
- Merges, R. P., & Nelson, R. R. (1994). On limiting or encouraging rivalry in technical progress: The effect of patent scope decisions. *Journal of Economic Behavior & Organization*, 25(1), 1–24. [https://doi.org/10.1016/0167-2681\(94\)90083-3](https://doi.org/10.1016/0167-2681(94)90083-3)
- Meyer, M. (2000). Does science push technology? Patents citing scientific literature. *Research Policy*, 29(3), 409–434. [https://doi.org/10.1016/S0048-7333\(99\)00040-2](https://doi.org/10.1016/S0048-7333(99)00040-2)
- Mitcham, C. (1978). Types of technology. *Research in philosophy and technology*, 1(1), 229–294.
- Moodysson, J., Coenen, L., & Asheim, B. (2008). Explaining spatial patterns of innovation: Analytical and synthetic modes of knowledge creation in the Medicon Valley life-science cluster. *Environment and Planning A*, 40(5), 1040–1056. <https://doi.org/10.1068/a39110>
- Murmann, J., & Frenken, K. (2005). Toward a Systematic Framework for Research on Dominant Designs, Technological Innovations, and Industrial Change. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.737063>
- Napolitano, L., Evangelou, E., Pugliese, E., Zeppini, P., & Room, G. (2018). Technology networks: The autocatalytic origins of innovation. *Royal Society Open Science*, 5(6), 172445. <https://doi.org/10.1098/rsos.172445>
- Narin, F., & Noma, E. (1985). Is technology becoming science? *Scientometrics*, 7(3–6), 369–381. <https://doi.org/10.1007/BF02017155>
- Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between U.S. technology and public science. *Research Policy*, 26(3), 317–330. [https://doi.org/10.1016/S0048-7333\(97\)00013-9](https://doi.org/10.1016/S0048-7333(97)00013-9)
- Nelson, & Winter. (1977). In search of useful theory of innovation. *Research Policy*, 6(1), 36–76. [https://doi.org/10.1016/0048-7333\(77\)90029-4](https://doi.org/10.1016/0048-7333(77)90029-4)
- Nelson, & Winter. (1982). *Evolutionary Theory of Economic Change*. Harvard University Press.
- Nemet, G. F. (2012). Inter-technology knowledge spillovers for energy technologies. *Energy Economics*, 34(5), 1259–1270. <https://doi.org/10.1016/j.eneco.2012.06.002>
- Newman, M. (2010). *Networks: An Introduction* [Publication Title: Networks]. Oxford University Press. Retrieved October 16, 2020, from <https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199206650.001.0001/acprof-9780199206650>
- Newton, I. (1675). Isaac Newton letter to Robert Hooke, 1675.
- Noailly, J., & Ryfisch, D. (2015). Multinational firms and the internationalization of green R&D: A review of the evidence and policy implications [Publisher: Elsevier BV]. *Energy Policy*, 83, 218–228. <https://doi.org/10.1016/j.enpol.2015.03.002>
- Noailly, & Shestalova. (2013a). Kennisspillovers van duurzame energietechnologieën, lessen op basis van patentverwijzingen | CPB.nl. Retrieved July 19, 2018, from <http://www.cpb.nl/en/publication/knowledge-spillovers-renewable-energy-technologies-lessons-patent-citations>

- Noailly, & Shestalova. (2013b). Wat zijn de kennisbronnen van innovaties op het gebied van duurzame energie? | CPB.nl. Retrieved July 19, 2018, from <http://www.cpb.nl/en/publication/on-which-technologies-do-renewable-energy-innovations-build-on>
- Ocampo-Corrales, D. B., Moreno, R., & Suriñach, J. (2020). Knowledge flows and technologies in renewable energies at the regional level in Europe [Publisher: Routledge _eprint: <https://doi.org/10.1080/00343404.2020.1807489>]. *Regional Studies*, *0*(0), 1–12. <https://doi.org/10.1080/00343404.2020.1807489>
- OECD. (2010). *Measuring Innovation*. OECD Publishing. <https://doi.org/10.1787/9789264059474-en>
- OECD. (2015). Frascati Manual 2015. Retrieved July 19, 2018, from https://www.oecd-ilibrary.org/science-and-technology/frascati-manual-2015_9789264239012-en
- Ogburn, W. (1922). *Social Change with Respect to Culture and Original Nature*. B.W. Huebsch, Incorporated.
- Pavitt, K. (1984). Sectoral patterns of technical change: Towards a taxonomy and a theory. *Research Policy*, *13*(6), 343–373. [https://doi.org/10.1016/0048-7333\(84\)90018-0](https://doi.org/10.1016/0048-7333(84)90018-0)
- Persoon, P. G. J., Bekkers, R. N. A., & Alkemade, F. (2020). The science base of renewables. *Technological Forecasting and Social Change*, *158*, 120121. <https://doi.org/10.1016/j.techfore.2020.120121>
- Persoon, P., Bekkers, R., & Alkemade, F. (2021). How cumulative is technological knowledge? *Quantitative Science Studies*, 1–27. https://doi.org/10.1162/qss_a_00140
- Pitt-Rivers, A. (2018). *The Evolution of Culture: And Other Essays (Classic Reprint)*. Fb&c Limited.
- Plum, O., & Hassink, R. (2012). Analysing the knowledge base configuration that drives southwest Saxony’s automotive firms: [Publisher: SAGE PublicationsSage UK: London, England]. *European Urban and Regional Studies*. <https://doi.org/10.1177/0969776412454127>
- Popp, D. (2017). From science to technology: The value of knowledge from different energy research institutions. *Research Policy*, *46*(9), 1580–1594. <https://doi.org/10.1016/j.respol.2017.07.011>
- Price, D. d. S. (1965a). Is Technology Historically Independent of Science? A Study in Statistical Historiography. *Technology and Culture*, *6*(4), 553. <https://doi.org/10.2307/3101749>
- Price, D. d. S. (1965b). Networks of Scientific Papers. *Science*, *149*(3683), 510–515. <https://doi.org/10.1126/science.149.3683.510>
- Price, D. d. S. (1976). A General Theory of Bibliometric and Other Cumulative Advantage Processes [Num Pages: 15 Place: New York, United States, New York Publisher: Wiley Periodicals Inc.]. *Journal of the American Society for Information Science (pre-1986)*; *New York*, *27*(5), 292–306. Retrieved May 18, 2020, from <http://search.proquest.com/docview/216627603/abstract/5D59167716B74D8APQ/1>
- Richerson, P., & Boyd, R. (2008). *Not By Genes Alone: How Culture Transformed Human Evolution*. University of Chicago Press.

- Ritchie, & Roser. (2020). Emissions by sector. Retrieved April 16, 2021, from <https://ourworldindata.org/emissions-by-sector>
- Rosenberg, N. (1976). *Perspectives on Technology*. Cambridge University Press.
- Schmidt, M. D. (n.d.). Generalized j-Factorial Functions, Polynomials, and Applications, 54.
- Schmidt, M. D. (2016). Combinatorial Identities for Generalized Stirling Numbers Expanding j -Factorial Functions and the j -Harmonic Numbers [arXiv: 1611.04708]. *arXiv:1611.04708 [math]*. Retrieved June 7, 2018, from <http://arxiv.org/abs/1611.04708>
- Schoenmakers, W., & Duysters, G. (2010). The technological origins of radical inventions. *Research Policy*, 39(8), 1051–1059. <https://doi.org/10.1016/j.respol.2010.05.013>
- Scotchmer, S. (1991). Standing on the Shoulders of Giants: Cumulative Research and the Patent Law. *Journal of Economic Perspectives*, 5(1), 29–41. <https://doi.org/10.1257/jep.5.1.29>
- Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2009). Comparative study on methods of detecting research fronts using different types of citation. *Journal of the American Society for Information Science and Technology*, 60(3), 571–580. <https://doi.org/10.1002/asi.20994>
- Simon, H. A. (1962). The Architecture of Complexity [Publisher: American Philosophical Society]. *Proceedings of the American Philosophical Society*, 106(6), 467–482. Retrieved September 14, 2020, from <http://www.jstor.org/stable/985254>
- Steinbock, C., Biham, O., & Katzav, E. (2019). Analytical results for the in-degree and out-degree distributions of directed random networks that grow by node duplication [Publisher: IOP Publishing]. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(8), 083403. <https://doi.org/10.1088/1742-5468/ab3191>
- Strumsky, D., & Lobo, J. (2015). Identifying the sources of technological novelty in the process of invention. *Research Policy*, 44(8), 1445–1461. <https://doi.org/10.1016/j.respol.2015.05.008>
- Suárez, D. (2014). Persistence of innovation in unstable environments: Continuity and change in the firm's innovative behavior. *Research Policy*, 43(4), 726–736. <https://doi.org/10.1016/j.respol.2013.10.002>
- Tennie, C., Call, J., & Tomasello, M. (2009). Ratcheting up the ratchet: On the evolution of cumulative culture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1528), 2405–2415. <https://doi.org/10.1098/rstb.2009.0052>
- Toynbee, A. (1963). *A Study of History: Introduction the Genesis of Civilizations*. Oxford University Press.
- Trajtenberg, M. (1990). A Penny for Your Quotes: Patent Citations and the Value of Innovations. *The RAND Journal of Economics*, 21(1), 172–187. <https://doi.org/10.2307/2555502>
- Trajtenberg, M., Henderson, R., & Jaffe, A. (1997). University Versus Corporate Patents: A Window On The Basicness Of Invention. *Economics of Innovation and New Technology*, 5(1), 19–50. <https://doi.org/10.1080/10438599700000006>

- Vaesen, K., & Houkes, W. (2017). Complexity and technological evolution: What everybody knows? *Biology & Philosophy*, *32*(6), 1245–1268. <https://doi.org/10.1007/s10539-017-9603-1>
- Valverde, S., Solé, R. V., Bedau, M. A., & Packard, N. (2007). Topology and evolution of technology innovation networks. *Physical Review E*, *76*(5). <https://doi.org/10.1103/PhysRevE.76.056118>
- Van der Loo, M. P. (2014). The stringdist package for approximate string matching. *The R Journal*, *6*(1), 111–122.
- Van Looy, B., Plessis, M. d., Magerman, T., European Commission, & Eurostat. (2006). *Data production methods for harmonised patent statistics: Assignee sector allocation*. [OCLC: 904336127]. Publications Office. Retrieved January 5, 2021, from <http://bookshop.europa.eu/uri?target=EUB:NOTICE:KSAV06001:EN:HTML>
- van Vianen, B., Moed, H., & van Raan, A. (1990). An exploration of the science base of recent technology. *Research Policy*, *19*(1), 61–81. [https://doi.org/10.1016/0048-7333\(90\)90034-4](https://doi.org/10.1016/0048-7333(90)90034-4)
- Vazquez, A. (2001). Statistics of citation networks [arXiv: cond-mat/0105031]. Retrieved May 17, 2020, from <http://arxiv.org/abs/cond-mat/0105031>
- Veefkind, V., Hurtado-Albir, J., Angelucci, S., Karachalios, K., & Thumm, N. (2012). A new EPO classification scheme for climate change mitigation technologies. *World Patent Information*, *34*(2), 106–111. <https://doi.org/10.1016/j.wpi.2011.12.004>
- Verbeek, A., Debackere, K., Luwel, M., Andries, P., Zimmermann, E., & Deleus, F. (n.d.). Linking science to technology: Using bibliographic references in patents to build linkage schemes. *Scientometrics*, *54*(3), 399–420. <https://doi.org/10.1023/A:1016034516731>
- Verhoeven, D., Bakker, J., & Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. *Research Policy*, *45*(3), 707–723. <https://doi.org/10.1016/j.respol.2015.11.010>
- Verspagen, B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in Complex Systems*, *10*(01), 93–115. <https://doi.org/10.1142/S0219525907000945>
- Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*(2), 365–391. <https://doi.org/10.1016/j.joi.2016.02.007>
- Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying Long-Term Scientific Impact. *Science*, *342*(6154), 127–132. <https://doi.org/10.1126/science.1237825>
- Wang, & Guan. (2011). Measuring science–technology interactions using patent citations and author–inventor links: An exploration analysis from Chinese nanotechnology. *Journal of Nanoparticle Research*, *13*(12), 6245–6262. <https://doi.org/10.1007/s11051-011-0549-y>
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks [Number: 6684 Publisher: Nature Publishing Group]. *Nature*, *393*(6684), 440–442. <https://doi.org/10.1038/30918>
- Web of Science Core Collection Help. (n.d.). Retrieved March 16, 2020, from https://images.webofknowledge.com/images/help/WOS/hp_subject_category_terms_tasca.html

- West, J. D., Bergstrom, C. T., & Bergstrom, T. C. (2010). The Eigenfactor Metrics: A network approach to assessing scholarly journals. Retrieved March 16, 2020, from <https://escholarship.org/uc/item/41h94387>
- Wilf, H. S. (1993). The asymptotic behavior of the stirling numbers of the first kind. *Journal of Combinatorial Theory, Series A*, 64(2), 344–349. [https://doi.org/10.1016/0097-3165\(93\)90103-F](https://doi.org/10.1016/0097-3165(93)90103-F)
- Wilson, C. (2012). Up-scaling, formative phases, and learning in the historical diffusion of energy technologies. *Energy Policy*, 50, 81–94. <https://doi.org/10.1016/j.enpol.2012.04.077>
- Winter, S. G. (1984). Schumpeterian competition in alternative technological regimes. *Journal of Economic Behavior & Organization*, 5(3), 287–320. [https://doi.org/10.1016/0167-2681\(84\)90004-0](https://doi.org/10.1016/0167-2681(84)90004-0)
- Wu, Y., Fu, T. Z. J., & Chiu, D. M. (2014). Generalized preferential attachment considering aging. *Journal of Informetrics*, 8(3), 650–658. <https://doi.org/10.1016/j.joi.2014.06.002>
- Yu, G. (2018). Scatterpie: Scatter Pie Plot. Retrieved December 21, 2018, from <https://CRAN.R-project.org/package=scatterpie>



*Eindhoven University of Technology
School of Innovation Sciences*