Knowledge Diffusion and the Dynamics of Citation Accrual

Kyle Higham

VICTORIA UNIVERSITY OF WELLINGTON

Te Whare Wānanga o te Ūpoko o te Ika a Māui

School of Chemical and Physical Sciences

Te Wānanga Matū

A thesis

submitted to the Victoria University of Wellington
in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in Physics.

Victoria University of Wellington

2019

Abstract

The diffusion of knowledge through society proceeds like an invisible ripple that moves between agents through multiple information channels. However, some types of knowledge are recorded, systematised and digitised for the benefit of everyone. Patents and academic articles are examples of such codified knowledge. These documents also contain a common element that is utilised for linking new and established knowledge: citations.

This thesis harnesses citations in patents and scientific articles as proxies for signifying the existence of knowledge flows between cited and citing documents, focusing primarily on the dynamics of citation accumulation and the mechanisms governing these dynamics. For this purpose, it is helpful to think of patents and their citations as nodes and links, respectively, in a network where new nodes join the network and distribute their citations among existing nodes. This mode of thinking leads directly to the question: How does the citation network grow? This thesis addresses that question both empirically and theoretically.

Two mechanisms that can explain much of the observed citation dynamics are preferential attachment and node ageing. The former mechanism reflects the tendency for successful nodes (by citation count) to become even more successful, while the latter captures the propensity for knowledge to become obsolete over time. The independence of these phenomena is nontrivial, but has generally been assumed. We put this assumption to the test for both patent and scientific-article citation networks and found it to be generally true if precautions are taken to account for important context surrounding the meaning of citations. Achieving a clear separation of these mechanisms is found to be very useful both mathematically and empirically, as they can now be studied independently.

Patents are particularly sophisticated documents, with various components holding specific legal meanings. Associating certain properties of these components with popularity in the form of citation accrual creates a rare opportunity to build a framework that can identify *ex-ante* node fitnesses and examine their effect on the growth of a citation network. We find that a significant portion of the preferential-attachment process observed in the patent-citation network can be attributed to basic properties of patents

determined by their time of grant. Besides suggesting novel approaches towards estimating patent quality, the results of our work also provide a platform for gaining a deeper understanding of the various mechanisms that underpin the success-breeds-success dynamics ubiquitously observed in complex systems.

Acknowledgements

I first want to express my deepest gratitude to my supervisors, Professors Michele Governale, Adam Jaffe, and Uli Zülicke. I know that I was an unconventional student working on an unconventional project for all of you, and I am grateful that you took this risk on me. I have always maintained that I would not have taken up this project with a less passionate and functional support team, and I am proud to have attracted the envy of my peers in this respect. You understood my strengths and weaknesses and reacted accordingly to ensure my smooth and successful candidature. Thank you for your encouragement, your mentorship, and, most of all, your patience.

My doctoral journey was funded by Te Pūnaha Matatini, and I would like to thank Shaun Hendy, Kate Hannah, Kathryn Morgan, and everyone else involved in this wonderful research centre. At the start of my candidature, it quickly became obvious that you are not just facilitating the transformation of a good student into a good researcher, but you have cultivated an environment to ensure that everyone, Whānau and investigators alike, is acutely aware of the social contexts within which research is conducted. I am incredibly proud to have been involved in this endeavour, and honestly believe that I am now a better human as a direct result of your efforts.

I would also like to thank the Te Pūnaha Matatini Whānau. My work was often academically isolating, as is often the nature of such trans-disciplinary research, but getting to know others in the same situation and being in a position to bring everyone together through my roles on the committee was very comforting and rewarding. Thank you for your camaraderie, you are all awesome people.

To my Fock Space colleagues, it has been a delight sharing an office with you. Thank you to my academic siblings Markus and Lei for great discussions and for providing welcome distractions. I am sorry I can't sustain an intelligent conversation about your research.

When I first started this project, I mostly worked at Motu Economic and Public Policy Research. I would like to thank everyone down there at your lovely institute for your help introducing me to the computational tools I needed for my research, and for continuing to invite me back to enjoy your company long after I vacated my desk.

Lastly, and most importantly, this thesis is dedicated to my wonderful wife Ella and our family. Mum, Dad, thank you for giving me everything I needed to explore my interests and your unwavering support along every path I have chosen to wander down. Ella, none of this would have been possible without you by my side. I can't wait for all of the adventures that await us.

Contents

1	Introduction	1
1.1	The Economy is a Complex System	1
1.2	Collaboration is Key	2
1.3	Building (Sturdy) Bridges	3
1.4	Real-World Impact	4
1.5	Open Questions	5
1.6	Philosophical Positioning and Scope	6
1.7	Thesis Structure	8
2	Patents and Innovation Dynamics	11
2.1	What is a Patent?	13
2.1.1	Legal Requirements	13
2.1.2	Patent Documents as Data	14
2.2	Patents as a Proxy for Inventive Activity	16
2.2.1	Patents and R&D	16
2.2.2	The Invention Machine	16
2.2.3	What Inventive Activity Do Patents Capture?	17
2.2.4	Bureaucratic Effects on Patent Grants	18
2.3	Patent Citations	19
2.3.1	The Citation Process	19
2.3.2	A Geometric Analogy	21
2.3.3	Citations as Indicators	22
2.3.4	Differences Between Academic Citations and Patent Citations . . .	26
2.3.5	Examiner Effects on Patent Citation	27
2.4	Innovation Dynamics	28
2.4.1	Knowledge Diffusion	28
2.4.2	Technological Obsolescence	30

3	Growing Directed Acyclic Networks	33
3.1	Characterisation of Citation Networks	34
3.2	Models of Citation Network Growth	36
3.2.1	A Simplified Measurement Framework	36
3.2.2	The Master Equation Approach	37
3.2.3	Random Attachment	41
3.2.4	Preferential Attachment	43
3.2.5	Node Ageing	45
3.2.6	Non-linear Preferential Attachment with Node Ageing	47
3.3	Structural Concerns and Alternative Models	48
3.3.1	Transitive Reduction	48
3.3.2	Copying Models	50
3.4	A Useful Illusion	51
4	The Dynamics of Patent Citations	53
4.1	Patent Citation Dynamics	53
4.1.1	Perspectives from Economics	53
4.1.2	Applications of Network Science	55
4.1.3	The Separability Assumption	57
4.2	Patent Citation Data	58
4.2.1	Data Sources	58
4.2.2	Technology Classification	59
4.2.3	Cohort Selection	60
4.3	Data Preparation	62
4.3.1	The Citation Lag	62
4.3.2	Citation Inflation	63
4.3.3	Binning	66
4.4	Modelling Strategy	68
4.4.1	Observation of Preferential Attachment and Ageing Phenomena	68
4.4.2	Model Fitting	69
4.5	Results & Discussion	70
4.5.1	Preferential Attachment	70
4.5.2	Ageing	72
4.5.3	The Degree Distribution and Related Quantities	74
4.5.4	Alternative Functional Forms of Rate Function Components	76
4.6	Conclusions & Outlook	80

4.7	Additional Tables & Figures for Chapter 4	83
5	The Dynamics of Scientific Citations	89
5.1	Scientific Citation Dynamics	89
5.1.1	Ageing	90
5.1.2	Preferential Attachment	92
5.2	Scientific Citation Data	94
5.2.1	The American Physical Society Dataset	94
5.2.2	A Note on Scientific Literature Classification Schemes	95
5.2.3	Cohort Selection	95
5.3	Data Preparation	98
5.3.1	Redefining the Citation Lag	99
5.3.2	Citation Inflation	99
5.4	Modelling Strategy	100
5.5	Results & Discussion	102
5.5.1	Preferential Attachment	103
5.5.2	Obsolescence	105
5.5.3	Evidence for Rate Separation in Citations to Scientific Articles . .	106
5.6	Conclusions & Outlook	109
5.6.1	Striking Similarities	110
5.7	Additional Figures for Chapter 5	113
6	Patent Quality and Network Growth	115
6.1	Patent Quality and Fitness in Networks	115
6.1.1	Measuring Patent Quality and Value	116
6.1.2	Node Fitness in Growing Networks	118
6.1.3	Patent Quality as Node Fitness	120
6.2	Ex-Ante Determinants of Patent Quality	122
6.3	Constructing a Fitness Metric	128
6.3.1	Dimensionality Reduction	129
6.3.2	Fitness Metric Composition	136
6.3.3	Calibration of Fitness Distribution	142
6.4	Integration of Fitness into a Growing Network Model	146
6.4.1	Consequences for the Preferential Attachment Strength	148
6.4.2	Consequences for Obsolescence	151
6.5	Conclusions and Outlook	154

6.6	Additional Figures for Chapter 6	157
7	Conclusions and Outlook	159
7.1	Conclusions	159
7.2	Biological Approaches to Technological Evolution	161
7.3	Changing Citation Behaviours & Network Structure	162
7.4	Comparisons to be Made (or Not)	163
7.4.1	Scientific Citations	163
7.4.2	European Patent Office Data	164
7.5	Extending Patent Fitness	165
7.6	Parting Words	170

Chapter 1

Introduction

In the age of big data and ever-increasing computational power, it is not surprising that fields of study with the potential to simplify and extract useful information from massive datasets have exploded in recent decades (e.g., [1–3]). In particular, scientists researching complex systems in areas such as statistical mechanics or evolutionary biology have now found themselves able to apply that technical knowledge to a vast range of problems in areas previously inaccessible to empirical rigour [4–9].

The field of economics is particular is known for its empirical rigour. Since the birth of the computer, economists have been processing large amounts of data in statistically robust ways [10]. However, many popular frameworks that have grown out of the study of complex systems have yet to enter mainstream economics. In particular, statistical mechanics has produced a number of new mathematical frameworks for thinking about large systems of interacting parts – the very definition of a complex system. Conventionally, these interacting components would be particles, spins, or harmonic oscillators. We have a good understanding of how each of these components behaves, the range of system states that these behaviours lead to, and ways to test any predictions made. The application of these mathematical tools, combined with existing econometric methodologies, certainly appears to be a potentially productive approach.

1.1 The Economy is a Complex System

The economy itself has long been understood to be a complex system [11]. It consists of not only many components, but many *types* of components, and an unquantifiable range

of behaviours and states displayed by each type. Components can be individual humans, companies, governments, or anything these entities produce. The work detailed in this thesis is focused on shining a light into a small corner of the broader topics of knowledge diffusion¹ and the dynamics of invention. The components under consideration in this complex system are *inventions* and *scientific articles*, and in particular those inventions with codified relationships to other components of the system - inventors, scientists, firms, and industries to name but a few. Because these relationships are systematised and digitised, there is little ambiguity about the existence of relationships between these components. Therefore, a complex-systems framework is well suited to help us to answer questions about the spread of knowledge between the different components of this system.

1.2 Collaboration is Key

While the *existence* of relationships between the components of our system is unambiguous, the *meaning* of those relationships is less certain. It is at this point that domain knowledge is vital to design and interpret the results of any analyses undertaken. This ingredient is so often missing from research applying complex-systems frameworks to economic questions, and this deficiency is one I hope to address in this work.

The most explicit relationships between technological inventions, in particular, are citations between patents, and indeed these are the relationships on which I will focus for the majority of this thesis. Most of the research done to date in bibliometric (document) networks, however, concerns scientific articles. This is not unexpected, after all, the scientists applying their knowledge of complex systems to these networks have a good understanding of the system within which these citations are made and, therefore, generally have both the domain knowledge and the skills with which to carry out the appropriate analyses. It also happens that patent citation networks look very similar to these scientific citation networks. As such, it is tempting to directly apply to them analyses that have proven fruitful when applied in scientific citation networks. However, patents and innovation economics generally lie outside of the knowledge domains of most complex-systems scientists, and this results in these topics getting little attention when compared to analogous work on academic research systems. Often when they do get

¹In the sense of Cabellero & Jaffe [12], referring to the adoption, augmentation, and improvement of ideas by people other than the originators of the idea.

attention, there are no domain specialists involved in the research.

Adding to the problems above, there is an apparent reluctance on the part of the economics community to experiment with new, complex-systems-based frameworks. While work utilising patent citations has been happening for about 40 years [13] (and even appearing in explicit network form very early on [14]), the application of new analytical tools from complex-systems science to innovation economics has yet to become widespread. The most impactful applications to date are often case-studies, and the tools utilised are rudimentary. In the words of Jaffe & de Rassenfosse [15],

“Studies by scholars using advanced network analysis tools offer little practical implications, whereas studies by scholars looking at real-world implications use quite basic network analysis tools. A promising way forward is to better integrate the technical and the practical aspects of network analysis.”

This statement expresses a desire to fill what appears to many as an obvious opportunity to leverage these new tools in order to close knowledge gaps through fruitful collaboration between network scientists and economists.

1.3 Building (Sturdy) Bridges

The reasons for this lack of integration are numerous. One particular problem that requires addressing is the apparent lack of awareness of or attribution to previous research in economics from new research applying tools from complex systems to economic systems [16]. It may be tempting to believe that because old concepts are being placed in a new framework, that the results will be, by definition, novel. This is often not the case, and sometimes these new frameworks merely offer complementary routes to conclusions that have been well understood in economics for many years. Additionally, just because we have an exciting new tool to deploy, does not mean that it can be directly applied in all situations, even where it may seem appropriate. Discussion or collaboration with domain specialists is vital to ensure the tool is appropriate to the problem, and if not, whether a modified approach may be developed to ensure the results are interpretable and useful [17].

The suggestions above allude to a practical approach to developing new models to understand real complex systems outside of the physical sciences; in short, the incor-

poration of relevant domain knowledge into models and their application at every step of the process – humans and our institutions are not amorphous particles and treating us as if we were may lead to misleading results. For example, generative models in network science are often constructed in order to predict or explain particular network behaviours or characteristics, where some plausible mechanisms can be shown to lead to the observed system properties. It is often challenging, however, to verify whether the proposed mechanisms reflect reality [18]. In the case of patent citation dynamics, as demonstrated in Chapter 4, a simple preferential attachment (PA) model appears to explain the observed patterns of citations very well. This does not, however, constitute evidence for PA as a mechanism. We know that many different mechanisms can mimic PA [19–22]; or rather, PA is an attractive model that can mimic real mechanisms that we do not yet understand. Unravelling the complexities of this system and discovering the real mechanisms driving the observed behaviours is much more difficult than testing a model on data [23, 24]. The reality is that human systems are often very context-dependent and annoyingly non-universal: our behaviour and institutions exhibit complex behaviours that often cannot be understood without moving away from the general and into the specific, and this should be reflected in our research [25]. It would be desirable to integrate the best of both worlds – the new analytical tools being developed to understand the behaviour of complex systems, combined with impetus grounded in the real-world problems that typify economics research.

1.4 Real-World Impact

The more we know about the process of invention and scientific research, the subsequent diffusion of the resultant information, and the development of relevant metrics to monitor progress, the better we can optimise the process of scientific and technological research and thus increase the potential to accelerate progress in all fields [26]. More specifically, the more understanding we have about what drives success in technology and science, the more we can focus on supporting these drivers and the faster we can improve the world around us.

Concerning research systems and their funding, the better informed on these topics we are as a society, the better decisions we can make. Appropriate metrics for research performance are essential for this purpose. These metrics must be grounded in sound knowledge of the complex systems in which research is embedded. In other words, if we

have a good understanding of the mechanisms by which important ideas are born, and how they subsequently spread, we can adopt more informative metrics that capture the aspects of research performance that are the most relevant to scientific and technological progress [27].

1.5 Open Questions

Citations to patents and academic papers are often considered to be indicative of knowledge flows [15, 28, 29] and are frequently used to assess the impact of technological and scientific output [30, 31]. It is well-known that highly-cited patents and papers will generally continue to attract more citations [32, 33]. It is also well-known that over the lifetime of these documents, they are generally at their most relevant at the time they are published [34]. However, the existing literature examining the citation dynamics of papers and patents, both empirical and theoretical, consistently assume that these two phenomena are wholly independent of one another [35–39].

This assumed independence is not at all self-evident, and further, the implications of such independence would be a significant finding, as it would facilitate the study of both the rise in popularity and the growing irrelevance of particular pieces of knowledge, despite these phenomena being captured by the same data (citations) and occurring at the same time. This kind of study requires the careful consideration of citation behaviour, which may differ between fields and change over time, while also keeping in mind the real significance of a citation in different contexts (in this case, patents and papers).

One avenue that could then be explored are the reasons for rising popularity in the first place. In many fields, success does indeed beget success. When the success of an academic article is proxied by citation counts, the general mechanisms responsible for this effect seem clear – the more a paper is cited, the more likely to be discovered and cited again. Patent citations, on the other hand, are strictly moderated and guided by the law. Therefore, the direct, popularity-reinforcing mechanisms prevalent in academia are almost non-existent, or at the very least, much more subtle in this context. These citations are thus likely to be highly related to intrinsic properties of the patent. This fact, in combination with the highly-structured nature of patent documents, means it is possible to incorporate some basic metadata into the analysis described in the previous paragraph to begin to understand the traits of these documents that can explain the

observed success-begets-success phenomenon.

1.6 Philosophical Positioning and Scope

Invention is a complex process with so many inputs, outputs, and interactions between agents in the innovation system that it is generally not possible to capture and incorporate all of this information into a coherent picture. When placed in a network framework, invention takes place in a highly multilayered and interconnected system, with interactions occurring between inventors, firms, attorney, examiners, and others, with little data about the information transferred across these connections or even the existence of connections in the first place. Thus, it is important to focus on particular aspects for which data exists in order to extract useful information from any empirical analysis of the system. Therefore, the available information delimits the scope of the research described herein. Patents, for example, represent a result of many long, convoluted, and unobserved processes and thus these units of observation are formally-structured bits of knowledge that embody the sum of these efforts – they are themselves a coarse-grain description of the inventive process up to that point. This view of the information contained in patents delimits the scope of this thesis, and applies equally to the scientific process and the resultant articles that are used as a point of comparison.

In particular, to answer any questions regarding the dynamics of knowledge diffusion (in the context of inventions, the incorporation of old inventions into new ones), many aspects of the innovation system will be convolved by necessity to obtain information about the average behaviours that are of interest here. This process not only allows for the identification of distinct behaviours of the system but also represents a starting point for more granular analyses. This approach is also less sensitive to idiosyncratic behaviours of the innovation system and is, therefore, more transferable across domains while facilitating the comparison of entirely different social systems – the methodologies and results of Chapters 4 and 5 have already been successfully applied in this regard [40].

Due to this aggregation of both time and units of observation, all analytic and empirical results presented throughout this thesis are not *dependent* on the use of a network framework, as no information about local network structure is utilised. However, the network framework facilitates the conceptualisation of citations as directed knowledge flows as well as providing language and mathematical results that allow both the comparison

with, and extension into, truly network-dependent research, so I borrow terminology from this literature where appropriate. Temporal networks [41] are an example of a more specific network-dependent framework into which this work may be extended. Applications of this kind of approach include field-specific case studies where relationships between inventors and firms may be established robustly [42], and there exists sufficient time-resolution for characterising these changing network structures. In addition, with more recent data it is possible to distinguish applicant-added citations from examiner-added ones, and so there may be an opportunity to reduce the signal-to-noise ratio with regard to measurable knowledge flows [43, 44]. The problem remains, however, that even the patent application dates that are available to us (and are utilised in this thesis) come after variable-length periods of ideation (see below), research, development, and drafting. The existence of these generally unobservable processes limits the amount of information that is available to extract from this data – even with perfect information about the inventor and firm networks on which knowledge is being diffused. The initial transfer of knowledge that started the invention process may occur months or years before the timestamp on the formal document.

Many informal processes occur before and after the formal patenting process, and large bodies of literature exploring these processes, that are not considered explicitly in this thesis. Chronologically, the first is *ideation* - the process of discovering or forming a good idea in the first place [45, 46]. Of relevance in the context of citation dynamics are the types of ideation that lead to both a patent application and citation to prior art (or in the case of scientific articles, publication and citations to prior literature), as these are visible to us, and the types and numbers of citations made can indicate the kind of ideation undertaken. Given that the average patent makes at least 5-10 citations to prior art, and often many more, one may hypothesise that most patents are either *derivative* ideas, which build on or modify an existing idea, or *symbiotic*, which combine different ideas in a novel way [45]. Therefore, while the ideas crystallised in patent documents in principle capture many types of inventive activity (see Section 2.2.3 for more detail), their citations will generally indicate their usefulness in the inception of ideas that emerge from derivative and symbiotic ideation processes. Also of note concerning the genesis of ideas is the relationship between the firm and the inventor: the amount of autonomy given to inventors by their employers (and the vast majority of patents are assigned to these employers) is likely to be very heterogeneous and complicates the study of ideation in industrial settings – some inventors will be more limited in their avenue of research than others. After a patent has been applied for, the patented material may

be used in some way to recoup (and are expected to exceed) its cost of development. This commercialisation process, and the *diffusion of innovation* that comes with it [47], is also reflected in patenting activity by competitors who would like to compete in this market. However, this feeds back into the ideation process of the competitors, which is unobservable in patent data and is therefore immune to an empirical treatment without additional (and generally private) information. This further justifies the aggregated approach – after all, this information is cannot be lost if it is not accessible in the first place.

Lastly, work contained in this thesis, while resulting in many exciting new findings, focuses on building empirical frameworks for the study of innovation dynamics and is inspired by methodologies commonly found in the physics and network science literature. While these frameworks are developed and applied to what is typically a concern of the economics community, no specific policy implications arise from this work – it is designed to study some of the characteristics of innovation without making a judgement on whether these characteristics are good for society at large. Those working on questions of policy may, however, be able to utilise the frameworks presented to answer specific policy-related questions. The broad policy areas in which these frameworks may provide future insights are noted where appropriate and generally concern the measurement of technological or scientific output. In short, the results emerging from this work are important for the understanding specific, fundamental aspects of the dynamics of the innovation system and does not aim to link the observed behaviours of the system to particular societal impacts.

1.7 Thesis Structure

With the scope and limitations of this thesis in mind, we can now turn to the particular topics that *are* studied. First, Chapter 2 presents the economics background relevant to patents, their citations, and their relationship with the dynamics of invention. This will provide a context within which subsequent analyses, particularly those in Chapters 4 and 6 are applied and interpreted.

Chapter 3 introduces the mathematics of the dynamics of citation accumulation. The mathematical frameworks that lay the foundation of much of the empirical work detailed in the remaining chapters are derived here. This includes the derivation of a novel and succinct master-equation-based solution to a separable rate equation for application to

the dynamics of citation accrual. This solution links specific citation accrual mechanisms such as the decay of relevance of knowledge with time and success-begets-success phenomena to the citation distribution and to any quantity that may be derived from the distribution, such as the average citation rate.

Chapters 4 and 5 apply these analytic results to real data, patent citations and scientific citations respectively. The empirical methodology developed in these chapters integrates the contextual knowledge detailed in Chapter 2 (and augmented for the scientific context in Chapter 5) into the mathematical framework developed in Chapter 3 to gain insight into the common drivers of the growth and stagnation of these citation networks. In particular, it is found that the mechanisms governing the growth and decay of citation accrual for both patents and physics papers are separable. This result has been frequently assumed in the literature; however, the results contained in these chapters constitute the first careful examination of this assumption. After field and time-varying extrinsic factors are controlled for, it is found that, indeed, the growth and decay processes of citation accrual are generally able to be separated empirically and are therefore likely to be mechanistically independent. The consequences of this finding as well as important differences that arise within and between the two sets of documents are discussed, respectively, in terms of field differences and real-world differences in citation practice.

Chapter 6 then delves into the development of a more detailed picture of the intrinsic properties of individual patents in order to understand the different dimensions of patent quality and how these relate to citation accrual, and hence knowledge diffusion. The citation growth model developed and tested in the previous chapters is augmented with an additional variable: patent fitness. It is found that a fitness metric that is carefully constructed from simple patent metadata can explain 20-30% of the cumulative advantage (preferential attachment) effect observed when patent fitness is not accounted for. This is the first demonstration of exogenous node fitness being incorporated into a growing network model and has ramifications for the interpretation of patent citations. Additionally, this approach provides a framework that can be used to understand the relative importance of cumulative advantage and node fitness effects in the growing networks examined in other fields. Chapter 7 concludes.

Chapter 2

Patents and Innovation Dynamics

In the widely cited work “Geography and Trade” [48], Paul Krugman contends that

“Knowledge flows [...] are invisible; they leave no paper trail by which they may be measured and tracked, and there is nothing to prevent the theorist from assuming anything about them that she likes.”

A lot has happened since those words were penned. Patents, and the citations between them, have become a ‘window into the knowledge economy’ [49] precisely because they leave a rich, varied, and easily accessible paper trail that can be used as a proxy to examine many different aspects of innovation dynamics when placed in their proper context. Therefore, in order to use this resource effectively, it is essential to understand both the relevant properties of patents and the complex relationship between these properties and the ideas we are testing.

Much of the invention and innovation that sustains and advances modern society is, in part, the result of patent protections. While in some industries the very idea of patent rights are mired in controversy (and sometimes justifiably so), they form the backbone of intellectual property protection in many sectors. Without patent protection, many firms would not bother to invest in research and development (henceforth R&D). Without the temporary, government-granted monopoly that patent rights bestow, there are often few barriers preventing competitors obtaining a product and replicating every useful aspect of it before commercialising their own version. In this scenario, firms would have to compete against others selling exactly the same product, and these competitors do not have to recoup the costs of inventing the product in the first place. In the absence of patent protection, firms would put themselves at a competitive disadvantage almost

every time they invest in R&D. This idea is so fundamental to technological progress in a capitalist environment that monopoly rights for inventors have been around in some form since at least the 15th century, when the architect Filippo Brunelleschi was granted a ‘patent’ for his unique river transport vessel used to transport marble for his engineering projects [50]. The patent was an exclusive right enforced by the State of Florence with the stipulation that anybody imitating this invention “shall be burned” [51]. Punishments for patent infringement have mellowed somewhat since this time.

The data source for the empirical research detailed in this thesis are those patents granted by the United States Patent and Trademark Office (USPTO), and we will, therefore, focus our attention on the relevant aspects of patent law and the patenting process specific to this office. Due to differences in patent law and patent office procedures between jurisdictions, it must be emphasised that the conclusions made in this work cannot automatically be generalised to universal phenomena – some of the results may be in part due to the eccentricities of the patent office and commercial environment particular to the United States of America [52].

At this point, it is important to justify the reasoning for choosing the USPTO as the primary source of data for this work. First, the patent data provided by the USPTO are readily available, free to access, and are accompanied by extensive descriptive documents (e.g. [53,54]) provided by The Office of the Chief Economist (OCE). The OCE is responsible for the research branch of the USPTO and therefore understand the requirements of those who would like to utilise patent data for research purposes. Second, as patent rights are restricted to the jurisdiction that granted the patent, firms worldwide will often apply for a patent in the jurisdiction with the largest market for potential customers – the United States of America is the largest market for many, if not most, commercial products. As a result, the USPTO receives more applications than any other patent office in the world except China, and the data are therefore one of the richest available by virtue of this volume. Lastly, much of the research in innovation economics is based on USPTO data (for the very reasons listed above), and this makes a comparison to earlier work possible.

The right of the government to grant patents (among other types of intellectual property) has been an important mechanism in the United States since its establishment as an independent country. The Constitution of the United States states [55]:

“The Congress shall have Power ... To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries.”

While details of this patenting process are laid out in the Patent Act of 1790 and have changed substantially in many respects, the fundamental aspects of this power have remained the same.

2.1 What is a Patent?

At its core, a patent is an agreement between the inventor and the US government whereby the inventor discloses the details of her invention. In exchange, the inventor is granted exclusionary rights by the government that prevent others from making, using, or selling the invention without permission for a set period of time. These exclusionary rights, effectively a temporary monopoly, allow the inventor time to recoup the cost of the resources used in the invention process with less concern for imitation.

This work will focus on *utility patents* which are defined by the U.S. Code [56] as “any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof.” From this point on, anything that falls under this definition will be referred to as an *invention*. While the vast majority of patents issued are utility patents, it is worth noting that the USPTO also issues patents for designs (which apply to the aesthetic properties of an object rather than their function), and plant patents (for newly discovered/invented plants that can be reproduced asexually). These types of patents are entirely excluded from all analyses in this work due to the differences in patentability requirements and examination procedures between these patents and utility patents.

2.1.1 Legal Requirements

There are a number of requirements that must be adhered to for a patent to be considered for grant. These requirements fall into five main categories [57]:

1. The invention must be patent-eligible in its subject matter, as determined by the USPTO.

2. The invention must have utility, that is, it must “perform a designed function or achieve some minimum human purpose” [57].
3. The invention must be novel. This criterion is largely determined by prior art, and therefore the establishment of novelty is one of the primary reasons for the presence of the citations listed on a patent.
4. The invention must not be obvious to a specialist in the relevant technology field.
5. The invention must be described in language that is definite and clear in its interpretation, and easily understood by a specialist in the relevant technology field.

After a patent is applied for, the inventor, their attorney, and the patent examiner generally go through a negotiation process to come to a compromise as to the extent of the technological aspects covered by the patent. The inventor would like the invention to be as broad as possible, in order to make it difficult for a competitor to invent around the patent by finding a different way of achieving the same utility as the original patent to compete without infringement. However, this goal is generally not compatible with the requirements listed above, which leads to a narrowing of patent scope until either an agreement is reached and the patent granted, or the applicant decides the new specifications are too narrow and the application is abandoned.

2.1.2 Patent Documents as Data

USPTO-granted patent documents have retained the same general structure for many decades. The strict, universal layout has not only made the transferral of patent information from paper archives into a digital database a mostly unambiguous exercise, but also means the same data is available in the same format for every granted patent since the mid-1960’s. One of the particularly useful aspects of this data from a research point of view is the fact that every piece of information on the patent is legally well-defined and relatively uniform across technology fields, which often makes the interpretation of results less subjective than for other bibliographic data.

The general structure of a utility patent document can be seen in Fig. 2.1. Additionally, there are metadata surrounding the patent that are also available to the general public. These metadata include both information about the patent before its issue, such as the details of examination procedure, and after its issue, such as whether the patent

<p>(12) United States Patent Callaghan et al.</p>	<p>(10) Patent No.: US 7,391,215 B2 (45) Date of Patent: Jun. 24, 2008</p>	<p>(d) (e)</p>																												
<hr/>																														
<p>(54) NMR APPARATUS</p> <p>(a) (75) Inventors: Paul Terence Callaghan, Wellington (NZ); Mark Warwick Hunter, Wellington (NZ); Robin Dykstra, Palmerston North (NZ); Craig David Eccles, Wellington (NZ)</p> <p>(73) Assignees: Victoria Link Limited, Kelburn, Wellington (NL); Massey University, Palmerston North (NL)</p> <p>(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 52 days.</p> <p>(21) Appl. No.: 10/520,862</p> <p>(b) (22) PCT Filed: Jul. 11, 2003</p> <p>(86) PCT No.: PCT/NZ03/00149</p> <p>§ 371 (c)(1), (2), (4) Date: Dec. 21, 2005</p> <p>(87) PCT Pub. No.: WO2004/008168</p> <p>PCT Pub. Date: Jan. 22, 2004</p> <p>(65) Prior Publication Data</p> <p>US 2006/0097725 A1 May 11, 2006</p> <p>(30) Foreign Application Priority Data</p> <p>Jul. 11, 2002 (NZ) 520114</p> <p>(c) (51) Int. Cl. G01V 3/00 (2006.01)</p> <p>(52) U.S. Cl. 324/319; 324/318</p> <p>(58) Field of Classification Search 324/319, 324/320, 318, 309, 307; 600/407, 410, 422</p> <p>See application file for complete search history.</p>	<p>(56) References Cited</p> <p>U.S. PATENT DOCUMENTS</p> <table border="0"> <tr> <td>4,498,048 A *</td> <td>2/1985</td> <td>Lee et al.</td> <td>324/309</td> </tr> <tr> <td>4,528,509 A *</td> <td>7/1985</td> <td>Radda et al.</td> <td></td> </tr> <tr> <td>4,538,130 A *</td> <td>8/1985</td> <td>Gluckstern et al.</td> <td>335/306</td> </tr> <tr> <td>4,580,098 A *</td> <td>4/1986</td> <td>Gluckstern et al.</td> <td>324/307</td> </tr> <tr> <td>4,710,713 A *</td> <td>12/1987</td> <td>Strikman</td> <td></td> </tr> <tr> <td>4,758,813 A *</td> <td>7/1988</td> <td>Holsinger et al.</td> <td>335/306</td> </tr> </table> <p>(Continued)</p> <p>FOREIGN PATENT DOCUMENTS</p> <table border="0"> <tr> <td>EP</td> <td>171931</td> <td>3/1989</td> <td></td> </tr> </table> <p>(Continued)</p> <p>OTHER PUBLICATIONS</p> <p>P.J. Prado, "NMR hand-held moisture sensor", Magnetic Resonance Imaging (pp. 505-508), (2001).</p> <p><i>Primary Examiner</i>—Brij B Shrivastav (74) <i>Attorney, Agent, or Firm</i>—Dann, Dorfinan, Herrell and Skillman, P.C.</p> <p>(57) ABSTRACT</p> <p>A magnetic assembly for a nuclear magnetic resonance apparatus includes a number of primary permanent magnets 1 disposed in an array about a longitudinal axis, the arrangement and/or characteristics of the plurality of magnets being such so as to create a zone of homogeneous magnetic field at some location along the axis forward of the array (and into the material when provided). A secondary permanent magnet 7 may be located along the longitudinal axis within the array of primary magnets and may be moveable. The primary magnets 1 have a north pole and a south pole with an axis therebetween, and may be arranged such that the axis between the poles is at an angle to the longitudinal axis.</p> <p style="text-align: center;">23 Claims, 12 Drawing Sheets</p>	4,498,048 A *	2/1985	Lee et al.	324/309	4,528,509 A *	7/1985	Radda et al.		4,538,130 A *	8/1985	Gluckstern et al.	335/306	4,580,098 A *	4/1986	Gluckstern et al.	324/307	4,710,713 A *	12/1987	Strikman		4,758,813 A *	7/1988	Holsinger et al.	335/306	EP	171931	3/1989		<p>(f) (g) (h) (i)</p>
4,498,048 A *	2/1985	Lee et al.	324/309																											
4,528,509 A *	7/1985	Radda et al.																												
4,538,130 A *	8/1985	Gluckstern et al.	335/306																											
4,580,098 A *	4/1986	Gluckstern et al.	324/307																											
4,710,713 A *	12/1987	Strikman																												
4,758,813 A *	7/1988	Holsinger et al.	335/306																											
EP	171931	3/1989																												

Figure 2.1: Example of the front page of a USPTO patent. Information pertinent to work in this thesis include (a) inventors, (b) application date, (c) CPC technology classification, (d) patent number, (e) grant date, (f) backward citations to USPTO patents, (g) backward citations to non-USPTO patents, (h) backward citations to non-patent documents, and (i) information about the total numbers of claims and figures. Note the assignees have mistakenly been given the country code for the Netherlands (NL) rather than New Zealand (NZ), this is later corrected in the database but not on this archival document.

has been renewed, litigated (although these data are limited), or whether the same invention is patented in other jurisdictions.

2.2 Patents as a Proxy for Inventive Activity

2.2.1 Patents and R&D

Schmookler [58] defines inventive activity as “work specifically directed towards the formulation of the essential properties of a novel product or process”. Assuming firms then patent these products or processes, patenting activity would appear to be a good measure of the level of inventive activity, where the inputs include resources such as time, money, and inventors. The usual way to quantify these inputs is by looking at R&D expenditure. Therefore, the first step in the validation of patent statistics as a measure of something economically interesting was to check the relationship between R&D spending and patent output. If more resources spent on inventive activity (as defined above) did not lead to increased levels of inventive output as indicated by patented inventions (providing propensity to patent remained constant), then there was little hope for their use as indicators of research activity. Much of the early work utilising patent data, therefore, aimed to link patent statistics to existing indicators of inventive activity in order to establish their usefulness.

The work of Jacob Schmookler in the 1950s and 60s was the first to attempt to empirically investigate the nature of invention and its relationship to both inputs and external economic forces [59, 60]. Patents were an obvious source – indeed one of the only sources – of data for this endeavour, and his work provided substantial evidence for the role of invention in the economic growth of a nation [58]. However, the first work to empirically examine the patent-R&D expenditure relationship directly was that of Pakes & Griliches [61]. The authors found a strong relationship between firms’ R&D investment and their patent output. Additional evidence for this claim was quickly found using other data sources and investigated in more depth [62–64], placing patent-based indicators on stable footing as a rich, plentiful, and easily accessible source of information on inventive activity.

2.2.2 The Invention Machine

In their influential 1984 work, Pakes & Griliches [65] formulate a knowledge production function that summarises the way in which research expenditure may be converted into inventions. This function may be visualised as an *invention machine* with mysterious and complex inner workings. First, firms input time, money, and other resources into the

machine, which converts them to ‘additions to economically valuable knowledge’ [66], and subsequently manifest themselves in a variety of ways as output. As these knowledge additions themselves are generally unobservable directly, we must find indirect ways of observing this output. Some of these indicators are financial, such as firm growth, productivity, or value as a function of the resources expended to produce the knowledge. However, these indicators are not only difficult to obtain data for, but they are also tied to many other economic factors other than invention that also affect output; this can present a challenge when attempting to establish causality.

Patents, on the other hand, are more like progress markers rather than explicit targets such as firm growth or productivity. They are indicators of advancement, and while they do provide legal benefits to the patent owners in the form of intangible capital [67], it is perhaps misleading to describe patents simply as outputs of the invention process. In our invention machine analogy, patents play the role of an array of meters on the side of the machine. One of the goals of empirical and theoretical research in innovation dynamics is to understand the relationship between input, output, and the readings on these meters. This is because the readings are so much more detailed and unambiguous - patenting is an unusually (but necessarily) public and transparent process. It is for this reason that much of this chapter will be spent detailing the eccentricities of the patenting process in the USA, from the ever-changing legal landscape to the incentives of the patent office to grant patents and even the characteristics of the USPTO workforce. All of these factors have consequences for the numbers and types of patent grants, the size of the inventive step, the lag between application and grant and its effect on firm patenting strategies, among countless other flow-on effects. Once these quirks are understood, and in some cases controlled for, it becomes possible to design effective analyses and evaluate the results within their proper context.

2.2.3 What Inventive Activity Do Patents Capture?

While all patents nominally represent an invention of some sort, patent rights are not the only way to protect intellectual property. The patenting process is costly, and firms or individuals will likely only patent inventions for which they expect returns greater than these costs [66].

One particular concern for the work contained in this thesis is the fact that differ-

ent industries have different propensities to patent their inventions. There are myriad reasons why firms choose not to patent an invention. While some industries may patent every incremental improvement that is believed to satisfy the patentability requirements, in other industries this kind of patenting is either not possible or not financially viable. In an early work by Mansfield [68], the author surveys 96 companies from 12 different industries in order to study the patterns in their propensities to patent and find significant variation between these industries. It is noted that there are many reasons for this variation, including rapid technological obsolescence rendering long-term patent protection impotent, difficulty identifying or prosecuting potential infringers, or that the subject matter is simply too costly to imitate in any case and so patents offer no additional benefit.

Some inventions are exclusive due to firms merely keeping them secret – the Coca-Cola ‘secret formula’ is the archetype of this kind of intellectual property protection. This type of protection is called a *trade secret* and has been declining in strength in many industries due to the new technology that can, for example, easily determine the structure of a new molecule or semiconductor [68, 69]. In these industries, therefore, patents may have become better indicators of inventive activity in recent decades as a firm’s patents reveal more information about their inventive activity than previously.

Patent statistics completely excludes some industries simply due to their subject matter not being patentable in the United States. For example, the Supreme Court of the United States ruled in 2013 that naturally occurring DNA sequences were not patentable [70].

2.2.4 Bureaucratic Effects on Patent Grants

Over the course of a research program, a firm will seek patent protection for as many aspects of the final product as possible, given the pros and cons of applying for this protection. In this view of invention, a patent can be interpreted as representing a quantum of knowledge, some minimum inventive step that meets the criteria for patentability. This concept is consistent with the above interpretation of patents as a progress marker; for a particular technology a patent represents a specific amount of technological progress - if the inventive step is too small then it may not meet the criteria for novelty, usefulness, or non-obviousness. If it is too large, the inventor may have risked competitors

patenting a similar invention and lost their first-mover advantage.

Of course, the inventive step is an abstract concept, and the minimum step size required for a patent is subjective and dependent on many factors. These include small administrative changes to the USPTO, legal precedent set by prominent court cases, and the establishment of a centralised appellate court. Together, these events contributed to the so-called the ‘patent explosion’ and, concurrently, the perceived relaxing of examination standards during the 1980s and 1990s [71, 72]. This phenomenon exemplifies the complex relationships between policy, administration, and patent output that need to be taken into account in order to compare patents granted at different points in time.

Patent law is both changing continuously with differing political philosophies about the purpose of patent rights [73], and adapting to incorporate or exclude certain types of inventions as the social and technological environments (in which patents are embedded) change. Additionally, the incentives and procedures of the patent office affect the number and quality of the patents that are granted, and both of these change over time as well. As an example of the latter phenomenon, in his 1990 review, Griliches [66] notes that some of the large-scale changes in patenting behaviour as observed in patent grants and applications are “nothing more than bureaucratic mirage”. The author of that work observes that it is statistically and graphically clear that, in the short-term, the number of grants made by the USPTO is more highly correlated with the number of examiners available than the number of applications. These kinds of observations are concerning to both researchers and policy-makers [71], as they force the conclusion that changes in the resources available to the USPTO may have significant effects on the examination times and potentially the quality of the patents that are granted. It is therefore essential that these effects be controlled for in order to place patent grants from different points in time on equal footing.

2.3 Patent Citations

2.3.1 The Citation Process

An essential part of the patent examination process is determining how a new invention is different from those that came before it, that is, establishing novelty. This is a negotiation between the inventor, attorney, and examiner, and exists primarily to delimit

the legal claims that a patent makes. Claims specify the scope of an invention and can be either independent or dependent. Independent claims are the novel aspects of an invention, while dependent claims narrow the scope of the independent claim with which they are associated. It is generally the dependent claims that differentiate inventions from those that came before, so-called *prior art*. Citations to prior art are often used by the examiner to justify additions to the list of dependent claims. These citations will henceforth be referred to as *examiner citations*.

Inventors are responsible for providing their own citations on application, due to the obligation known formally as the ‘Duty of Disclosure, Candor, and Good Faith’, but usually shortened to ‘duty of candor’. The duty of candor requires all applicants to provide the USPTO with all material related to the invention of which the applicant is aware. This not only includes scientific articles or other non-patent material but most importantly, it requires the applicant to disclose relevant patents (including patents granted in other jurisdictions) that the applicant is aware of and that may restrict the scope of the final patent. These kinds of references to this prior art will henceforth be referred to as *inventor citations*.

Inventors and examiners may cite prior art for a number of different reasons. First, citations can directly call into question the novelty of a patent. These citations generally justify narrowing the scope of the prospective patent by, for example, adding additional dependent claims. This avoids any overlap in the technological space (see Section 2.3.2) delimited by the claims of two different patents. Second, the size of the inventive step implied by a particular claim may be challenged by a combination of patents. Taken as a group, this combination may show that an inventive step is not as significant as it may seem when the application is compared to each of these pieces of prior art individually. Lastly, a patent may cite prior art that provides some technical background in order to clarify the size of the inventive step and potentially establish utility, if this is questioned.

Recent work by Cotropia et al. [44] calls into question the importance of inventor citations in testing the validity of patent claims, and therefore the necessity of the Duty of Candour. The authors of that work collected information about the prior art used to justify a narrowing of the scope of 1192 patents in the form of 3358 ‘novelty’ or ‘non-obviousness’ rejections (a patent can be rejected multiple times). The prior art that was used in this justification is then added to the list of citations. Since 2001, the USPTO has made explicit (on the patent document) which citations were made by the inventor

as part of their Duty of Candor and which were added by the examiner. The authors were, therefore, able to identify the origin of the documents used to narrow the patent claims. It was found that 87.2% of rejections were based on examiner-added citations, despite examiner citations only making up 26.5% of all citations in the sample. This result suggests that while inventor citations are generally seen as more likely to indicate knowledge flow [43], they do not seem to play a large part in determining the boundaries of the patent in technology space.

2.3.2 A Geometric Analogy

One way of thinking about the role of patent citations is to first imagine that technology exists in some high dimensional space, and that patents specify a region in which the technology described in the patent resides. These regions are government-granted plots which may not be trespassed without explicit permission from the landowners. Patent citations in combination with the claims work to delimit the technological space that a firm can, if they wish, legally defend from trespass for as long as the patent is in force, after which time the space becomes public. In particular, the citations indicate the locations of nearby properties, while the claims in both the citing and cited patents together establish a reasonably well-defined boundary between two properties. The size of these properties differs widely from patent to patent, and one indicator of this size is by looking at how many neighbours abut the property. While the citations to prior art as displayed on a particular patent, henceforth *backward citations*, are the initial indicators of this size, there are many directions in which the boundaries are not so clear. A better indicator of this size is the number of people who cite the patent in question, henceforth *forward citations*. These are often patents that have had their claims narrowed to prevent overlapping intellectual property. In this view, a patent's true impact may only become apparent when we discover many other patents attempting to encroach on what is likely valuable intellectual property. This situation can often happen when a patent is located on the boundary or frontier of human knowledge, with prior art only limiting the size of this property on one side. This analogy is not perfect by any means; however, it gives a nice geometric analogy to assist in the imagining of the legalities involved in the ownership of intellectual property and the role citations play in this picture.

2.3.3 Citations as Indicators

Patent citations have become very common indicators in a wide variety of economic applications. Not only are citations indicators of relationships between inventions, but they contain information about the relationships between inventors, firms, industrial sectors, and more. This Section summarises the most prominent work done to present in establishing citations as a valid tool for measuring the flow of knowledge, economic value, technological quality, and social impact. Without establishing these relationships, it is impossible to properly evaluate the empirical results contained in this thesis and relate them back to these concepts.

Citations as Indicators of Knowledge Spillovers

One of the first applications of patent citations, primarily due to their economic importance, was the analysis of the extent to which knowledge spillovers occurred; that is, how ideas diffuse between people, firms, and industries. This included a desire to understand how and when the knowledge contained in newly patented inventions is utilised by other, potentially competing, firms [28]. If an inventor utilises the information in another patent, an acknowledgement is usually made in the form of a citation. However, there exists a complicated relationship between citations made and knowledge flow. There are different mechanisms for knowledge transmission (see Section 2.4.1), and these manifest themselves in different ways on the patent document. For example, an examiner citation is often assumed not to be indicative of a knowledge flow at all, but this view ignores the informal ways in which an inventor can learn about aspects of another invention without ever knowing that this information exists in the form of a patent. To this point, it has been argued that one should be comparing the technological content of two patents to establish the existence of knowledge flow in order to capture more informal forms of knowledge transfer that are not reflected in citation data [74].

The first work that links patent citations to localised knowledge spillovers is that of Jaffe et al. [28]. It is well known that industries tend to cluster together, the existence of Silicon Valley being an archetype of this type of firm behaviour. What was not well understood was the extent of any knowledge flows within these localised industries, *given* their geographical distribution. In other words, because industries cluster and citations often flow between firms in the same industry, the question is how much of this citation flow is because these firms are geographically nearby over and above the expected citation

flow due to technological similarity. It was discovered that geographic proximity indeed had a significant effect on firms' citation behaviour and constituted good evidence for citations as indicators of informal knowledge flow. This result provided the first suggestion that citations were useful as indicators of something other than raw technological impact – citations have the potential to be used to track the diffusion of ideas in time and space.

Jaffe et al. [28] has been corroborated in many other contexts, notably by Jaffe & Trajtenberg [75] (international knowledge flows) and Maurseth & Verspagen [76] (European knowledge flows). The latter also provided the opportunity to examine the effect that national borders and language differences have on these localised knowledge flows. Localised geographic spillovers between universities and industry have also been discovered [77].

Citations as Indicators of Economic Value

Patents can be used to measure a component of the economic value of a firm, often referred to as the value of 'knowledge stock'. The idea here is that, just like other investments made by a firm, a firm will generally expect a positive return on their R&D investments. These returns are often difficult to measure precisely, as data on R&D spending are notoriously difficult to collect and compare between firms. This opacity occurs for a couple of reasons: publicly listed firms have different ways of accounting and reporting their R&D spending, and private companies have no obligation to make these numbers public. However, patents offer a way to consistently compare inventive activity and knowledge stock among firms in the same industry. If this information is combined with stock market data, then it is possible to link additional citations to increased firm value [67] and even associate stock market movements to the grants of particular patents [78].

There are also proxies that are often used to estimate the economic value of patents that correlate with forward citations. Two notable proxies that have been shown to correlate with forward citations are litigation events [79] and renewal events [80, 81]. Litigation events are deemed a good proxy for economic value under the assumption that firms do not litigate against worthless patents, as this is an expensive exercise and therefore litigated patents, on average, can be assumed to have a higher economic value

than non-litigated patents¹. Finding a correlation between litigation probability and forward citations would, therefore, provide an indirect link between economic value and technological impact. In one of the first studies to examine this relationship, Lanjouw & Schankerman [79] found that, indeed, litigated patents were significantly more highly cited than non-litigated patents. The authors also observed that litigated patents generally contained more claims than non-litigated ones, and so proceeded to run the same analysis after normalising the number of forward citations by the number of claims. They concluded that even after controlling for the the number of claims of the invention, litigated patents are still cited significantly more than non-litigated patents.

Renewal events, on the other hand, reflect a willingness on the patent owner's part to continue to pay the patent office in order to keep their patent in force. The fee to do this, called maintenance, increases with time – a characteristic that researchers often use to enrich their analyses. Much of this research assumes that firms will renew only when they believe that keeping the patent in force will be more valuable than the price of the maintenance fee, thus providing a measure of minimum patent value as assessed by the firm holding it. The best-known studies that utilise these statistics as a proxy for economic value come from Europe, where the maintenance fees are generally due on a yearly basis until the patent term is complete. This payment scheme allows for much better time resolution when compared to the same obligations in the USA, which occur every four years until the twelfth year. En route to a survey of the private economic value of a patent (see below), Harhoff et al. [80] found that patents that were renewed annually for their full term were more highly cited than those left to lapse prematurely. In a more recent work by Maurseth [81], the author also finds that additional patent citations do indeed correlate with a longer lifetime. However, when these citations are disaggregated by classifying them as cross-class citations or within-class citations, it is discovered that cross-class citations drive this effect and that within-class citations actually negatively correlate with lapse time. This observation may be due to the within-class citations being more likely to represent knowledge spillovers to competing firms, and thus represent creative destruction [83] that leads to the patent losing its value much faster. This particular work provides insight into the complexities and counter-intuitions that are encountered when dealing with data representing the output of a complex economic

¹The last decade or so has, however, seen an increasing number of frivolous lawsuits filed by non-practicing entities who accumulate patents and target small businesses with the hope that the target's resources are not sufficient for full litigation proceedings, and therefore force settlement [82]. This abuse of patent rights likely makes litigation probability less useful as a indicator of patent value.

system and acts as a cautionary tale, particularly when both the independent and dependent variables are proxy indicators for unobservable processes.

The most direct way of obtaining an estimate of patent value is by simply asking the firm or inventor how much they believe the patent is worth. The best-known research of this type was undertaken by Harhoff and collaborators [80, 84]. Harhoff et al. [80] conducted a survey to establish the perceived value of 772 German patents, as determined by the firms that held them. The authors found that not only were these value estimates significantly correlated to renewal probability, but that the high-value patents were also much more likely to be highly cited. In a later work, Harhoff et al. [85] examined this relationship in more detail. It was found that not only were the number of backward citations positively related to patent value, the number of non-patent citations was also positively correlated with value in particular technology fields.

Citations as Indicators of Social Impact and Technological Quality

Throughout this work, it is assumed that citations are a noisy indicator of the quality of an invention, and this assumption is vital to the interpretation of the results of the analyses in this thesis. The meaning and measurement of quality in the context of patent documents is discussed in more detail in Section 6.1.1; there exist many dimensions of patent quality and not all of these dimensions are reflected in citation counts. An important relationship to first establish is the link between patent citations and the social impact of an invention. Impact is necessarily measured in different ways for different types of inventions, and so the best evidence available comes from case studies of particular technologies. The first major work in this literature is due to Trajtenberg [86], in which the author constructed a measure of consumer surplus gains related to particular improvements in Computed Tomography scanning technology that could be linked to particular patents. Consumer surplus gains were measured to be associated with improvements with larger numbers of citations. This finding was among the first indications that citation numbers could be used as a measure of the societal influence of an invention. In a more recent work, Moser et al. [87] directly observe a strong positive relationship between performance (e.g. yield) of patented hybrid corn varieties and the number of citations to those patents. It is also noted that citations that were added by the examiner are not directly related to the performance of the corn and it is suggested that the reason for this is that examiners are not adding citations based on the perfor-

mance of the cited inventions, but rather to establish technological similarity. At least in the case of hybrid corn, citations to patents based purely on their performance are done mostly by the applicants and this observation is consistent with the notion that citations do in fact reflect knowledge flows.

Another line of research investigates the relationship between the intrinsic nature of particular inventions and their subsequent impact. Trajtenberg et al. [88] showed that citation rates to patents assigned to universities were higher than those to private companies' patents. This work was done with the assumption that there was a difference between the natures of these two groups of patents, namely that the patents assigned to the university tended to be somehow more fundamental discoveries rather than incremental improvements on existing technologies, and that patents for these fundamental discoveries are generally more impactful than incremental patents. These assumptions were indeed borne out, and this work became one of the first to present findings that indicate a possible relationship between patent citations and an intrinsic property of the patented invention (in this case, proximity to fundamental scientific research), rather than to external social or economic indicators of impact.

Validation studies utilising expert knowledge have also been conducted. Carpenter et al. [89] leveraged expert knowledge indirectly by comparing the citation rates of patents that underlay products awarded the IR100 award, a list published in the Industrial Research and Development journal designed to highlight the most "important, unique, and useful" products of the previous year and chosen by the Editorial Advisory Board. This board presumably constituted a good representation of a carefully selected panel of experts. The authors found that the patents in the award group were cited more than twice as frequently as a set of control patents, and was an important step in the establishment of patent citation statistics as meaningful indicators of technological impact, or at the very least perceived importance. Albert et al. [30] used a survey of experts in silver halide technology to assess the relationship between expert opinion and citation rates to particular Kodak patents – a strong correlation was found yet again.

2.3.4 Differences Between Academic Citations and Patent Citations

Much of the work in the field of bibliometrics has been conducted using citations to published academic articles. Therefore, it is pertinent to briefly outline the main differences

in citing practice and the meaning of citations in each of these areas.

First, there are very specific circumstances under which a patent must give a citation to prior art, whereas academic citations are not bound by these kinds of regulations. Second, there is a legal meaning inherent in a patent citation that is absent from academic citations. These citations, therefore, have real consequences for the citing patent – they can not only affect the value of the patent, but an improper citation can be legally problematic and may even lead to invalidation. On the other hand, academic citations have few consequences for the citing article – perhaps with the exception of the annoyance of an uncited contemporary. Lastly, patent examiners have a much more significant impact on the composition of the reference list than do academic referees or editors. Examiners have the power to add or remove citations according to their mandate, whereas instances of academic referees or editors doing the same are rare. This examiner autonomy has significant consequences for the meaning of patent citations when compared to academic citations concerning their objectivity.

These differences also manifest themselves in the structure of the citation networks that are the results of the citation process. Some of these structural differences (and similarities) are further explored in Sections 3.3.1 & 5.6.1.

2.3.5 Examiner Effects on Patent Citation

Many citations are added by the patent examiner, and thus understanding the effects that examiners have on the final citations that end up on the patent is vital to the understanding of their meaning. In the United States, these effects are somewhat muted by Duty of Candor. This means there are many more inventor citations than in other large jurisdictions such as the European Patent Office, where no such disclosure obligation was in place until 2011. Additionally, the USPTO started to include data on which party added each citation listed on the patent document from 2001, and this has made possible the analysis of examiner effect on patent citations, with many interesting findings emerging.

A 2002 work by Cockburn et al [90] explored a dataset which incorporated particular examiners' examination histories and litigation probabilities of the examined patents. Their first observation was that primary examiner characteristics are highly heterogeneous – there was substantial variation in their average number of citations added per

patent and their examination speeds; these individual characteristics accounted for 7.7% of the overall variance observed in the numbers of citations made and 8.3% of the approval time for individual patents, after controlling for technology and patent age effects. The authors also found that the average number of citations made per patent had significant positive effects on both the number of citations those patents received and the probability that they would be found to be valid after a litigation event. Lemley et al. [91] found a negative relationship between the experience of an examiner and the number of citations made on their examined patents, while Alcacer et al. [92] found that the fraction of examiner-added citations varied across a number of cross-sections including technological field, applicant location, and applicant portfolio size. While examiner citations are not examined in isolation in any analysis in this thesis, these results point to some potentially fruitful avenues for future research (see Section 7.5).

2.4 Innovation Dynamics

Innovation economics is a rich and varied field and as such attention will now be focused on the facets of invention that are most relevant to this work: knowledge diffusion and technological obsolescence. These phenomena act in opposite directions as time goes on; on the one hand, ideas spread and grow popular, on the other they become irrelevant. We expect both of these effects to manifest themselves in the form of citation dynamics. As an invention gains popularity and enjoys broader impact, citations to the associated patent are expected to accelerate - this effect is what we refer to as knowledge diffusion, and we will pay particular attention to the temporal dimension of this diffusion. At the same time, the invention is becoming ‘old news’; at some point most inventors who can use the idea in their own work have already done so, and in any case the technology is being replaced by something better.

2.4.1 Knowledge Diffusion

The obvious first step to understanding the diffusion of ideas and its relationship to patent citations is to understand the mechanisms that drive diffusion. That is, how exactly is information transmitted between inventors? Only when we have an answer to this question is it possible to begin answering questions about how patent citations relate to these different mechanisms and therefore the types of knowledge diffusion captured

by citations. While this Section will focus on the mechanisms of knowledge diffusion and its relationship to the temporal dimension of patent citations, there are also related spatial patterns [28] that are not empirically examined in this thesis.

As a step toward identifying specific mechanisms of knowledge diffusion, a systematic survey of inventors was conducted by Jaffe et al. [43], wherein the authors asked about 400 inventors various questions regarding either the backward citations from their patent or the forward citations to their patent. The citing inventors were asked about three patents, two of which were cited by the inventor and differed only in their age, and one ‘placebo’ patent that matched the earliest cited patent by age and technology class. This matching not only allowed comparison between cited patents of different ages to observe recency effects, but more importantly, it allowed comparison between two very similar patents, only one of which was actually cited by the inventor. Differences observed in the latter comparison were therefore likely to reflect real information transmission, or knowledge spillovers, encoded in the citations. Many promising results emerged from this data: cited patents were significantly more likely to be remembered as a source of ideas than the placebo, the perceived technological significance of the cited patents was significantly correlated with accrued citations to those patents, and more recent patents were more likely to be remembered. With the stipulation that about one-half of the citations did not correspond to a knowledge spillover, the authors conclude that their results were “consistent with the notion of citations as a noisy signal of the presence of spillovers.” Additionally, the inventors noted that “word of mouth, personal interaction or viewing a presentation or demonstration” was a significant influence in about a quarter of all spillovers, and information contained in “technical or patent literature” in about one fifth. These results indicate that a significant number of spillovers, which constitute a significant number of citations, are due directly to information transmission between inventors.

In Roach & Cohen [93], the authors undertook a survey focused mostly on understanding the specific types of knowledge flow that may *not* be captured by patent citations. The two types of measurement error identified as affecting one’s ability to measure knowledge flows with citations were ‘errors of omission’ and ‘errors of commission’. The former represents knowledge flows that we expect citations to reflect but are not materialised, and the latter represents everything that citations capture that are entirely unrelated to knowledge flows. Errors of omission identified include contract-based linkages between academia and industry, as well as fundamental research that

contributes heavily to a patented invention but is either not patented itself or is simply not patentable material. Errors of commission have been implicitly discussed above and in Section 2.3.1, and include citations resulting from firm strategy and some examiner citations. Both this work and that of Jaffe et al. [43] suggest a degree of caution when using citations as indicators of knowledge flow.

2.4.2 Technological Obsolescence

The growing irrelevance of ideas is perhaps more socially relevant in the case of technological innovations than it is to scientific advances due to the economic consequences of this phenomenon. New innovations that emerge due to technological advancement bring about creative destruction, that is, the displacement of one product by a better one. However, this can be viewed as distinct from so-called *technological obsolescence*, or *knowledge obsolescence*. Cabellero & Jaffe [12] distinguish technological obsolescence from creative destruction by identifying the different mechanisms driving each of these phenomena. Whereas creative destruction can be thought of as new products displacing old products, technological obsolescence is the process of new knowledge displacing old knowledge – this can be thought of as the advancement of the technological frontier whereby old inventions, while they may still be useful, become less and less relevant to the production of new knowledge. While the processes are related, creative destruction is a result of technological obsolescence and not the other way around, and therefore the latter can be seen as the more fundamental process. Patents themselves can be thought of as encapsulations of technical knowledge and do not necessarily contribute directly to the commercialisation of a product. It is more likely, therefore, that citations are more reflective of knowledge obsolescence than creative destruction, as defined.

The structure of the patent citation process means that the slowdown of citations to even highly cited patents is much more universal than for academic publications where some seminal works achieve a kind of immortality [94]. This phenomenon is generally not the case for patents due to an apparent transitivity that exists among the citations of cited and citing patents. What this means in practice is that when patent B cites patent A, most of the restrictions that are placed on A due to its citations are also placed on B [95]. In academia, however, there is no such convention. Papers important to a field will continue to be cited indefinitely, even if every other paper cited also makes the same citation [94]. The removal of these redundant links (e.g. Fig. 3.4), transitive reduc-

tion [95], reveals differences in mesoscale network structure between patent and academic citation networks that is not visible by looking at citation rates – this is discussed in more detail in Section 3.3.1. These kinds of effects suggest that thinking about the accumulation of patent citations in terms of network growth may be a worthwhile endeavour.

Chapter 3

Growing Directed Acyclic Networks

Network science is generally considered to be a new field, accelerated in the late 1990's by new insights made possible by rapidly increasing computational power [1, 96–99]. However, the intersection of networks and scientific citations began to occur in the 1960's with advances in both graph theory [100] and computational techniques [101] alongside a strong desire by some to improve the measurement of scientific output using these new methods [102–106]. Interest in patent citation networks came much later, with a small amount of interest coming in the late 1970's and early 1980's [14, 107, 108]. However, the incorporation of citation networks into more mainstream economics would take longer to occur – as late as 1990, an influential work utilising patent citation data regarded work to date on patent citation networks to be “not quite in the realm of economics proper” [86]. More recent works, including examples of the limited diffusion of citation networks into “economics proper”, are examined in Section 4.1.

Citation networks of all types share special properties which, when leveraged appropriately, simplify the analysis of their structure and evolution. This fact, in combination with knowledge about the institutional restrictions and mechanisms which moderate their growth, in turn perhaps makes both the empirical analyses of particular growth mechanisms and the interpretation of the results of these analyses more straightforward than for other real-world networks. (See e.g. [18, 23, 109, 110] for examples of confounding effects and misuse of network science in other fields). This chapter will first characterise citation accumulation as network growth and introduce the mathematical tools used to analyse the network growth dynamics before covering several generic growth models.

Much of the work on citation dynamics, including this thesis, does not depend on the network framework to function; however, this framework makes the notion of knowledge flows manifesting themselves in the form of citations easier to conceptualise (in the spirit of Price [105]), and more easily allows for extension in future work where specific network structures or characteristics are essential. The term *network* in this thesis is not only used in order to be consistent with the bibliometrics literature, but also as an accurate descriptor of the systems being studied. Whether or not the local properties of the network are examined in particular, the accumulation of citations occurs when new documents join the system *and* citations are made between the new documents and the existing documents. In this sense, citation accumulation *is* network growth by its very nature. The use of the network framework and its associated terminology, therefore, serve additionally as useful reminders that citation accumulation occurs at a microscopic scale despite the mostly macroscopic approach detailed here. For this reason, it is necessary to first briefly review the basic terminology of network science before characterising citation networks – in particular to identify their special properties that allow for considerable simplifications in their analysis.

3.1 Characterisation of Citation Networks

There are many ways to characterise networks based on the nature of the links, the nature of the nodes, and structural properties both local and global. It is therefore important to define some of the network terminology used in this chapter, as there are field differences in the use of many of these terms:

- Network: an ordered pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consisting of a set of nodes $\mathcal{V}$ and a set of links $\mathcal{E}$ representing relationships between pairs of nodes. In graph theory, a network is also called a graph;
- Node: a fundamental unit in a network. In this work, the nodes are patents or scientific articles. In graph theory, nodes are called vertices;
- Link: a relationship between two nodes. In this work, links are citations between patents or scientific articles. In graph theory, links are called edges;
- Degree: the number of links to which a node is incident.

Citation networks are special in that there are some restrictions on the way the network can grow, and these lead to simplifications which make it easier to explore both their theoretical and empirical properties. In particular, citation networks are (generally) unipartite, directed, acyclic networks. These properties are defined as follows:

- Unipartite networks are those whose nodes are all of the same type and links are generally possible between any two nodes in the system. Contrast these with, for example, the bipartite inventor-patent network where links represent inventorship and thus can only form between the nodes of different types, that is, inventors and patents. Unipartite networks are simpler in most aspects, however ‘projections’ of a bipartite network are possible [111], which in the above example would yield two unipartite networks, one of which would be a co-invention network. Unipartite networks are sometimes referred to as one-mode networks;
- Directed networks have directed links, which means each link implies a one-way relationship (though a pair of nodes may have multiple links between them, pointing in any direction). In the context of citation networks, this allows the distinction between the citing document and the cited document. These types of networks split degree into two distinct types: in-degree (links pointing *to* a node) and out-degree (links pointing *from* a node);
- Acyclic networks do not have cycles - following directed links through an acyclic network will never result in a node being visited more than once. In the context of a directed network, this implies that it is not possible to draw a path through a series of nodes (in the direction dictated by the link directions) and visit the same node twice. This kind of network is only possible when there exists a topological ordering [112], a linear ordering after which the links all point in the same direction. For patents, the mechanism responsible for this ordering is time, whereby a granted patent may only cite patents that came earlier.

The properties above describe an idealised citation network. However, under certain conditions, the acyclic nature of citation networks may be broken. The simplest way this can occur is by the citation, intentional or not, of an unpublished document. While this does not occur between two patents granted in the same jurisdiction (except for errors), it is common practice in some fields to cite a forthcoming work, usually by the same author. This event is inevitably followed by a citation in the forthcoming work to the original work, thereby creating a cycle between the articles. Empirically, this issue can be solved by removing citations directed to a forthcoming article, and in any case,

many journals disallow this type of citation.

A more complicated situation arises when one considers citations between patents in the context of knowledge flows. It is possible to choose the particular events that characterise the citation lag between cited and citing patents. For example, the granting of patent A may inspire others to incorporate this work into their own research, which they then submit to the same patent office as an application for a patent of their own, patent B. The lag between the initial information transfer and the first indication that an inventor has utilised this knowledge would correspond to the time between the grant of patent A and the application of patent B. A complication arises when the knowledge contained in the patent is transferred between inventors via some other means, and the inventor of patent B learns of, and utilises, the information in patent A before it is granted (via word of mouth, pre-grant application publications, or other means). Additionally, examiners may add their own citations to patent B that don't correspond to any knowledge flow, and the same situation will arise. These citations will lead to the situation where the application of patent B occurs before the grant of patent A, which creates a citation that points forward in time for this definition of lag. In this work, these negative-lag citations are aggregated to form an initial citation distribution.

3.2 Models of Citation Network Growth

There are some mathematically convenient aspects of citation networks that not only make the construction of their base models much more straightforward but also simplify augmentations to these models. This Section will cover the basic models of citation network growth, and the *master equation* approach that will be utilised to derive the macroscopic network properties from specific microscopic network growth mechanisms.

3.2.1 A Simplified Measurement Framework

The models and analytics that follow are based on a measurement framework that does not attempt to examine the network as a whole. Due to the changeable nature of patent law, economic environments, firm incentives, and other extrinsic and intrinsic factors that change patenting behaviour with time, it is essential to select a cohort of patents that were granted at about the same time to facilitate assessment of the differences in

citation accumulation between different patents fairly. Therefore, while these patents are embedded in a growing network, the properties of the individual patents that cite a patent in the cohort are not taken into consideration. This distinction is important, as much of the existing literature (and theoretical work in particular) considers the entire network, an approach which may complicate the measurement of particular dynamic network properties.

The results that follow are derived such that they accommodate any fixed subset of nodes embedded in a vast, growing network. This measure is to ensure that our theoretical results are compatible with important practical considerations concerning the real patenting (and citing) behaviours examined later in this thesis, while also simplifying the measurement of relevant network properties. While citation networks have directed links, we can apply all analytic results below to a fixed subset of any growing network with equivalent growth rules – outgoing directed links (backward citations) from the nodes in the cohort are generally not relevant to network growth dynamics, and as such we can treat incoming links as undirected without additional considerations. To retain some semantic generality in the current chapter, the in-degree of a document is henceforth shortened to *degree*, which in later chapters is simply equal to the number of forward citations.

3.2.2 The Master Equation Approach

The master equation is a very general mathematical tool that can characterise the evolution of a system wherein each component occupies a particular state, and may move from one state to another with a particular rate. Generally, a master equation takes the form of a differential equation that can be solved to provide a probabilistic description of the configuration of the system as a function of time. This Section provides an overview of this mathematical framework that both facilitates comparison to existing network-science literature on network growth, and provides a means to check the accuracy of the mathematical assumptions and functional forms of the citation accrual model presented in this thesis (see Sections 4.1.3 and 4.4). A very general version of the master equation is presented initially, which is refined (and consequently simplified) as more citation-network-specific properties are incorporated. The resulting analytic form of the degree distribution may then be used to calculate many global properties of the network, including the average rates of citation accrual and the average degrees of nodes

in the citation network, as well as allowing an easy comparison of the dynamics of these properties across networks with different growth mechanisms.

To establish the general master equation, we first define the time-dependent transition matrix $W(t)$, whose elements $W_{ji}(t)$ represent the probability of transition (per unit time) to k_j from state k_i , which are elements of a set of m discrete states, denoted K . We can now describe the time evolution of the system as the rate of change in the probability $n(k_i, t)$ that a randomly chosen element of the system occupies state k_i at time t

$$\frac{dn(k_i, t)}{dt} = \sum_{j=1}^m W_{ij}(t)n(k_j, t) - W_{ji}(t)n(k_i, t) \quad . \quad (3.1)$$

Eq. (3.1) simply constitutes a gain-loss equation; the change in the probability of a component being in state k_i is the number of components transitioning *to* k_i from other states k_j (gain) minus the number of components transitioning *from* k_i into other states k_j (loss).

From this point, it is possible to place restrictions on this master equation that reflect the fact that the system in question is a growing citation network and k_i represents the number of citations accrued by node i (its *in-degree*):

- Nodes attract links as part of a continuous-time stochastic point process;
- Nodes cannot receive multiple new links simultaneously (simultaneous citations, where two citations occur on the same day to the same document but originate from different documents, can be considered an artifact of finite time resolution);
- Nodes cannot have a negative number of links;
- Nodes cannot decrease their in-degree, that is, lose links.

The consequences of these restrictions are that the only transitions possible in this system are $k_i - 1 \rightarrow k_i$ for the gain component of the master equation and $k_i \rightarrow k_i + 1$ for the loss component, where the state space K now corresponds to the natural numbers plus zero. That is, for a given state k_i , there is only one state to transition from and one state to transition to. Therefore, $W(t)$ can now be simplified to a rate function $\lambda(k, t)$ which corresponds to a single column in the transition matrix $W(t)$ with $k_i \rightarrow k$. This rate function incorporates an initial assumption that the probability that a node will

attract a new link in the next time window is a function only of the node's age, t , and the current number of accrued links by this age, $k(t)$. As there is only one state from which it is possible to move into state k , and only one state to move to, the sum in Eq. (3.1) reduces to

$$\frac{dn(k, t)}{dt} = \lambda(k-1, t)n(k-1, t) - \lambda(k, t)n(k, t) . \quad (3.2)$$

Additionally, a boundary condition is put in place to reflect the above criterion that a node cannot have a negative number of links. We can express this condition as

$$\frac{dn(0, t)}{dt} = -\lambda(0, t)n(0, t) \quad , \quad (3.3)$$

equivalent to setting $\lambda(k, t) = 0$ for $k < 0$. Thus, for any rate function $\lambda(k, t)$ it is possible to construct a pair of differential equations which may be solved to determine the distribution over the states in K for any time $t > 0$ – the degree distribution $n(k, t)$ [113, 114].

Empirical evidence (as seen in Chapters 4 & 5) suggests that the rate may be decomposed into two mutually independent functions $f(k)$ and $A(t)$ such that $\lambda(k, t) = f(k)A(t)$. This greatly simplifies the process of finding a general solution to Eq. (3.2).

Eq. (3.2) can be solved generally in a straightforward fashion via variation of parameters followed by induction. Solving the homogeneous equation

$$\frac{dn(k, t)}{dt} = -\lambda(k, t)n(k, t) \quad (3.4)$$

yields the complementary solution

$$n_c(k, t) = n_0(k) \exp \left[- \int_0^t \lambda(k, t') dt' \right] \quad , \quad (3.5)$$

where $n_0(k)$ is the proportion of nodes with degree k at $t = 0$. The particular solution is then simply

$$n_p(k, t) = C(t) \exp \left[- \int_0^t \lambda(k, t') dt' \right] \quad , \quad (3.6)$$

where $C(t)$ is an unknown function which has absorbed $n_0(k)$. The full solution is then $n(k, t) = n_c(k, t) + n_p(k, t)$, which is substituted into Eq. (3.2) to obtain

$$n(k, t) = n_0(k)[\gamma(t)]^{f(k)} + [\gamma(t)]^{f(k)} \int_0^t \lambda(k-1, t') n(k-1, t') [\gamma(t')]^{-f(k)} dt' \quad (3.7)$$

with $\lambda(k, t) = f(k)A(t)$

$$\text{and } \gamma(t) = \exp\left[-\int_0^t A(t') dt'\right] \quad .$$

The boundary condition, Eq. (3.3), requires that $\lambda(k-1, t) = 0$ for $k < 1$ which leads directly to the solution for $k = 0$,

$$n(0, t) = n_0(0)[\gamma(t)]^{f(0)} \quad . \quad (3.8)$$

The solutions for $k > 0$ are then obtained in an iterative manner, where the solution for $n(k, t)$ is found by substituting the solution for $n(k-1, t)$ into the integral in Eq. (3.7) and solving the resultant integral. The results for $k = 1$ and $k = 2$ are therefore

$$\begin{aligned} n(1, t) &= n_0(1)[\gamma(t)]^{f(1)} + \\ & n_0(0)f(0)\left\{\frac{[\gamma(t)]^{f(0)}}{f(1)-f(0)} + \frac{[\gamma(t)]^{f(1)}}{f(0)-f(1)}\right\} \quad \text{and} \quad (3.9) \\ n(2, t) &= n_0(2)[\gamma(t)]^{f(2)} + \\ & n_0(1)f(1)\left\{\frac{[\gamma(t)]^{f(1)}}{f(2)-f(1)} + \frac{[\gamma(t)]^{f(2)}}{f(1)-f(2)}\right\} + \\ & n_0(0)f(0)f(1)\left\{\frac{[\gamma(t)]^{f(0)}}{[f(0)-f(1)][f(0)-f(2)]} + \frac{[\gamma(t)]^{f(1)}}{[f(1)-f(0)][f(1)-f(2)]} + \right. \\ & \left. \frac{[\gamma(t)]^{f(2)}}{[f(2)-f(0)][f(2)-f(1)]}\right\} \quad . \quad (3.10) \end{aligned}$$

This pattern, and the final degree distribution $n(k, t)$, can then be succinctly summarised as

$$n(k, t) = n_0(k) [\gamma(t)]^{f(k)} + \sum_{l=0}^{k-1} n_0(l) \left(\prod_{m=l}^{k-1} f(m) \right) \sum_{q=l}^k \frac{[\gamma(t)]^{f(q)}}{\prod_{\substack{m=l \\ m \neq q}}^k [f(m) - f(q)]} \quad . \quad (3.11)$$

Eq. (3.11) is exact for any separable rate function $\lambda(k, t) = f(k)A(t)$ with any initial degree distribution $n_0(k)$ for which the following conditions are true:

- The integral $\gamma(t) = \exp[-\int_0^t A(t')dt']$ is finite;
- The degree-dependent component $f(k)$ is not constant (e.g., Eq. (3.11) cannot be applied to the random attachment example below);
- Time t is continuous;
- Degree k is discrete, integer-valued, and non-negative.

The degree distribution $n(k, t)$ is also useful in deriving other basic properties of the growing network as they pertain to the focal cohort. These include the average degree of nodes in the network, $K(t) = \sum_k n(k, t)k$, and the average link accumulation rate, $\Lambda(t) = \sum_k n(k, t)\lambda(k, t) = dK(t)/dt$.

3.2.3 Random Attachment

New publications join the citation network by citing earlier works. This process results in a directed, acyclic network as described above. Early models of bibliometric network growth asked the question “how do new publications distribute their citations?” [105]. After all, the number of backward citations on a patent or article is much smaller than the total number of existing documents. The above question leads us to a class of network models known as *generative models*, as they describe mechanistically how a network is constructed from the ground up. By starting with a generative model, the properties of a network can be analytically derived (such as the degree distribution $n(k, t)$) directly from the particular mechanisms defined in the model (e.g., $f(k)$ and $A(t)$).

The first step when faced with a problem such as this is to derive a null model that corresponds to new nodes linking to existing nodes with complete disregard to their properties, a purely random growth model. If the network properties are significantly different from this random model, then the network likely grows in a not-entirely-random way. The most well-known version of the random growth model is that of Erdős & Rényi (1959) [115]. This model sets the number of nodes at N , then adds links between pairs of nodes at random. The resultant degree distribution is binomial, however this distribution is very well approximated by a Poisson distribution which is independent of network size N and has average degree $K(t)$ as its only parameter if the network is *sparse* at a given point in time – a network is sparse when $K(t) \ll N$, and is characteristic of most real networks. The Poisson distribution is mathematically simpler to work with than the binomial and is generally used as a reference distribution to compare the various

properties of other networks to a truly random one.

The degree distribution $n(k, t)$ within a fixed cohort for the random attachment model can be calculated from Eqs. (3.2) & (3.3) by setting the rate to a constant $\lambda(k, t) = r$ such that

$$\frac{dn(k, t)}{dt} = r n(k-1, t) - r n(k, t) \quad (3.12)$$

and

$$\frac{dn(0, t)}{dt} = -r n(0, t) \quad . \quad (3.13)$$

Eq. (3.13) leads to the probability of a node having zero links after some elapsed time t being the simple exponential decay

$$n(0, t) = n_0(0) e^{-rt} \quad , \quad (3.14)$$

which we can substitute into Eq. (3.12) with $k = 1$ and solve by the method of variation of parameters to obtain

$$n(1, t) = n_0(1) e^{-rt} + n_0(0) r t e^{-rt} \quad . \quad (3.15)$$

This process can be iterated for $k \rightarrow \infty$ to obtain the degree distribution

$$n(k, t) = \sum_{l=0}^k n_0(k-l) \frac{(rt)^l e^{-rt}}{l!} \quad , \quad (3.16)$$

which for the initial distribution $n_0(k) = \delta(k)$ (all nodes having $k = 0$ at $t = 0$) gives the Poisson distribution for a given time t ,

$$n(k, t) = \frac{(rt)^k e^{-rt}}{k!} \quad . \quad (3.17)$$

Eq. (3.17) is equivalent to the probability of a random variable n having value k after time t in a Poisson counting process with rate r , which is consistent with the Erdős-Rényi model [115].

3.2.4 Preferential Attachment

Intuitively, random attachment does not appear to be a realistic model for citation networks, as there are many publication attributes that may change the likelihood of attracting a link from an incoming node. The first, and perhaps the strongest, effect to be noticed and incorporated into bibliographic models is the *cumulative advantage* or *preferential attachment* phenomenon introduced by Price [116] which significantly alters the growth dynamics of the network (Fig. 3.1). This effect is the bibliometric analogy of the *rich-gets-richer* or *success-breeds-success* effect observed in other fields [117, 118]. In the context of a growing citation network, the preferential attachment mechanism reflects the increased probability of a new publication citing an older one which already has many citations. On its face, this is a perfectly reasonable thing to expect; not all publications are made equal, and a meritocratic view of the bibliometric world would suggest that those of high quality and high relevance to the research community should rightly obtain more citations than lower quality works. This notion and its validity is the subject of Chapter 6.

The model of Price [116] assumes a preferential attachment that is a linear function of accrued citations k with an initial citation rate equal to unity, that is,

$$f(k) = k + 1 \quad . \quad (3.18)$$

To obtain the (time-varying) degree distribution that results from such a preferential attachment function, Eq. (3.18) is substituted into the degree distribution expressed in Eq. (3.11) with no ageing ($A(t) = 1/\tau_0 = \text{constant}$). It is straightforward to demonstrate that for the initial condition $n(0, 0) = 1$ (all nodes start with $k = 0$),¹

$$n(k, t) = e^{-t/\tau_0} (1 - e^{-t/\tau_0})^k \quad . \quad (3.19)$$

In the standard preferential attachment model described above, all nodes have the same underlying quality, and as a result, the nodes with the greatest number of links are generally among the first nodes added to the system. This phenomenon is known

¹All analytic results derived in this chapter pertain to a set of focal nodes, and assume that the addition of new nodes does not affect the link accumulation rate to these focal nodes, that is, there is no *structural* ageing built into these results. This choice introduces mathematical differences between the results in this thesis and those in the early network science literature (e.g. [119]) – in particular this formulation lacks a stationary distribution as $t \rightarrow \infty$. Structural ageing is discussed in more depth in Section 3.2.5.

formally as the first-mover advantage, and this property alone makes preferential attachment as a generative model inappropriate for the modelling of many real network growth phenomena. However, it has remained popular not only due to its convenience as a starting point for more complex models but also because the analytic predictions made as a result of assuming this underlying generative mechanism appear to fit many types of real-world data. This empirical agreement is in part because it is challenging, and in some cases impossible, to distinguish between the properties of a growing network generated by a PA mechanism and those generated by a model with other linking mechanisms [1, 120, 121]. The uncertainty surrounding microscopic mechanisms in the growth of real networks has led to a debate around the drivers of preferential attachment-like results and the role of luck and quality, the influence of which means that preferential attachment may be treated as an entirely emergent phenomenon in some real networks [19, 122–124].

Preferential attachment in citations to scientific articles is, however, a plausible generative mechanism in an environment where highly cited papers gain more visibility due to those citations (and new citations are not intensely scrutinised). That is, with current bibliographic search tools it is possible to explicitly search for the most cited papers in a subject area under the assumption that the papers discovered represent the most influential and important papers. While this assumption may be valid in some respects, there is also a case to be made that this mechanism constitutes a reversal of the presumed causality: the highly cited papers become ‘important’ due to their citations, whereas many would like to think papers are highly cited due to their importance. In this way, a relatively narrow quality gap between two papers published at the same time in the same field can become a considerable citation gap that exaggerates the quality difference. This distortion has been observed, for example, in the World Wide Web [125]. This phenomenon in the context of patents is explored in detail in Chapter 6.

Figure 3.1: A comparison of the first three models presented in this chapter: **Random Attachment** (solid), **Price model** (dashed), and **Price model with ageing** (dot-dashed). **Top:** Degree distributions at times $t/\tau_0 = 1, 2, 3$ (left to right). **Bottom:** Time evolution of the fraction of nodes with degree $k = 0, 1, 5$. Random attachment produces narrow degree distributions, while the Price model leads to a very fat-tailed distribution without the dampening effect of node ageing. The form of node ageing used here is expressed in Eq. (3.20).

3.2.5 Node Ageing

Knowledge generally doesn't stay relevant indefinitely; it becomes obsolete as it is improved on or entirely replaced by more recent advances. As citations can be thought of as indicators of knowledge flow [28], this phenomenon is reflected in the slowing of the citation rate to documents as they age, and acts in opposition to the preferential attachment effect (which unabated leads to explosive growth in accrued citations). In network growth models, this propensity for information to become irrelevant is often encoded as a multiplicative rate modifier [34, 35, 126]. This node ageing significantly changes the growth dynamics displayed in generative models (Fig. 3.1), as it mutes or entirely negates the first-mover advantage, and the extent to which this occurs depends on the functional form of the ageing function. An example of this effect is illustrated

in Dorogovstev et al. (2000) [35], where the authors find that the degree distribution narrows significantly (with preferential attachment proportional to k), becoming exponential under certain conditions.

Depending on the generative model (analytically) or real growth mechanism (empirically), some of the observed ageing may be due to *structural* ageing. That is, as the network grows, more nodes can be cited, and thus the probability of a particular node gaining a link decreases with time. In random generative models such as preferential attachment with ageing [35], the ageing of nodes is, therefore, the combination of structural ageing and intrinsic ageing due to an explicit ageing function. In real networks such as the patent citation network, these two types of ageing both have a common cause – the addition of a new node both increases the network size and contains information that is more cutting-edge and relevant to future patents than those that came before it. However, due to the legal meaning and strict moderation of citations in this context (and the resultant transitivity), the hypothesis that the increasing size of the network has an impact on the probability of accruing new citations does not seem likely. If this proposition is accepted, therefore, then technological irrelevance and the generally technologically transitive nature of patent citations are assumed sufficient to explain any observed ageing, with the corollary that structural ageing as defined above is assumed to be negligible while these other mechanisms are at play.

We can explore the effect of node ageing on a growing network by choosing a specific ageing function $A(t)$. It is straightforward to show that for a Price [116] preferential attachment mechanism $f(k) = k + 1$ where all nodes start at $k = 0$, the degree distribution is

$$n(k, t) = \gamma(t)[1 - \gamma(t)]^k \quad (3.20)$$

with $\gamma(t) = \exp[-\int_0^t A(t')dt']$ as usual. For an ageing function that is a simple exponential decay with characteristic time τ , $A(t) = e^{-t/\tau}$,

$$\gamma(t) = \exp[-\tau(1 - e^{-t/\tau})] \quad . \quad (3.21)$$

Fig. 3.2 displays the drastic differences in the growth dynamics caused by a change in the characteristic node ageing time τ – the degree distribution is much steeper for small τ , while the differences in the average citation rate and average in-degree observed when τ is changed are much larger than proportional.

Figure 3.2: Degree distributions and dynamics of the Price model with ageing (Eqs. (3.20) & (3.21) respectively) for $\tau = \tau_0$ (solid), $2\tau_0$ (dashed), $3\tau_0$ (dot-dashed). **Top-left:** Degree distributions at time $t = 3\tau_0$ on a log-linear scale. **Top-right:** Time evolution of the fraction of nodes with $k = 5$. **Bottom-left:** Time evolution of $\Lambda(t)$, the average link accumulation rate. **Bottom-right:** Time evolution of $K(t)$, the average in-degree of nodes.

3.2.6 Non-linear Preferential Attachment with Node Ageing

The models above are necessarily simple in order to obtain closed-form analytic expressions for $n(k, t)$ for all possible k and t . In reality, the preferential attachment function $f(k)$ may not be so simple and has numerous refinements that attempt to capture more complex behaviours. In this thesis, the form used is $f(k) = k^\alpha + f_0$. The reasons for this particular form and its advantages over other common forms is primarily of empirical concern, and is discussed in Section 4.5.4.

While a closed form solution for the degree distribution may not exist for $\alpha \neq 1$, the asymptotic behaviour of this non-linear preferential attachment model was explored very soon after the original Price-Barabasi-Albert model [1] was suggested [113, 114]. While

these models did not consider node ageing, the authors found that the shape of the degree distribution depended critically on the exponent α in the long-time limit [113]. For $\alpha < 1$, the degree distribution tends towards a stretched exponential distribution, and for $\alpha > 1$ a single node begins to dominate. In fact, for $\alpha > 2$, there is a non-zero probability that one node is connected to every other node in the network. Additionally, in the linear case with $\alpha = 1$, the initial link attraction probability (proportional to f_0) can tune the exponent on the resultant power law distribution to any value from 2 to ∞ [114]. In short, the degree distribution is very sensitive to the value of the parameters in the generative model and care should be taken when interpreting the degree distribution of any network which exhibits a success-breeds-success mechanism. Fig. 3.3 shows what is observed in networks grown with different preferential attachment strengths. While the results above [113, 114] pertain to the long time limit with no ageing (which significantly changes the growth dynamics [127]), Fig. 3.3 shows that even for short times there are significant differences in network growth dynamics for different values of α .

3.3 Structural Concerns and Alternative Models

It is possible to model many of the features of citation networks with the simple combination of preferential attachment and ageing, as empirically explored in Chapters 4 and 5. However, as a generative model, preferential attachment also results in features that are not seen in real citation networks and can disguise essential differences in network structure between different types of citation networks (e.g. patents, scientific articles, or case law) [95]. Additionally, there are more plausible (but much more analytically complex) generative mechanisms that are not explicitly covered in the remaining chapters of this thesis (e.g. [123, 128–131]). Structural properties uncovered by transitive reduction, as well as an alternative generative model, are explored here as examples of the limitations of the analytic framework described above.

3.3.1 Transitive Reduction

In Section 5.6.1, some significant differences in the meaning of citations to scientific and patent documents are outlined. Naturally, these differences in citation practice lead to differences in network structure. One of the more striking examples of this difference is

Figure 3.3: Degree distributions and dynamics of the non-linear preferential attachment model $f(k)$ with ageing ($\tau = 2\tau_0$) for $\alpha = 0.7$ (solid), 1 (dashed), 1.3 (dot-dashed). **Top-left:** Degree distributions at time $t = 3\tau_0$ on a log-linear scale. Note that changing α only changes the dynamics for nodes with $k > 1$. **Top-right:** Time evolution of the fraction of nodes with $k = 5$. **Bottom-left:** Time evolution of $\Lambda(t)$, the average link accumulation rate. **Bottom-right:** Time evolution of $K(t)$, the average number of links per node.

discovered upon a transitive reduction of the networks [95, 132].

Transitive reduction (TR) is a process that iteratively removes links in a directed acyclic network such that a path between every pair of nodes that one could travel between in the original directed network still exists. The process continues to remove links until there is only one way to travel between each of these pairs. Fig. 3.4 shows an example of the result of this process. If patent citations did indeed implicitly carry over from cited to citing patent, then TR would not remove any links at all. In reality, TR removes about 15% of links. However, when TR is applied to the article citation network, about 80% of links are removed. This disparity can in part be explained by the convention to cite the influential papers in a field as an acknowledgement to those who paved the way

Figure 3.4: An directed acyclic network **(a)** before and **(b)** after transitive reduction.

for the citing paper, where it does not matter that many of the other citations made in the citing paper also cite these same influential papers. Indeed, preferential attachment combined with the triangle-forming behaviour of scientific citations calls into question the real influence of highly-cited papers. This result also suggests that the two citation networks have very different local growth mechanisms.

3.3.2 Copying Models

A common criticism of preferential attachment as a generative model is that it requires, in its pure form, knowledge of the degrees of every node in the network to calculate the probability of linking to a particular node. Variants of a *copying model* can reproduce preferential attachment-like global network properties with only local knowledge. This popular model was first introduced in Kleinberg et al. (1999) [133] to replicate some of the network features observed in the World Wide Web, and has been mathematically formalised and modified by many authors since [20, 122, 134–138].

The generative algorithm for this type of network growth is generally some variation of the following: an incoming node randomly selects an existing target node, then forms a link with this target node *and*, with some probability, its neighbours. This process is random and can, therefore, produce a wide range of directed acyclic graphs. The degree distributions that result from this kind of growth are often power laws (e.g. [122, 129]),

as this one of the features the inventors of these models are often attempting to reproduce to be consistent with empirical observations.

A feature that is often exhibited in scientific citation networks, but is not well modelled with a preferential attachment generative mechanism [139], is the large number of triangles [140]. One of the primary features of the copying mechanism is that it forms lots of triangles (by construction) – which can be used to measure the level of clustering in a network [96, 141]. These triangles are dismantled under transitive reduction, and so the differences in the results of transitive reduction between patent and paper citation networks is a direct reflection of the clustering [142] present in each of these networks. This fact suggests that while a copying model may be suitable as a generative model for scientific citation networks, it is likely a poor representation of network growth in patent citation networks.

3.4 A Useful Illusion

Throughout this chapter, we have focused on obtaining a dynamic degree distribution from microscopic attachment rules. While the degree distribution and the other global quantities derived from it are important, the results above clearly demonstrate that many local quantities may contain more relevant information about how a network grows.

From an empirical standpoint, however, preferential attachment remains the most parsimonious model to test and compare the growth of different networks. The fact that the *real* network growth mechanisms may be very different to preferential attachment does not diminish its usefulness in this context. These real growth mechanisms are myriad and complex – what is important is that they all appear to sum to a success-breeds-success effect.

This thesis asks readers to think of preferential attachment as a proxy for the aggregate effect of all the real mechanisms that add up to the success-breeds-success phenomena observed in citation networks, and not as a generative model describing real citation behaviour. It is true that preferential attachment has many shortcomings, but these are more than balanced by the simplicity and parsimony of this model. Chapters 4 and 5 demonstrate that preferential attachment adequately captures the success-breeds-success phenomena observed in two different citation networks and that the strength of

this effect is generally unaffected by the passage of time. Chapter 6 takes a step towards answering the question “how does a patent become successful in the first place?”. This question goes beyond asking for a generative model that attempts to capture specific features of the network. After all, we do not cite randomly; we cite research that is important, useful, or merely similar to the research we are doing. Perhaps we go out of our way to cite our friends or, more likely, ourselves. If we are applying for a patent, we may not want to cite anyone if we can get away with it. Citation behaviour is complex and context-dependent; it is tied up in unwritten rules, search behaviours, social networks, economic incentives, and laws. What is known is that, somehow, certain ideas and inventions end up with more influence than others. The remainder of this thesis is dedicated to adding to the current understanding about how and why this happens – first by characterising the growth of this influence.

Chapter 4

The Dynamics of Patent Citations

4.1 Patent Citation Dynamics

4.1.1 Perspectives from Economics

In the economics literature, the analysis of patent citation dynamics has mostly been based on the framework laid out in Caballero & Jaffe [12], and twice extended by Jaffe & Trajtenberg [75, 77]. The underlying assumptions behind this model are that there are two competing (and generally non-interacting) effects that govern the dynamics of the probability of a citation between two patents: technological obsolescence and the diffusion of knowledge. The former accounts for the tendency of inventions to be supplanted or become irrelevant and the latter reflects the propensity for ideas to spread, with more and more inventors building on or around a technology as the idea becomes more well known. These effects are usually modelled as an exponential decay and an exponential saturation function respectively (as in Eq. (4.1)), together these are commonly referred to as a ‘double exponential’ in the economics literature.

This model is most explicitly specified in Jaffe & Trajtenberg (1999) [75], which assumes the probability $p(k, K)$ of patent K , granted in year T , citing a particular older patent k , granted in year t , is

$$p(k, K) = [1 + \gamma D(k, K)] \alpha(k, K) \exp^{-\beta_1(k, K)(T-t)} [1 - \exp^{-\beta_2(T-t)}] \quad . \quad (4.1)$$

Eq. (4.1) takes into account the technological distance between the patents and its importance ($D(k, K)$ and γ respectively), the rate of technological obsolescence (β_1 , with

potential dependencies on patent attributes), and the rate of diffusion (β_2). The function $\alpha(k, K)$ captures any other patent attributes that may affect the probability of citation between K and k , including the location of their assignee, grant years, and technological categories. In Jaffe & Trajtenberg (1999) [75], these last two attributes act as controls for variable grant and citation rates at different times and across different technologies, while the first is a variable of interest as it provides information about the magnitude of the international knowledge flows after which the paper is named.

This general model gained much traction in the economics literature due to its applicability and flexibility in form, and has mostly been applied to problems mirroring its original context - studying technological knowledge spillovers (i.e. diffusion) across geographical or institutional borders using patent data [143–150] (however it is not unheard of to apply the model to other knowledge-diffusion-related phenomena [151]).

Other approaches taken to understand patent citation dynamics include hazard estimation [152] and diffusion of innovation-inspired models¹ such as the Bass model [47, 153–155]. While these models have produced useful results and are able to model particular aspects of technological diffusion [154, 155] as well as abstract properties of the inventions that are being diffused [152], they are relatively narrow in their applicability and therefore lack the flexibility of the Cabellero & Jaffe [12] model.

The words flow and diffusion in physics generally imply some dynamic process reflecting movement or exchange of some unit of interest. However, except for the models listed above, the dynamic aspect is often ignored in research studying the ‘diffusion of knowledge’. Diffusion of knowledge can, for example, refer to the mechanisms by which knowledge is transferred [156], wherein dynamics is either difficult to measure or merely irrelevant. In light of this difference in language use, it is important to note that most work in the field of knowledge flow, diffusion, or transfer is, while informative, generally unrelated to the research detailed in this thesis.

In summary, there are surprisingly few empirically-tested models of knowledge diffusion and technological obsolescence in the economics literature. The network-growth framework appears to be a more all-encompassing mathematical framework and can

¹In the context of diffusion of innovation, *innovations* generally refer to the implementation of inventions or ideas. *Inventions* are not always commercialised, or even utilised.

facilitate research not only on growth dynamics [157], but other important aspects of the invention ecosystem such as local structure of technological or industrial networks (e.g. clustering) [158, 159], patent heterogeneity (Chapter 6), collaboration between inventors and between firms [160], etc. These properties are generally considered to be related to each other, and the network-based framework makes it possible to start to understand the interactions between the many different facets of invention as different interconnected layers of a larger system (e.g. [161, 162]).

4.1.2 Applications of Network Science

The network-growth framework in the context of knowledge flows and citation dynamics has historically been applied to scientific citation networks, for the reasons discussed in Section 5.1. While many of the modelling techniques used in that field are appropriate for the analysis of patent networks, both the method of application of these techniques and the interpretation of any results must be tailored to the context of patents in order to obtain useful information from them. It is the lack of this consideration that most often obfuscates research in this new field, although this has also been framed as a difference in scientific objectives between the economists and network scientists studying knowledge diffusion [25].

As an example of these differences in scientific objectives, we will examine the definition of the citation lag. This concept is discussed in more depth in Section 4.3.1, however, the citation lag provides a useful illustration of the different approaches used by economists and network scientists in the context of citation dynamics.

A citation is generally regarded as a good proxy for knowledge flow [66]. However, the timing of the citation is dependent on a number of factors and requires a choice of definition. There are a few definitions to choose from, which are either time-based (based on the application or grant dates) or sequence-based (based on the order of applications or grants). The time-based lags are simply measured as the amount of calendar time that passes between the application/grant dates of the cited and citing patents. Generally, these definitions are one of two:

- The time between the grant dates of the cited and citing patent; or
- The time between the grant date of the cited patent and the *application* date of the citing patent.

While both of these definitions are found in the economics literature [163], the first incorporates the period between the application and grant of the citing patent - an additional source of (mostly) random noise with regard to knowledge diffusion [164]. Complementary to this fact, the second definition more accurately reflects the time from an inventor learning about a new invention to the time that this information is acted on and a related invention is completed. For this reason, the second definition is generally preferred when attempting to model knowledge diffusion [163].

In the network literature, however, there is a long tradition of growing a network one node at a time, with time parametrised by the number of nodes added to the network thus far. Because each patent is given a unique number in a monotonically increasing sequence, patent numbers offer a convenient way to implement models of this type [37, 165, 166]. While useful for comparing the patent citation network to other growing networks, time loses most of its meaning and makes it difficult to compare or integrate this research with the economics work discussed above. So, while network approaches are good at teaching us about network topology and growth, any discussion about knowledge flow will be very limited without a more contextualised approach. This reflects the differences in scientific objectives mentioned above.

In addition to the modelling of the patent citation network directly, there may be more direct and informative network-based ways of measuring knowledge flow. In particular, aspects of technological diffusion can be modelled as a social phenomenon, and can, therefore, be studied as information spreading through a social network. This line of research is undoubtedly a fertile one, and it is being catalysed by increasing amounts of data becoming available on the structure of inventor-collaboration networks [42]. Some influential studies on smaller or more aggregated networks have been done in this field [167, 168], and attempts to model and measure large-scale knowledge spreading phenomena on these collaboration networks are sure to follow. For example, one recent work [169] finds that the community structures in co-invention networks affect the timing of citations to new patents granted to inventors within these communities. That work illustrates the need for the application of more advanced network analysis tools than simple social distance approaches [167, 168] to understand information diffusion in the co-invention network.

4.1.3 The Separability Assumption

Previous research on the dynamics in both patent and scientific citation networks often seems to assume, for mathematical convenience, a rather profound fact: the mechanisms responsible for the diffusion of knowledge and the mechanisms responsible for the obsolescence of knowledge are entirely independent of one another.

The existence and strength of the preferential attachment mechanism is something we often want to measure, as the network growth dynamics is highly sensitive to nonlinearities in this phenomenon (as shown in Section 3.2.6). We also know that knowledge obsolescence takes place; most ideas or inventions do not stay relevant forever, and they are generally most relevant when they are first proposed. The network growth dynamics is also very sensitive to the functional form of this node ageing, and to the speed of this ageing (see Section 3.2.5).

In particular, previous empirical measurements of citation dynamics assume that the success-breeds-success mechanism $f(k)$ isn't dependent on time, and that knowledge obsolescence $A(t)$ isn't dependent on the success of the idea [36–38, 94, 126, 170–173]. This scenario is often assumed in theoretical models as well [35, 39, 137, 174, 175], often without acknowledgement that this implied separability of the network growth dynamics is an assumption that may be a significant finding if found to be true in specific real-world networks. The probability $\Pi(k, t)$ of receiving a citation, in its most basic form, would be

$$\Pi(k, t) \propto f(k)A(t) \quad . \quad (4.2)$$

The formulation in Eq. (4.2) makes measuring both preferential attachment and node ageing much easier but ignores the possibility of, for example, a popular patent (or paper) ageing more slowly due to this popularity. When the parameters governing network growth and node ageing are measured, it is often observed that the parameters governing the preferential attachment mechanism can change significantly with time [38, 165]. This variability is a problem for these growth models, and this chapter attempts to uncover the extent of this problem after properly considering the context in which these citations are made. In particular, we focus on the evidence for the stability of parameters for groups of patents published at roughly the same time in the same technology area – citation network research generally assumes stable growth parameters for all cohorts at all times (when they are not case studies). Based on the results detailed

in this chapter (see Section 4.5), for example, it would appear unlikely that the rate of knowledge obsolescence is the same for inventions related to mechanical engineering in the 1960s than for semiconductors in the 1990s. This observation means aggregating data from disparate technology categories is likely to bury many exciting results.

There are many nice properties of a separable rate function that could be taken advantage of if found to be a valid assumption. If Eq. (4.2) were approximately true, it is possible to find an analytic expression for the citation distribution and related quantities such as the total citation rate or the average number of citations for a particular cohort (as demonstrated in Section 3.2.2). Additionally, this separation would mean that the mechanisms responsible for growth and obsolescence could be treated as independent mechanisms - an empirically and conceptually useful feature.

Lastly, a separable rate function would also constitute a useful model for comparing different cohorts in terms of their node ageing or network growth characteristics; these dynamics could differ by technology category, time of grant, or any other patent attribute. It would, therefore, be a productive line of enquiry to check whether Eq. (4.2) held true when the patent citation network was analysed in its proper context.

4.2 Patent Citation Data

4.2.1 Data Sources

All data used in this chapter are downloaded from the United States Patent and Trademark Office (USPTO)² and consists exclusively of USPTO-granted patents and their metadata (including citations). The patent classification system used in this thesis is the Cooperative Patent Classification (CPC),³ a system designed as a harmonisation of the classification schemes in use by the USPTO and the European Patent Office (EPO). As of 2015, both patent offices have been using the CPC exclusively, and as such, the decision to use the CPC scheme ensures the work detailed in this thesis is consistent

²All bibliographic data available to download directly from the USPTO website at <https://bulkdata.uspto.gov/data/patent/ptmtdvd/>.

³CPC data for both USPTO patent grants and patent applications can be found at <https://bulkdata.uspto.gov/data/patent/classification/cpc/>. Patents are retrospectively matched via their existing concurrences with the European Classification System, on which CPC is largely based.

with any future research in this area. The CPC scheme is updated at bimonthly intervals in order to reflect the ever-evolving nature of technological development – the CPC scheme adopted for this thesis was the version published in November 2016. The fact that this scheme is now out of date is not expected to impact the results of any analysis conducted in this thesis, due to the historical nature of the patents considered here as well as the technological aggregation level of the cohorts considered. The technology categories used in this thesis are listed in Table 4.3.

4.2.2 Technology Classification

Different industries may have different rates of development, different knowledge management strategies, or even different knowledge diffusion mechanisms. There is a myriad of reasons why some technologies have higher rates of patenting and different citation practices that manifest themselves in the data. For example, it would be reasonable to expect different network growth dynamics for patents classified as “Human Necessities” when compared to those classified as related to “Electricity”, simply due to the different technological life cycles and speed of development in these industries [77, 163, 176]. It is therefore essential to split up patents by technology category to control for the ways these differences might affect the way in which the citation network grows.

At the same time, however, it is not obvious that the parameters that characterise preferential attachment or knowledge obsolescence should be significantly different between these technologies. In particular, different industries may have different (or incomparable) minimum inventive step heights required for a patent, and this would affect the speed of obsolescence measured using patent data. For some technologies, this may work in the opposite direction to obsolescence, where a large inventive step counters a fast obsolescence time. Thus, even the observed node ageing may be the convolution of many node heterogeneity-driven effects, presenting a good case for further investigation of a more fine-grained approach to studying the citation network dynamics than previously undertaken.

There is generally not an obvious choice for the system of technology classification that is most appropriate for a given application. In this thesis, the system common to the USPTO and the EPO is used (CPC), as this system is empirically unambiguous, easily accessible, and likely to be the most common classification system used in future work

in this field.⁴ Table 4.3 describes all top-level technology categories in the CPC scheme, as well as the largest sub-categories within each category.⁵ However, there exist other, more data-driven approaches that may be better for specific purposes that require fine-grained classification or patent matching (and are therefore not suitable for the analyses conducted in this thesis); these include approaches based on semantic similarity [177] or network-based methodologies [178–180].

4.2.3 Cohort Selection

To control for grant time, a cohort of patents published at about the same time must be chosen; this means choosing a span of time whereby any patents granted during this period are in the cohort. A number of considerations need to be balanced while making this decision given the following, often conflicting, requirements for the adequate measurement of network growth, including:

1. A sufficient number of patents in the cohort;
2. Enough time to conduct a robust analysis of the citation network evolution; and
3. Relative homogeneity across the cohort concerning new legislation or procedures affecting patenting behaviour.

The first requirement asks for a large cohort, and due to the ‘patent explosion,’ [72] this suggests choosing as late a cohort as possible with many years of patent grants included. The second requirement asks for many years during which citations can accrue and the network growth dynamics measured, that is, as early a cohort as possible. The last requirement asks that our cohort not overlap with points in time when patenting behaviour changes, such as the implementation of new procedures that change the

⁴Note that throughout this thesis the term *technology category* is used to describe the different categorisations of technology in the CPC classification system. In the CPC itself, however, these top-level categories are officially termed *Sections*. As this word is not typically associated with the description of different technology types, *category* will continue to be used in order to both avoid confusion and ensure consistency with the economics literature.

⁵In this work, patents classified as belonging to multiple classes are treated as belonging to each of these classes independently. The rationale for this decision concerns the comparison of results for different classes: if two classes share many patents, then these classes likely have similar technological properties and, therefore, similar citation dynamics.

mechanisms of knowledge transfer [181, 182], or laws that change the patent output of particular sectors [183, 184]. Additionally, there are secular changes in patenting strategy that occur gradually over time that do not have a singular cause [71]. Restricting the size of the cohort by narrowing the time period used to define it will mitigate this effect. The arguments above all suggest the necessity to restrict the number of years used to define the cohort. We will now examine the practical considerations that must be made in parallel to these empirical requirements.

While our patent grant data is available until the end of 2014, there is a *grant lag*. Because the citation lag in this thesis depends on the application date of the patent (see Section 4.3.1), the time between application and grant affects the number of citations observed. This interval is called the grant lag and means that complete data is not available for the whole period of observation. For example, while many patents were applied for during 2014, we do not know which ones have been granted. Therefore, our citation counts for 2014, as measured by the application dates, are incomplete. While the citation inflation calculation (see Section 4.3.2) will correct for most of this effect, if subcategories are heterogeneous with regard to *both* their grant lag and network growth dynamics, then any measurements taken during these incomplete periods will be biased. The average grant time being between two and three years (see Fig. 4.1), we will consider the data suitable for our purposes if we disregard the final three years of the data available – we therefore include in our dataset the citations to our cohort until the end of 2011.

We consider a minimum of ten years of citation accrual sufficient to measure the growth dynamics, which means the cohort must not extend beyond the end of 2001. This is convenient, as this is also the time at which the USPTO started publishing patent applications that were more than 18 months old (irrespective of the application status), which would add a new avenue for knowledge transfer that may increase the initial citation rate of patents granted after this time and may distort results.

The mid-1980's saw the start of an explosion in the rate of patenting (see Fig. 4.1), reflecting a substantial change in firm's strategies in response to new legislation and USPTO procedures [71]. This variation suggests a restriction of the number of years included in our cohort in order to mitigate any unwanted effects this turbulent environment may have. Three years is manually chosen as an appropriate cohort size to both control for this effect while maintaining a large sample of patents.

Figure 4.1: Total patent grants (bars) and average grant lags (line) for patents granted from 1967-2011. The cohort of patents selected for this thesis is highlighted in green.

This rise in the rate of patenting also means there is much more data contained in the same-width time window at the end of the 1990s than at the beginning. With all of the above factors considered, the three year period 1999-2001 (inclusive) was chosen as the cohort of patents representing the best balance of sample size, time for citation accrual, and grant-time homogeneity concerning changing external pressures. Summary statistics for this cohort can be found in Table 4.1.

4.3 Data Preparation

4.3.1 The Citation Lag

In line with previous work in economics on patent citation dynamics [163,185], the time between the grant of the cited patent and the application of the citing patent was chosen as the definition of the citation lag. The primary reason for this choice is that, if knowledge flow is what is being represented by patent citations, what we would like is the date of invention of the citing patent. As this date is not generally public information, the application date is used as a proxy, being the closest date available to the time of invention [185].

Sometimes the date of application of the citing patent is *before* the cited patent is

granted, leading to a negative citation lag. This is possible through both examiner citations added during the examination process and through self-citations, as the applicant must provide information on pending patent applications as part of their duty of candor. In this thesis, these negative-lag citations are aggregated to form the initial distribution $n_0(k)$ and are not considered to be indicative of knowledge diffusion between inventors.

4.3.2 Citation Inflation

Patenting rates change over time in response to changing political and economic environments unrelated to the intrinsic characteristics of the patents granted. With variable numbers of patents being granted comes variable probabilities of receiving a citation from a patent granted in any given year. It is therefore important to correct for this, as the measured growth dynamics (and technological obsolescence in particular) will be affected by changing rates of patent grants. While the rate of progress may be variable, this variability is not related to the properties of the patents in the selected cohort. In order to compare patents assigned to different technology categories or granted in different time periods it is necessary to correct for any changes in patenting behaviour that occur after the grant of the cohort. In this thesis, a version of the fixed effects approach suggested in Hall et al. [185] is followed; this is quickly becoming standard practice in longitudinal bibliometric analysis [186]. With additional statistical assumptions about the citation lag, it is possible to take a quasi-structural approach [185] by fitting a full model to the citation lags (such as that of Caballero & Jaffe [12]). Controlling for citing year, cited year, and technology field, deflators can be extracted and applied to and from any patent from any technology field. However, as the citation lag structure is precisely what we are modelling, this is unsuitable for use here.

As we will be correcting for citation inflation and not patenting inflation, there exist two different phenomena occurring simultaneously that each need to be considered in isolation: the changing number of patents being issued, and the changing number of citations being made on each patent. Generally, both of these quantities are increasing with time. With these considerations in mind, it is important to reiterate that citation inflation should correct for any external driving forces that change patenting behaviour over time - thus changing the relative value of a citation.

We would ideally like to control for extrinsic effects (such as the changeable rate of

Figure 4.2: Patent applications filed by Intel Corporation over a three-year period starting in 1999. A clear quarterly surge in patent applications can be observed. By aggregating citations in three-month blocks, most unwanted artefacts due to these types of cyclical patenting behaviours are mitigated. Note that this is a particularly extreme case of periodicity.

patenting), but not the intrinsic properties of the system (such as the number of citable patents). As proxies, we treat increasing numbers of patent grants as extrinsic and increasing numbers of citations per patent as intrinsic. Citation inflation is then simply a matter of choosing a benchmark time period $[T_0, T_0 + \Delta T]$ and then adjusting the value of a citation from any other time period $[T, T + \Delta T]$ to reflect the difference in the number of patents issued in this period when compared to the benchmark period. For reasons discussed in Section 4.2.3, the work presented in both this chapter and in Chapter 6 uses $T_0 = 1$ January 1999, while the analyses in Chapter 5 uses the same T_0 for related reasons.

This process requires a choice of time resolution ΔT for the required corrections accounting for citation inflation. As there is no shortage of data, even a small time window would be sufficient for this task. However, in order to avoid any unforeseen empirical problems related to periodic reporting periods (see Fig. 4.2), and to smooth much of the noise in the patenting rate, this window is chosen to be $\Delta T = 3$ months.

The value of a citation may also depend on the technology field from which it is coming. Citation counts from specific technology categories are more volatile (or grow-

ing/dying faster) than others, so it is important to control for citation inflation for each technology category independently. If a citing patent i is assigned multiple technology categories (defining the set J_i), then the value of a citation from that patent is the weighted average value of a citation from each of those categories. This choice is due to the ambiguity of the ordering of technology categories listed on a patent, and also ensures flexibility with regard to both the classification scheme and the chosen level of the scheme's hierarchy. In this thesis, we use the top-level CPC categories to discriminate between technology types (categories listed in Table 4.3).

Mathematically, the value of a citation from patent i applied for in the period $[T_i, T_i + \Delta T]$ is then

$$c_i(T_i; T_0, \Delta T) = \frac{1}{|J_i|} \sum_{j \in J_i} \frac{N_j(T_0, \Delta T)}{N_j(T_i, \Delta T)} \quad , \quad (4.3)$$

where J_i is the set of technology categories to which patent i has been assigned, $N_j(T_i, \Delta T)$ is the number of patents applied for in the time interval $[T_i, T_i + \Delta T]$ that were assigned to the technology category $j \in J_i$, $N_j(T_0, \Delta T)$ is the number of patents applied for in the time interval $[T_0, T_0 + \Delta T]$ that were assigned to the same category $j \in J_i$, and $|J_i|$ is the cardinality of the set J_i . From this point on, any measurement of citations to a patent refers to these inflation-corrected citation values as shown by category in Fig. 4.3. Statistics for the inflation-adjusted citations to the 1999-2001 cohort can be found in Table 4.1. Patents which are classified into multiple categories are treated as belonging to each category equally. For example, if a patent is classified into seven subcategories, five under category A and two under category B, the patent belongs equally to each of categories A and B. The reasoning for this is due to uncertainty in the breadth of subcategories; categories may differ in their level of technological discrimination at the subcategory level. Returning to the previous example, it would be difficult to judge whether the patent was more likely to resemble other patents in A (rather than B) if this category simply had many more subcategories to which it could be assigned, or less distinct boundaries between these subcategories. For this reason, there is no such weighting in the citation value calculation.

Another approach to control for varying numbers of new nodes joining the network over time is implemented in the context of citations to scientific articles in Wang et al. (2013) [187], where the authors approximate the functional form of the increase in published articles as an exponential rise. This method turns out to be mathematically

Figure 4.3: The value of patent citations, $c(T; T_0, \Delta T)$, for different technology categories over the course of the ten-year period during which citation accrual is observed, calculated as described in Eq. (4.3) for each single-category case. Citation values for individual patents are calculated as the average of the values for all categories to which each patent is assigned.

convenient as the parameter characterising the rise can be absorbed into another parameter in the model. However, while this may be an appropriate (if strong) approximation to the real rise in scientific publications over an extended period of time for highly aggregated cohorts, no such simple form is able to model the changing numbers of patent grants to technology specific cohorts for the time period examined here; the methodology described above and used throughout this thesis is intentionally much more flexible.

4.3.3 Binning

In order to test the assumption of separability of the rate function $\lambda(k, t) = f(k)A(t)$ empirically, we must split the data into bins that cover all possible combinations of degrees and times.

After citations have been inflation-adjusted according to the procedure described in Section 4.3.2, all citations to patents in the cohort are assigned a time t corresponding to the citation lag between cited and citing patents. Two time series are then constructed: the number of citations patent i has accrued by time t , denoted by $k_i(t)$, and the number of citations patent i gains in the next period $(t, t + \Delta t]$, denoted as $\Delta k_i(t)$.

Category	Total Patents	Total Citations	Mean Citations	Median Citations
A	88 721	1 636 104	18.4	7.2
B	103 488	1 050 223	10.1	5.5
C	79 028	787 118	10.0	4.3
D	8 184	70 269	8.6	4.0
E	16 738	189 929	11.3	6.5
F	47 304	388 891	8.2	5.0
G	131 603	2 294 008	17.4	7.7
H	124 582	1 972 189	15.8	7.5
Y	99 170	1 350 949	13.6	6.3

Table 4.1: Summary statistics for the cohort of patents granted by the USPTO from 1999-2001 and the citations to this cohort (inflation-adjusted as per Eq. (4.3)) with an application date of 2011 or earlier.

Our further analysis will be based on the assumption that the rate $\lambda(t)$ at which individual patents gain citations is a function of both k and t ,

$$\lambda(t) = \bar{\lambda}(k(t), t) \quad . \quad (4.4)$$

As an empirical measure for $\bar{\lambda}(k, t)$, we use the average citation rate for the group of patents with k citations in the interval $(t, t + \Delta t]$, that is, $\bar{\lambda}(k(t), t) \approx \overline{\Delta k_i(t + \Delta t)} / \Delta t$. In this manner, we obtain an array of citation rates for each technology category with every entry corresponding to the average rate of citation to the group of patents assigned to that technology category and having k citations at time t . We therefore place each patent in a particular bin based on its accrued citations, k_i , at the end of each time period. As described above, the binning in the time dimension is simply a linear scale where each bin has the width Δt . For the same reasons as the choice of citation-inflation resolution, Δt is set to three months. Note that this does not correspond directly to the three-month blocks for which the citation-inflation adjustment is made, as t depends on the month of the cited patent's grant.

In the k dimension, partial logarithmic binning is used [188]. This means each k -bin has the same width on a logarithmic scale for values of k larger than some minimum, which is appropriate for the observed functional form for the k dependence of $\bar{\lambda}(k, t)$ (see Section 4.5). The minimum is necessary due to the discrete nature of citations (even after inflation correction); in this thesis, the minimum is set to $k = 5$, after which any effects

due to discretisation become unnoticeable. The size of the logarithmic bins depends on the amount of data available – if there are N patents in a cohort then the bin size is proportional to $\ln(1 + 100000/N)$, reflecting the fact that most categories have approximately 100 000 patents in the cohort, with significantly larger bins only necessary for categories that are much smaller than 100 000 patents. This scaling was chosen manually with the goal of having a sufficient number of bins to model the rate function accurately (as described in Section 4.4) while ensuring each bin generally contains enough patents to minimise the effect of outliers. Other methods to bin data include cumulative density or quantile approaches, however these are not used here due to the discrete nature of citation data. At early times, in particular, many patents have no citations at all and this may restrict the information accessible via a quantile approach. This problem can be easily rectified in the logarithmic approach described above by weighting bins by the number of patents in each bin for any regressions conducted.

4.4 Modelling Strategy

4.4.1 Observation of Preferential Attachment and Ageing Phenomena

In order to observe the functional forms of preferential attachment and ageing, $f(k)$ and $A(t)$ respectively, we can fix one variable while measuring the effect the other has on the rate $\bar{\lambda}(k(t), t) = f(k)A(t)$. The k and t dependencies of the citation rate are then observed to be power-law and exponential respectively (alternate forms are examined in Section 4.5.4), such that

$$A(t) = A_0 \exp\left(-\frac{t}{\tau}\right) \quad \text{for } t \geq t_0 \approx \tau/2 \quad , \quad (4.5a)$$

$$f(k) = k^\alpha + f_0 \quad , \quad (4.5b)$$

where τ is a decay constant, α an exponent representing preferential-attachment strength, and f_0 is a constant contribution that allows for a non-zero citation rate for patents with zero citations (the additive nature of this parameter means its importance to growth dynamics quickly diminishes as k increases).

By fitting Eqs. (4.5a) & (4.5b) for different values of the held-constant variable, we can measure the consistency of the values of the parameters τ , α , and f_0 across this

variable. For example, by measuring $A(t)$ for various fixed values of k , we can check whether the fitting parameter τ changes with k , or in other words, whether the rate of knowledge obsolescence depends on the number of accrued citations. Then we repeat this procedure with $f(k)$ by fixing t and measuring α and f_0 . If none of these parameters are shown to vary systematically, then the separability Ansatz is valid.

4.4.2 Model Fitting

To fit the above functional forms to the observed citation rates and obtain estimates of the network growth parameters, a weighted non-linear logarithmic regression is used. This regression takes the logarithm of the observed rates (to reduce the effect of outliers on the fitting parameters) before conducting a non-linear least-squares regression to the transformed data. In the case of the exponential obsolescence function $A(t)$, this is equivalent to a linear regression with gradient $-1/\tau$ and intercept $\ln(A_0)$.

Each observation in these fits represents the average rate for a set of patents which are in the same k -bin at the same time t . As such, the bins with larger numbers of patents should be given a larger weight in the regression. However, because the data is extremely bottom-heavy (there are many patents with small k), and the preferential attachment strength parameter α is best determined by a broad range of degrees, weighting the regression with the raw number of patents represented by each observation is not appropriate. For this reason, observations are weighted by the logarithm of the number of patents in each observation, which results in a very good fit across the whole range of k -bins. The same methodology is applied to the $A(t)$ fits with similarly satisfactory results.

Before fitting, the data is trimmed to exclude those patents that accrue a number of citations higher than the 99th percentile after ten years. For example, if after ten years the 99th percentile in accrued citations is measured to be 100 citations, then patents are excluded from measurements for times after they accrue 100 citations (and not before). In the case of the preferential attachment measurements, this means that for most times at which the parameters are estimated, the number of excluded patents constitute much less than one percent of the sample. The reason for this truncation is because the number of patents in the largest k -bins becomes exceedingly small and only amass enough patents for an accurate rate measurement at very late times, if at all. If these patents are included in the regression described above, it is found that the estimated parameter

values remain the same while uncertainties increase. This observation is consistent with the recognition that these patents only increase the noise level in our measurements when using the described methodology, and justifies their omission from the parameter estimation procedure.

Concerning the preferential attachment curve fitting, α is much more important to network structure and growth than is f_0 , however fitting both parameters at the same time introduces a small interdependence between the two that is exacerbated by the relatively high uncertainty in the measurement of f_0 . For this reason, we remove codependency of α and f_0 in the fitting procedure using a two-step regression process. The first step fits both α and f_0 to obtain a time series of estimates for each parameter. The second step then refits the curves with f_0 fixed to the average value obtained from the first step. This process results in the removal of any noise in the measurement of α caused by the simultaneous fitting of f_0 . This process is only appropriate when there is no systematic variability in f_0 , which is observed to be true for our data (see Figs. 4.5, 4.10-4.12). The parameters listed in Table 4.2 are the averages and uncertainties for f_0 and α obtained from the first and second fitting stages respectively, and the uncertainties represent a 95% confidence interval; the average is the arithmetic mean weighted by the inverses of the variances in each parameter measurement, and the variance in this mean is the bias-corrected variance of the parameter measurements.

4.5 Results & Discussion

4.5.1 Preferential Attachment

Preferential attachment, represented by the function $f(k)$, appears to be an excellent model for the functional dependence of k on the rate function⁶ (with the conceptual caveats outlined in Section 3.4), see Fig. 4.4. As shown in the exemplar in Fig. 4.5 (equivalent figures for all other categories can be found in Section 4.7), the growth parameters α and f_0 show no systematic deviations from a constant value for most technology categories and are therefore generally consistent with the separable rate function Ansatz.

⁶An alternate form of the preferential attachment model is examined in Section 4.5.4, with similar results. The justification for the form of preferential attachment used in this thesis is also given there.

Figure 4.4: Fitting the time and citation-number dependences of the empirical citation rate for the Human Necessities (A) technology category. **Left:** Mean number of additional citations, $\Delta k(t + \Delta t)$, as a function of time t for fixed citation bins with log-midpoints at $k = 1, 16$ and 80 on a log-linear scale. Hollow symbols are derived from the citation data, while lines represent the fit to the form of $A(t)$ given by Eq. (4.5a). Deviations from exponential-ageing behaviour occur only at short times. **Right:** Mean number of additional citations, $\Delta k(t + \Delta t)$, as a function of cumulative number of citations $k(t)$ for various fixed times $t = 2, 5$ and 9 years on a log-log scale. Hollow symbols are derived from the citation data, while lines represent the fit to the form of $f(k)$ given by Eq. (4.5b).

There does not appear to be much variation in the preferential attachment strength α between categories, with $\alpha \approx 1.1$ (see Table 4.2). This value is slightly lower than those typically found previously (e.g. [165]) but still implies super-linear growth dynamics. There is some variation in the parameter f_0 between categories, reflecting differing probabilities of being cited while having very few citations, however, the uncertainty in this variable means that all measurements are generally consistent with one another. The observed technology-independent universality of the preferential-attachment-type growth behaviour points to the existence of common drivers for knowledge flow as proxied by citations.

The “Physics” category is a notable exception with regard to stable growth parameters. For this category, both α and f_0 increase slowly with time. The reasons for this behaviour are still unknown.

Category	α	f_0	τ [years]	A_0 [years ⁻¹]
A	1.13 ± 0.02	1.21 ± 0.29	5.20 ± 0.48	0.33 ± 0.04
B	1.10 ± 0.04	1.46 ± 0.43	5.38 ± 0.47	0.25 ± 0.02
C	1.11 ± 0.03	0.95 ± 0.19	5.14 ± 0.47	0.32 ± 0.03
D	1.14 ± 0.14	1.33 ± 0.72	5.29 ± 0.97	0.24 ± 0.06
E	1.14 ± 0.14^a	1.89 ± 0.81	5.52 ± 1.99^b	0.21 ± 0.03
F	1.13 ± 0.05	1.66 ± 0.56	5.77 ± 0.64	0.21 ± 0.03
G	1.09 ± 0.08^a	1.19 ± 0.28	5.19 ± 0.91^b	0.28 ± 0.02
H	1.08 ± 0.06^a	1.30 ± 0.31	5.19 ± 0.70	0.25 ± 0.02
Y	1.09 ± 0.03	1.32 ± 0.35	5.42 ± 0.36	0.27 ± 0.02

^a Here α showed a moderate residual time dependence.

^b Here τ showed a moderate residual dependence on the cumulative number of citations.

Table 4.2: Parameter values for preferential-attachment-type growth (α , f_0) and exponential ageing (τ , A_0), extracted for patent cohorts from the individual CPC technology categories listed in Table 4.3.

4.5.2 Ageing

The node ageing function $A(t)$ also appears to be a good model of obsolescence in the growing citation network, as shown in Fig. 4.4. For some technology categories there is some deviation from the exponential form at small times (see Fig. 4.6), however an exponential decay as in Eq. (4.5a) provides a good fit for all technology categories for $t \geq \tau/2 \approx 2.5$ years. The parameter τ is approximately constant across the whole range of k -bins for most categories (see Fig. 4.5 and Section 4.7), again with the exception of the Physics category,⁷ and similar results are found for the scale parameter A_0 . While there is variation between technology categories for τ , the uncertainties are proportionally larger than for α . As such, while the differences shown in Table 4.2 are interesting, they are mostly not statistically significant. The most obvious differences are those between the stereotypically high-tech categories such as the Physics and Electricity categories and the low-tech categories such as the Fixed Constructions and Mechanical Engineering - the former categories age faster than the latter. While this result fits in well with one's intuition about the nature of these technologies, more exploration needs

⁷Hypotheses for citation lifetime increasing with popularity include an increased importance of auto-correlations in this category, less predictable citation histories, or perhaps different knowledge diffusion mechanisms reflecting a closer tie to academic research.

Figure 4.5: Disentangling preferential-attachment-type growth from ageing in the citation rate of patents in the Human Necessities technology category. **Left:** Obsolescence time τ extracted for each bin of patents with fixed number of citations k from fitting the citation rate to the form of Eq. (4.5a) for times $t \geq t_0 = 2$ years. **Centre:** Preferential-attachment exponent α , extracted for fixed times t from fits of the citation rate to the form of Eq. (4.5b). **Right:** Constant contribution f_0 to the preferential-attachment part of the citation rate, extracted for fixed times t from fits to Eq. (4.5b). Circles are fit-parameter values, solid lines their weighted averages, and black dashed (red dotted) curves show 95% confidence intervals for fit-parameter values (weighted averages).

to be done to determine the true magnitude of these differences and check whether they are also manifested in the small-scale structure of the citation network.

For some k -bins in some categories there exist small systematic deviations from the exponential ageing function, as shown in Fig. 4.6. While these deviations are much smaller than those observed in Higham et al. (2017) [157] due to improved methodology, they are significant enough to require a correction term when calculating the analytic form of the degree distribution (see Section 4.5.3). These deviations require further exploration; however, it is later observed in Chapter 6 that this deviation may be accounted for by heterogeneities in patent quality. The model used in this chapter assumes no differences in intrinsic patent quality, and therefore most patents have roughly the same chance of being cited at small times (some patents may have accrued negative-lag citations, however). If there exists a quality distribution with non-zero variance, this assumption is violated, and this may affect measurements of node ageing until enough time has passed such that accrued citations are more reflective of patent quality. Alongside this effect, the mechanisms of knowledge transfer may also change over time, and this may change the functional form of knowledge obsolescence. Both of these hypotheses are consistent with two distinct node ageing regimes that are more significant for certain types of technology.

Figure 4.6: Even when deviations from the exponential ageing function in Eq. (4.5a) are large, as for the Electricity category (H), their effects remain confined to the first three years of citation accrual. **Left:** Mean number of additional citations, $\Delta k(t + \Delta t)$, as a function of time t for fixed citation bins with log-midpoints at $k = 1, 15$ and 80 on a log-linear scale. Hollow symbols are derived from the citation data, while lines represent the fit to the form of $A(t)$ given by Eq. (4.5a). The deviation from this exponential form is greater than for the Human Necessities category shown in Fig. 4.4. **Right:** Average citation rate $\Lambda(t)$ as a function of time t . Empirical data for patents in the Electronics technology category are shown as hollow circles. The red dashed line shows the theoretical rate values calculated as described in Section 4.5.3 with $\Gamma_0 = 0.78$. Parameters used are those listed for the Electricity category (H) in Table 4.2.

As for the preferential attachment function, there are also alternative forms of the ageing function. While the two-parameter exponential decay is the most parsimonious form found in the literature, there are other, perhaps more popular, forms of the ageing function⁸ [34]. These alternate forms and the justifications for avoiding their use in this thesis can be found in Section 4.5.4.

4.5.3 The Degree Distribution and Related Quantities

The degree distribution $n(k, t)$ (reiterated in Eq. (4.6)) depends only on the preferential attachment function $f(k)$, the ageing function $A(t)$, and the initial citation distribution $n_0(k)$. However, the deviations from the exponential ageing function observed for some categories mean that $n(k, t)$ will not accurately model the true citation distribution or related quantities such as the average citation rate when $A(t)$ is modelled as a purely

⁸These forms are usually applied either to citations between scientific articles or citations between patents where the citation lag has been calculated differently from its definition in this thesis (Section 4.3.1).

exponential decay.

In order to correct for this discrepancy and verify whether our fitting parameters can accurately model the degree distribution, we need to account for this deviation at small times. Remembering the degree distribution,

$$n(k, t) = n_0(k) [\gamma(t)]^{f(k)} + \sum_{l=0}^{k-1} n_0(l) \left(\prod_{m=l}^{k-1} f(m) \right) \sum_{q=l}^k \frac{[\gamma(t)]^{f(q)}}{\prod_{\substack{m=l \\ m \neq q}}^k [f(m) - f(q)]} \quad , \quad (4.6)$$

we can specialise the time dependent component $\gamma(t)$ to the exponential decay in Eq. (4.5a), such that

$$\gamma(t) = \Gamma(t) \exp \left\{ \tau A_0 \left[\exp \left(-\frac{t}{\tau} \right) - 1 \right] \right\} \quad . \quad (4.7)$$

Here, $\Gamma(t)$ quantifies any systematic short-time deviation from exponential decay and is defined as

$$\Gamma(t) \equiv \exp \left\{ - \int_0^t dt' \left[A(t') - A_0 \exp \left(-\frac{t'}{\tau} \right) \right] \right\} \quad . \quad (4.8)$$

Because we observe exponential ageing at long times, $\Gamma(t)$ goes to a constant Γ_0 for $t \geq \tau/2$.

Extracting Γ_0 empirically is a numerical process. We first measure the deviation from the fitted exponential decay as a function of time, after which we observe the saturation point and average all measurements after this to obtain an estimate of Γ_0 . Note that except for the measurement of Γ_0 , all of the parameters used to determine the analytic curves for the citation distribution $n(k, t)$ and the average citation rate $\Lambda(t) = \sum_k n(k, t) k$ are the parameter averages that are listed in Table 4.2.

In practice, Γ_0 is calculated at each point in time by finding the value that Γ_0 must take in the model (Eq. (4.7) and $\Lambda(t) = \sum_k n(k, t) \lambda(k, t)$) to fit the real total rate at that time. As the exponential decay in Eq. (4.5a) fits the observations well after a few years, these measurements will eventually reach a steady value. However, the calculation of A_0 requires knowledge of α and f_0 , as the exponential fitting is carried out for different values of k and at $t = 0$ the model rate is equal to $A_0(k^\alpha + f_0)$. If, at small times, the parameters α and f_0 are systematically smaller or larger than their averages shown in

Table 4.2, then Γ_0 will not be calculated correctly at these times – which are precisely the times during which the deviation exists. This is because the averages for α and f_0 are the values inserted into the model to calculate Γ_0 in the first place. For this reason, Γ_0 cannot be compared between categories where at least one of those categories has α and f_0 systematically different to their averages. This is the case for Category A, as shown in Fig. 4.5, where f_0 is systematically low for the times during which the excess rate exists, and explains why the Γ_0 measured for this category is the same value as for Category H, despite the former displaying noticeably less deviation from exponential decay⁹ (compare Figures 4.4 (left) and 4.6 (left)). In categories for which it is accurately measured, Γ_0 may provide some information about the nature of knowledge diffusion at short time, as it quantifies the deviation from the model that describes the long-time diffusion process [157]; however, this feature of knowledge obsolescence (also seen for scientific articles, see Chapter 5) requires further work to understand its cause.

The good agreement between the real citation distribution and the analytic one for Category A at small times (e.g. the curve at $t = 2$ in Fig. 4.7) is, therefore, coincidental, as the deviation from exponential at small times is almost exactly balanced out by the systematic deviation in f_0 . A good fit is also found for the (more aggregated) total citation rate $\Lambda(t)$ for almost the whole citation accrual period, also seen in Fig. 4.7 – again, coincidental at small times but showing good agreement for longer times as well. For other categories, such as Electricity, the deviation from exponential ageing is initially on the order of 20-30%. After measuring Γ_0 and correcting for this deviation, however, a good fit is achieved for $t \geq 3$ as expected (see Fig. 4.6).

4.5.4 Alternative Functional Forms of Rate Function Components

Initial Attractiveness

A common alternate form of $f(k)$ that has been used to capture non-linearities in citations networks is $f^*(k) = (k + k_0)^{\alpha^*}$. Note this model only differs significantly from

⁹The estimated values of the parameter f_0 for Category H do not exhibit any systematic deviation from the average listed in Table 4.2 (see Fig. 4.11).

Figure 4.7: **Left:** Distribution function $n(k, t)$ for patent citations, plotted as a function of the number k of citations for fixed times $t = 2, 5$, and 9 years. **Right:** Average citation rate $\Lambda(t)$ as a function of time t . Empirical data for patents in the Human Necessities technology category are shown as hollow symbols. Curves have been calculated from the general theoretical expression in Eq. (4.6) (with $\Lambda(t)$ calculated as described in Section 4.5.3) with functional forms for the ageing and preferential attachment functions given in Eqs. (4.5a) & (4.5b), respectively. To capture deviations from exponential ageing at short times, the factor Γ_0 has been introduced as explained in Eqs. (4.7) & (4.8), with $\Gamma_0 = 0.78$. Parameters used are those listed for Human Necessities (category A) in Table 4.2.

$f(k)$ for small k , as

$$\lim_{k \rightarrow \infty} \frac{f(k)}{f^*(k)} = 1 \quad (4.9a)$$

$$\text{and} \quad \lim_{k \rightarrow 0} \frac{f(k)}{f^*(k)} = \frac{f_0}{k_0^{\alpha^*}} \quad (4.9b)$$

There are strong reasons for selecting $f(k)$ over $f^*(k)$. The first is practical: the preferential attachment strength parameter α should be independent of the initial citation rate proportional to f_0 . By allowing α^* and k_0 to be dependent on one another as fitting parameters (as for the function $f^*(k)$ for small k), their meaning becomes more ambiguous. Secondly, the deviation from preferential attachment with $f_0 = k_0 = 0$ for nodes with small degree should reflect a mechanism by which a node can start to accrue new links. However, the effect of the super-linear ($\alpha > 1$) preferential attachment mechanism generally observed in citation networks should increase the chance of a new citation based *only* on citations already accrued. By including k_0 in the argument, the initial rate needlessly depends on α^* ; more seriously (as α^* is much more consequential for network structure at long times), α^* depends needlessly on the initial rate. While

Figure 4.8: Parameter estimates for the alternate functional form of preferential attachment $f^*(k) = (k + k_0)^{\alpha^*}$. **Left:** Exponent α^* extracted from the fitting of $f^*(k)$. **Right:** Constant contribution k_0 extracted from the fitting of $f^*(k)$. Circles are fit-parameter values, solid lines their weighted averages, and black dashed (red dotted) curves show 95% confidence intervals for fit-parameter values (weighted averages).

using the $f^*(k)$ form does not (with enough data) significantly impact the measurement uncertainty of the growth parameters, the results of this dependence is that when $\alpha^* > 1$ it takes longer for k_0 to become insignificant and this will generally bias α^* upwards slightly. Thus, while the empirical ramifications of choosing $f^*(k)$ over $f(k)$ are not great with large data sets with a large range of degrees with which to measure α , networks in which this range is not available may suffer from a fitting bias. The fitting results for the Human Necessities technology category for the alternate form $f^*(k)$ can be seen in Fig. 4.8, with similar results found for all other categories.

Power-Law Ageing

Many functional forms of node ageing have been suggested in the literature. In theoretical work the form is chosen for mathematical flexibility or convenience [35, 126]; in empirical work (including this one) the form is usually chosen based on observation, which itself depends on both how ageing is measured and how time is parametrised.

However, even in empirical work, convenience can dictate the form of the ageing function. John von Neumann purportedly said “With four parameters I can fit an elephant, and with five I can make him wiggle his trunk” [189]. This exercise has even been carried out [190]. Over-fitting is always a concern in empirical work, and in this thesis, it is argued that it is worth sacrificing some goodness-of-fit in exchange for reducing

model complexity on the grounds that deviations from this simple model are perhaps the result of changing knowledge diffusion mechanisms or unobserved heterogeneities between patents within in each cohort such as fitness [191]. These suspicions are largely confirmed in Chapter 6.

In previous research where network growth models are applied to citation networks, the goal is often to produce a predictive model or test a new generative model, rather than to understand knowledge dynamics (e.g. [187, 192]). In this situation, a three-parameter model such as a modified power-law [38] or some model-dependent analytical distribution may be appropriate. To compare these models of node ageing with the fit obtained through Eq. (4.5a), a modified power-law of the form [38]

$$A^*(t) = \frac{A_0^*}{(1 + t/\tau^*)^\beta} \quad (4.10)$$

is tested on the data for the Electricity category (see Fig. 4.9), where A_0^* , τ^* , and β are fitting parameters. The Electricity technology category was chosen due to the patents in this category displaying a larger deviation from exponential ageing, and thus provided the best opportunity for the model in Eq. (4.10) to perform well. While this form provides a better fit, the uncertainties (displayed in Fig. 4.9) in the estimated parameters are unacceptably large even for purposes of providing evidence for or against the separability of the rate function, let alone the comparison of different technological categories or the study of knowledge transfer mechanisms.

If time is parametrised using the patent sequence numbers, then the ageing function is not monotonic, instead taking on a Weibull-like form [37, 165]. In some ways, this parametrisation is convenient, as it ensures a constant rate of patent grants, something achieved in this thesis via the corrections to citation inflation. However, patent numbers also significantly distort any measurements of knowledge flow due to large variances in the grant lag within and between technology categories. For example, Valverde et al. [37] find that the peak citation rate in the Weibull ageing model is after about 180 000 subsequent patents, which for the period around 1999-2001 corresponds to about 14 months. The average grant lag for this period is about 27 months (see Fig. 4.1), implying that it is highly likely that most of the citing patents represented in this peak were applied for before the grant of the patents they cite. The fact that there exists an observed peak in citation rates in that work can, therefore, be seen as an artefact of the time parametrisation and thus does not correspond to any real phenomena, and

Figure 4.9: Power-law vs. exponential ageing: The three-parameter power law shown in Eq. (4.10) fits the data superficially well but yields very uncertain parameter values. Results shown for Category H (Electricity). **Left:** Comparison of model fits for the case $k = 15$, where the power-law fit (red solid curve) yields $\beta = 2.9 \pm 0.7$ and $\tau^* = (8.6 \pm 6.0)$ years (relative uncertainties of 24% and 69% respectively). This can be contrasted with the exponential fit (black dashed line) that gives $\tau = (5.01 \pm 0.40)$ years – a relative uncertainty of 8% – while providing a fit that is almost indistinguishable from the power-law fit for times $t \geq 2$. **Centre:** Parameter τ^* as obtained from the modified power-law of the form in Eq. (4.10). **Right:** Parameter β as obtained from the modified power-law of the form in Eq. (4.10). For the two graphs on the right, solid red lines represent weighted averages, and black dashed (red dotted) lines represent 95% confidence intervals for the fit-parameter values (weighted averages). Due to their very large values, 95% confidence intervals for the bin $k = 53$ have been left off the figures for readability.

that only the tail of this ageing function (which is a stretched exponential) contains useful information about knowledge diffusion. While this approach may shed light on the growth of real citation networks, there is little scope to connect this work to existing economics literature on the diffusion and obsolescence of technical knowledge.

4.6 Conclusions & Outlook

The citation accrual model used in this chapter implements a different framework from that of Caballero & Jaffe (1993) [12] and the subsequent work discussed in Section 4.1.1. However, the models share the functional form of knowledge obsolescence – an exponential decay – which provides the opportunity for comparison. The parameter governing this behaviour in our model is τ , which is equivalent to $1/\beta_1$ in the model used in Jaffe & Trajtenberg (1999) [75] and seen in Eq. (4.1). The authors in that paper report that $\beta_1 = 0.190$ for the set of all patents granted between 1963 and 1993. Inverting this value gives an equivalent τ of 5.26 years, a similar value to those displayed in Table 4.2.

The results described in this chapter provide the first evidence for separability of the rate function in the growth of the patent citation network, which means that it is possible to move forward from this point with the knowledge that the mechanisms driving obsolescence are likely to be different from those that drive technological diffusion. Additionally, these effects manifest themselves in the network structure such that they can generally be treated as independent. In other words, this implies not only that the rate at which new technology becomes irrelevant does not depend on the popularity of said technology, but the *reasons* technology becomes irrelevant appear to be independent of the mechanisms by which the ideas spread. Additionally, these mechanisms appear manifest themselves across all technologies with equal strength – both long-term obsolescence rates and preferential attachment strengths display very similar values across technology categories.

Measurements of the parameters governing preferential attachment and node ageing in the patent citation network also provide some insight into differences in network growth dynamics between technology types. The magnitude of these differences is not large; this may reflect the high level of aggregation present in the technology categories used, where many patents belong to multiple categories. Because large quantities of data are required for accurate measurement of the network growth parameters, analysing the CPC-defined subcategories (moving down the scheme hierarchy) is almost always impossible without significantly increasing both the time-window that defines the cohort and the size of the k -bins to the point where cross-category comparisons are inconclusive. However, one intriguing avenue for future work is the examination of preferential attachment and obsolescence based on the technological category of the citing patent: do citations inter-category citations display different dynamics to intra-category citations?

In this work, we have not examined any of the mesoscale structural differences between the various technology categories such as clustering or community structure within categories, or indeed whether the CPC categories used are a good representation of a partition that *would* produce differences in the network structure. Some categories, such as Electricity (category H) or Textiles (category D) may be quite self-contained as a collection of patents with similar diffusion and obsolescence characteristics, whereas Operations and Transport (category B) casts a much wider net and may include a much more diverse set of patents (see Table 4.3). Categorisation by community structure or semantic content may produce much more homogeneous technology groups and, upon analysis, may result in more significant differences in network growth parameters be-

tween categories.

Many of the research methods detailed in this chapter were first published by the author of this thesis in Ref. [157]. That work utilised different technological classifications (those listed in Ref. [185]) and a smaller cohort, and many of the methodologies described therein have been significantly improved upon by those described in this chapter. The results here remain qualitatively similar to those described in Ref. [157].

The growth of the patent citation network is often compared to that of the scientific citation network. The substantial differences in the citation processes that generate these networks would, at least on the surface, appear to preclude any comparison. However, both processes involve both the obsolescence of knowledge and some mechanism(s) that lead to good ideas gaining more traction than bad ones. On the other hand, it is reasonable to expect that scientific knowledge is forgotten (as reflected by citations to this knowledge) on a different timescale to the one that governs technological irrelevance, given the differences in the incentives for discovery, knowledge organisation systems, and citation conventions. In the next chapter, we ask: what happens when we apply the same methodology to a scientific citation network?

4.7 Additional Tables & Figures for Chapter 4

Table 4.3: CPC technology categories (officially *Sections*) and the three largest sub-categories (officially *Classes*) in each.

Category	Name	Largest Sub-Categories	Description
A	Human Necessities	61	Medical or Veterinary Science; Hygiene
		01	Agriculture; Forestry; Animal Husbandry; Hunting; Trapping; Fishing
		47	Furniture; Domestic Articles or Appliances; Coffee Mills; Spice Mills; Suction Cleaners in General
B	Operations & Transport	60	Vehicles in General
		01	Physical or Chemical Processes or Apparatus in General
		65	Conveying; Packing; Storing; Handling Thin or Filamentary Material
C	Chemistry & Metallurgy	07	Organic Chemistry
		08	Organic Macromolecular Compounds; Their Preparation or Chemical Working-up; Compositions Based Thereon
		12	Biochemistry; Beer; Spirits; Wine; Vinegar; Microbiology; Enzymology; Mutation or Genetic Engineering
D	Textiles	06	Treatment of Textiles or the Like; Laundering; Flexible Materials Not Otherwise Provided for
		21	Paper-making; Production of Cellulose
		04	Braiding; Lace-making; Knitting; Trimmings; Non-woven Fabrics

Continued on Next Page...

Table 4.3 (Continued)

Category	Name	Largest Sub-Categories	Description
E	Fixed Constructions	04	Building
		21	Earth Drilling; Mining
		05	Locks; Keys; Window or Door Fittings; Safes
F	Mechanical Engineering	16	Engineering Elements and Units; General Measures for Producing and Maintaining Effective Functioning of Machines or Installations; Thermal Insulation in General
		02	Combustion Engines; Hot-gas or Combustion-product Engine Plants
		01	Machines Or Engines In General; Engine Plants In General; Steam Engines
G	Physics	06	Computing; Calculating; Counting
		01	Measuring; Testing
		11	Information Storage
H	Electricity	01	Basic Electric Elements
		04	Electric Communication Technique
		03	Basic Electronic Circuitry
Y	General	10	Technical Subjects Covered by Former USPC
		02	Technologies or Applications for Mitigation or Adaptation Against Climate Change
		04	Information or Communication Technologies Having an Impact on Other Technology Areas

Figure 4.10: Disentangling preferential-attachment-type growth from ageing in the citation rate of patents granted by the USPTO. Rows represent technology categories (excluding Category A, for which the parameter estimates can be seen in Fig. 4.5) while columns represent the parameter measurements.

Rows: Parameter estimates for the cohorts of patents assigned, from top to bottom, categories B, C, and D.

Columns: **Left:** Obsolescence time τ extracted for each bin of patents with fixed number of citations k from fitting the citation rate to the form of Eq. (4.5a) for times $t \geq t_0 = 2$ years. **Centre:** Preferential-attachment exponent α , extracted for fixed times t from fits of the citation rate to the form of Eq. (4.5b). **Right:** Constant contribution f_0 to the preferential-attachment part of the citation rate, extracted for fixed times t from fits to Eq. (4.5b). Circles are fit-parameter values, solid lines their weighted averages, and black dashed (red dotted) curves show 95% confidence intervals for fit-parameter values (weighted averages).

Figure 4.11: Disentangling preferential-attachment-type growth from ageing in the citation rate of patents granted by the USPTO. Rows represent technology categories while columns represent the parameter measurements.

Rows: Parameter estimates for the cohorts of patents assigned, from top to bottom, categories E, F, and G.

Columns: **Left:** Obsolescence time τ extracted for each bin of patents with fixed number of citations k from fitting the citation rate to the form of Eq. (4.5a) for times $t \geq t_0 = 2$ years. **Centre:** Preferential-attachment exponent α , extracted for fixed times t from fits of the citation rate to the form of Eq. (4.5b). **Right:** Constant contribution f_0 to the preferential-attachment part of the citation rate, extracted for fixed times t from fits to Eq. (4.5b). Circles are fit-parameter values, solid lines their weighted averages, and black dashed (red dotted) curves show 95% confidence intervals for fit-parameter values (weighted averages).

Figure 4.12: Disentangling preferential-attachment-type growth from ageing in the citation rate of patents granted by the USPTO. Rows represent technology categories while columns represent the parameter measurements.

Rows: Parameter estimates for the cohorts of patents assigned categories H (top) and Y (bottom).

Columns: **Left:** Obsolescence time τ extracted for each bin of patents with fixed number of citations k from fitting the citation rate to the form of Eq. (4.5a) for times $t \geq t_0 = 2$ years. **Centre:** Preferential-attachment exponent α , extracted for fixed times t from fits of the citation rate to the form of Eq. (4.5b). **Right:** Constant contribution f_0 to the preferential-attachment part of the citation rate, extracted for fixed times t from fits to Eq. (4.5b). Circles are fit-parameter values, solid lines their weighted averages, and black dashed (red dotted) curves show 95% confidence intervals for fit-parameter values (weighted averages).

Chapter 5

The Dynamics of Scientific Citations

5.1 Scientific Citation Dynamics

There is often a focus, especially in the network science literature, on the particular shape of the scientific citation distribution [138, 193–197]. This chapter, as for Chapter 4, will focus instead on the dynamics of scientific citations and assumes that the citation distribution is simply the result of microscopic network-growth mechanisms acting over time – it is these growth mechanisms that are of interest. The analyses presented in this chapter mirror those presented in Chapter 4 and act as a comparison between the dynamics of scientific citations and patent citations.

The analysis of scientific citation dynamics has a longer history and has generated a lot more interest than that of patent citation dynamics, and this is particularly true of network-science-based work. As such, the following sections will briefly review empirical work in scientific citation dynamics, and in particular the growth and ageing dynamics, from both a traditional bibliometric viewpoint and a network science viewpoint. The theoretical work outlined in Chapter 3 applies equally to scientific and patent citation networks, and will not be reiterated here. However, note that the network-science literature in this field tends to focus on modelling the mechanisms that advance the growth of the network rather than those that restrict growth (e.g. node ageing).

Gaining a full understanding of the dynamics of citation accrual by scientific articles and patents has been hampered by the need to account for various closely entangled, and

sometimes mutually counteracting, influences. This suggests a need to disentangle these influences before the real-life mechanisms behind them can be understood. This chapter serves two purposes: to check that the rate separability assumption applies to scientific works as it did for patents (see Chapter 4); and to directly compare the extracted rate parameters for obsolescence and preferential attachment between these two networks. Much of the discussion around the separability assumption in Section 4.1.3 is based on prior work on the growth of scientific citation networks. As such, that discussion is particularly relevant to this chapter as well.

5.1.1 Ageing

The obsolescence, or ageing, of scientific ideas has been of empirical interest since at least 1960 when the idea of a “half-life” of scientific ideas first appeared [198]. Since this time, two distinct but related literatures have developed with regard to the measurement of obsolescence [199–201]. The most common way to measure the growing irrelevance of ideas in most early work is known as the *synchronous* or *retrospective* approach (e.g. [171, 174, 202–204]), and involves the analysis of the age distribution of backward citations. This method was much more practical than the analysis of forward citations, known as the *diachronous* or *prospective* approach (e.g. [187, 205, 206]), in the absence of the large data sets available today – after all, it would be possible, given a set of articles, to construct the complete age distribution of the backward citations and gain a good understanding of the age structure of the field characterising the articles. On the other hand, an (unbiased) diachronous approach would require knowledge of all forward citations to an article from all possible sources – certainly a mammoth task without digital databases and the world wide web, and aspects of which are still problematic today.

One of the questions that is often asked in works studying obsolescence is the particular shape of the ageing function [34, 171, 204]. This depends critically, of course, on both how the data is processed and which of the approaches above (synchronous or diachronous) is used – the various choices of which can produce decays that are exponential [36, 126, 170, 171, 207], log-normal [34, 187], power-law [32, 38, 174], Weibull [37, 208] or combinations of these [192]. To address the second point of difference, Yin & Wang [34] conduct an extensive literature review of both approaches and conduct their own analysis in order to consolidate the myriad conflicting results. Once the distributions of citation

ages had been normalised to account for changing numbers of potentially citing/citable papers, the authors found that a log-normal decay in the probability of citation worked well for both synchronous and diachronous measurements. The results discussed in this chapter are not necessarily inconsistent with this finding, as the articles examined in this chapter are relatively homogeneous in publication time and more field specific than those used in Ref. [34], which are an aggregate of many different academic fields which may differ considerably in their ageing characteristics. The modelling approach detailed in Section 5.4 reflects the fact that the use of a model with more than one parameter would not lead to interpretable results, and in any case would not be directly comparable to the results for patents described in Section 4.5. Additionally, while the authors of Ref. [34] note an exponential increase in publications in their data, this is not observed here and there is in fact distinctive variation in the number of publications by particular journals that can be traced to changes in journal policy (see Section 5.2.3).

It is very difficult to empirically test models of real phenomena that lead to the observed ageing functions. This is perhaps not surprising, as the citation process, the various motivations for citation, and the diffusion mechanisms by which ideas spread are all acting at the same time; these effects may interact with one another, and data that can be used to investigate them may be hard to come by.

Very little has been done to incorporate obsolescence into growing networks in a way that is substantively different from the traditional approaches; after all, explicit and global node ageing is *not* a network-exclusive phenomenon. This is because obsolescence, as defined here and in many other works (e.g. see the review in [34]), is generally assumed to be indiscriminate with regard to nodes in different local network environments. In most cases where a citation model is dependent on network structure, ageing is built in as a side-effect of the model construction (e.g. structural ageing, as discussed in Section 3.2.5), rather than being dependent on any individual-node-specific properties.

However, while not the goal of this thesis, there are many exciting avenues of research to be undertaken in this area that do integrate network phenomena into ageing dynamics or attempt to model real obsolescence mechanisms. These include the integration of social or collaboration networks as conduits for knowledge diffusion [209], models that incorporate network-based node properties such as centrality measures [210] or copying mechanisms [138], and the application of sociological [211–213] or economic principles [214–216] to explain specific behaviours.

While there is still much interest in the ageing of scientific literature from a traditional bibliometric standpoint, the focus appears to have shifted somewhat to the categorisation of article citation histories in order to find so-called ‘sleeping beauties’ or ‘flashes in the pan’ and identifying the differing characteristics of the articles in these categories [217–222]. This research is not very relevant to the work in this chapter and is not discussed further.

5.1.2 Preferential Attachment

Preferential attachment is the network-science version of the concept variously known in other fields as cumulative advantage [116], the Matthew effect¹ [118], or simply ‘the rich get richer’ and ‘success breeds success’ [223]. In the context of bibliometrics, cumulative advantage was first investigated by Price [116], building on mathematical models developed earlier by Yule [224] and Simon [117]. In this model, Price makes the reasonable assumption that the probability of an article gaining a new citation is linearly proportional to the number of citations already obtained, with an initial attractivity of unity chosen for practical reasons. The resulting citation distribution is shown to be in the form of a Beta distribution, which was consistent with real citation data.

Barabasi & Albert [1] then independently rediscovered a special case of this mechanism as a feasible explanation for the power-law distributed tails of the degree distributions observed in many real networks. The authors found that a linear preferential attachment mechanism results in a power-law distributed degree distribution with exponent equal to -3 at long times.² This (very popular) paper came very shortly after a study of the citation distribution that observed a power law distribution with exponent -3 [225], suggesting that citations too, while not considered in the paper of Barabasi

¹The Matthew effect takes its name from the Bible verse Matthew 25:29, which states “For unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken away even that which he hath.” Note that the preferential attachment mechanism as defined in the network science literature generally does not remove links, as implied by the last clause of this verse.

² Note that this distribution was the result of a mean-field approach, mathematically equivalent to a model with continuous node degrees [114], and is therefore an approximation to the degree distribution for networks with discrete node degrees (including the distributions of the networks the authors were trying to account for). This issue was quickly noticed and rectified by two groups simultaneously [113, 114].

& Albert [1], might be generated by a process resembling preferential attachment. Of note is that neither of these papers cited Yule, Simon, or Price, the latter of whom was a pioneer in the study of citation distributions (e.g. [105]). Surprisingly, the fact that preferential attachment was a rediscovery was not acknowledged in the literature until more than three years after the publication of Ref. [1] by White [226], a mathematical sociologist. In the physics literature this took longer to acknowledge; the first physicist (a group within which network science was quickly gathering momentum) to cite both Ref. [1] and Price’s original paper [116] was Newman [227], 4 years after the publication of Ref. [1], by which time this paper had already garnered over 1000 citations.

Models of preferential attachment in citation networks, however, don’t often stray from the model proposed by Price [116] – a paper in which the ‘citation network’ was used as a descriptor for the system under examination and no network-specific mathematical tools were implemented. Even in his paper “Networks of Scientific Papers” [105], Price simply used the concept of a network as a mental framework for his analysis. We follow this mode of thinking in both this chapter and the previous. For example, the preferential attachment mechanism defined in this thesis is only augmented by two parameters: an initial attractiveness in the spirit of Price [116], and an exponent to allow for non-linear growth (although this exponent is still generally close to unity and, therefore, the linear case). Note that preferential attachment (as for node ageing) as defined in this thesis is a mechanism that applies equally to all nodes regardless of any local factors such as the characteristics of a node’s neighbours or its community membership – by definition this means that we don’t actually require the network framework for these analyses. There are many toy models that can imitate the preferential attachment effect without the need for a network framework, including the Pólya urn model [228] or the Chinese restaurant process [229, 230]. A useful survey of generative models of this type can be found in Ref. [231].

After its rediscovery in the era of accessible, digital datasets, preferential attachment proliferated in many new and exciting fields, including the field of network science as applied to bibliometrics. The new resources available in this field were ripe for testing new algorithms and statistical methodologies – this is perhaps unsurprising given that the scientific citation network was something that researchers were already embedded in, both professionally and literally. The process of scientific citation was not foreign to those who wished to study it, and there was already a large body of work in bibliometrics, pioneered by the likes of Garfield [103], Merton [118], and Price [232] on which to draw

inspiration and with which to compare results. The ‘new’ science of networks [233], however, did offer unexplored avenues of research that depended on the presence of network structure, such as collaboration networks [234], co-citation [235], homophily [236], and clustering [237]. Without imagining scientific articles as being embedded in a growing citation network, the analysis and measurement of these phenomena would be difficult, if not impossible, to conceive.

5.2 Scientific Citation Data

5.2.1 The American Physical Society Dataset

The bibliometric and citation data set used in this chapter is provided by the American Physical Society (APS) and, in its entirety, consists of article metadata and citation pairs dating back to 1893 [238]. The subset of this data set that is focused on here is the cohort of articles published in the years 1999–2001 (the justification for using this particular subset can be found in Section 5.2.3) in the research-field-specific APS journals Physical Review A, B, C, D, and E (from this point onward abbreviated as PRA, PRB, etc.), as well as the APS’s multidisciplinary-physics letters journal Physical Review Letters (PRL). Secondary cohorts from earlier times are also analysed for temporal comparisons to this focal cohort, and include articles published in the same journals (excluding PRE) from 1978–82 and 1988–92. These earlier cohorts are discussed further in Section 5.2.3.

Citation rates are measured using all citation pairs whereby the cited article in one of these cohorts is linked to a citing article published in the years 1999–2015 in any APS journal. For all of the specialized journals (such as PRA–E), the fraction of citations originating from articles published in the same journal is quite high, justifying the approach to use these journals to be representative of different research fields. As expected, this is not the case for the multi-disciplinary letters journal PRL, which is included as a benchmark for useful comparison.

There are a number of benefits and limitations presented by this dataset. The primary advantage is that, while limited, this data does allow for some degree of quality control on the citing articles due to the prestige of these journals, and therefore citations to a particular article may constitute a reasonable measure of its impact within

the physics community. The main limitation, which may also be framed as a convenient filter, is that only citations between APS articles are included, and, therefore, we cannot measure the impact of articles in the dataset outside of APS journals. This fact particularly impacts the more interdisciplinary journals such as PRB and PRE, where much of the impact may be in either specific related fields that are not covered by APS journals such as physical chemistry (in the case of PRB) or in non-physical science fields generally (in the case of PRE).

5.2.2 A Note on Scientific Literature Classification Schemes

There are many ways to classify scientific literature. As mentioned above, the informal classification system used in this chapter is simply defined by the scope of the different APS journals, as listed in Table 5.1. There exist more systematic, hierarchical classification systems such as the Physics and Astronomy Classification Scheme (PACS) that may be preferable to the simple groupings used here, however this data is not available for this dataset as PACS is no longer maintained. In any case, the data is limited enough without breaking it down further into smaller groups for analysis and therefore the use of PACS would likely impact the accuracy with which network growth parameters are measured, and therefore severely restrict the conclusions able to be drawn from such an analysis. Note that a system like PACS would be preferable with a larger data set; however with regard to the cohorts examined in this chapter, further disaggregation would severely limit meaningful interpretation of results due to lack of data.

5.2.3 Cohort Selection

In the case of the cohort selected to analyse the growth of patent citations in Section 4.2.3, three criteria were considered: cohort size, time to accrue citations, and the minimisation of changes external to the invention process that affected patenting or citation behaviour. The first two criteria are just as applicable to scientific articles as they are to patents, however, the last criteria is less relevant. As scientific articles and their citations do not have legal consequence in the same way that patents and their citations do, there are no procedural changes originating outside APS itself that directly changes citation behaviour. As changes in the way research is evaluated may change publication behaviour [239], however, it remains prudent to restrict the cohort length to mitigate the effect of long-term systematic variation caused by these policy changes. Additionally, as

Journal	Fields Covered	Published from
Physical Review Letters	All Physics	1958
Physical Review A	Atomic, molecular, and optical physics, and quantum information	1970
Physical Review B	Condensed matter and materials physics	1970
Physical Review C	Nuclear physics	1970
Physical Review D	Particle physics, field theory, gravitation, and cosmology	1970
Physical Review E	Statistical, nonlinear, biological, and soft matter physics	1993

Table 5.1: Descriptions of the journals that make up the cohorts whose citation growth properties are analysed in this chapter. Note that incoming citations are *not* restricted to this group, and may originate from any journal published by the American Physical Society.

each journal published by APS has relative autonomy over any guidelines that may effect publication or citation procedure (e.g. [240, 241]), it is not possible to control for these small changes by choosing a cohort during which time no significant changes occurred. Therefore, for the same reasons described in Section 4.2.3, the cohort 1999-2001 was chosen for the analysis in this chapter – this balances the need for a large data set with the length of time required to analyse the dynamics of citation accrual while allaying the effect of secular changes in publication behaviour.

For comparison, two older cohorts are also included for analysis in this chapter, articles published in the same journals (excluding PRE³) between 1978-82 and between 1988-92. The time intervals between each cohort is therefore about 10 years, which facilitates comparison of temporal changes between cohorts. Table 5.2 shows the cohort statistics for all cohorts, wherein it is clear that the 1988-92 cohort has a comparable number of articles to the latest cohort, while the number of citations per article is much smaller. As all of the analyses in this chapter are dependent on being able to measure

³PRA was originally subtitled “General Physics”, and was split into the current versions of PRA and PRE in 1993. PRA in the two older cohorts can therefore be viewed as the combination of PRA and PRE as described in Table 5.1.

Year	Journal	Total Articles	Total Citations	Mean Citations	Median Citations
1999-2001	PRA	4 449	40 850	9.2	4.2
	PRB	14 645	134 844	9.2	4.5
	PRC	2 635	26 394	10.0	5.3
	PRD	5 885	68 909	11.7	5.7
	PRE	6 565	40 066	6.1	3.2
	PRL	9 334	192 903	20.7	11.1
1988-92	PRA	9 519	51 017	5.4	2.7
	PRB	19 189	128 918	6.7	3.1
	PRC	3 617	19 137	5.3	3.1
	PRD	5 232	27 897	5.3	2.4
	PRL	9 598	127 883	13.3	6.7
1978-82	PRA	3 571	15 120	4.2	2.4
	PRB	7 820	40 503	5.2	2.5
	PRC	3 221	18 994	5.9	2.9
	PRD	4 291	25 040	5.8	2.2
	PRL	5 786	52 685	9.1	5.2

Table 5.2: Summary statistics for the three cohorts of APS-published articles analysed in this chapter. Citation statistics are calculated from the inflation-adjusted citations (from all APS journals) to each cohort 14 years after the publication of the cited articles within the cohort.

citation rates, lower citation rates limit the accuracy and precision with which the network growth parameters are able to be measured. The 1978-82 cohort has this limitation as well, while at the same time having half as many articles, compounding this issue. Therefore, while these earlier cohorts are included for comparison to the 1999-2001 cohort, only the large and consistent differences between cohorts will be noted.

As for patent citations, changes in publication rates are controlled for (as detailed in Section 5.3.2), while changes in the average length of reference lists are not – and there have been few specific changes to APS procedures or author guidelines that would change the number of scientific publications that are cited. The notable exception to this was a change to the length limit for PRL articles, which switched from a four-page limit to a 3750 word limit in July 2011, and removed the need to exclude pertinent references if more space was required for the main body of the article. This resulted in a marked increase in the number of references listed in articles published in this journal

Figure 5.1: The mean number of backward citations made by PRL articles by month between 1999 and 2015. In July 2011, PRL switched from a page-limit to a word-limit, allowing for longer reference lists and leading to a clear change in citation behaviour due to this change in publication policy.

after this time, see Fig. 5.1. The large external change that may be of concern is the shift to digital search for literature, which one would expect to result in longer reference lists due to the increasing ease of discovery of relevant work [242]. This is not controlled for in this chapter, and in any case, it is not clear that this would have a large impact on internal APS citations.

5.3 Data Preparation

The data preparation for scientific articles is nearly identical to the data preparation for patents, as the same analysis is conducted on each set of documents. Therefore, with the exception of the minor differences discussed in the following sections, the procedure for processing the APS data is the same as the procedure detailed in Section 4.3.

5.3.1 Redefining the Citation Lag

For the analysis of patent citations, it was important to define the citation lag as the time between the grant of the cited patent and the application of the citing patent (see Section 4.3.1). This was because the application date is closer to the date of invention, and therefore the date that the information contained in the cited patent was built on or otherwise utilised to invent something novel. Additionally, the grant lag is not only very long, but extremely variable from patent to patent.

While the APS dataset does not include the dates that each article was submitted to the journal in question, which would be ideal, the lag between submission and publication is nowhere near as long as for patents. Therefore, while the use of publication dates is the only option to define the citation lag, this is unlikely to have a large impact on the results of subsequent analyses. Additionally, as many physics papers are published on a preprint server before being published by a journal, much of the information content of the published articles are disseminated to the community before their publication. This allows other researchers to utilise this information and publish their own articles sooner than would be expected if, in analogy to patent grants, publication in an APS journal was the first time the work was made public.

All analysis and results in this chapter, therefore, have the citation lag defined as the time between the publication of the cited article and the publication of the citing article. Note also that because all citations are internal (they do not include citations to preprints or other non-finalised works), all citation lags are positive, unlike the patent case. It follows that all articles have zero citations at $t = 0$, and the initial citation distribution is thus simply $n_0(k) = \delta(k)$.

5.3.2 Citation Inflation

Both the number of articles produced and the average number of citations made by articles vary over time. In the long term, the combined effect of these factors causes a citation inflation that can mask the trend of obsolescence. As previously, we consider the changing rate at which articles are published an exogenous factor, as such changes in research productivity can be expected to be largely determined by the availability of resources, general policy decisions, or other influences that do not reflect the utility of prior knowledge. In contrast, any change in the average number of citations made

by each article is indicative of the need to cite more or less of the currently relevant knowledge and, thus, is intrinsic to the information ecosystem.

The differences between citation inflation as detailed here and in Section 4.3.2, therefore, constitute a simple change in the meaning of the notation (emphasised in *italics*) to reflect the different data used in each analysis. Slightly modifying Eq. (4.3), we can account for the growth of published articles and therefore a changing probability of citation due to the purely exogenous forces driving these changes by rescaling citations such that a citation from article i published in journal J in the time period $[T_i, T_i + \Delta T]$ has the value

$$c_i(T_i; T_0, \Delta T) = \frac{N_J(T_0, \Delta T)}{N_J(T_i; \Delta T)} \quad , \quad (5.1)$$

where $N_J(T_i; \Delta T)$ is the number of *articles published* in the time interval $[T_i, T_i + \Delta T]$ in the *journal* J , and $N_J(T_0, \Delta T)$ is the number of *articles published* in the time interval $[T_0, T_0 + \Delta T]$ in the same *journal* J . Note that the averaging process used in the case of patent citations is not necessary for scientific articles as each article can only be published in one APS journal. The variations in the value of citations originating from different journals is shown in Fig. 5.2.

5.4 Modelling Strategy

As per the previous chapter, the simultaneous influences of knowledge-diffusion-driven growth and obsolescence-related decay on the citation rate can be captured by postulating the functional form $\bar{\lambda}(k, t) = A(t) f(k)$, where $A(t)$ is a purely time-dependent ageing function, and the growth kernel $f(k)$ embodies the PA mechanism. The time variable t here represents the time in years from publication of the articles in the cohort (c.f. T in Eq. (5.1) which denotes calendar time).

We again observe that the observations are well described by the functional forms

$$A(t) = A_0 \exp\left(-\frac{t}{\tau}\right) \quad \text{for } t \geq t_0 \approx \tau/2 \quad , \quad (5.2a)$$

$$f(k) = k^\alpha + f_0 \quad , \quad (5.2b)$$

where A_0 , τ , α , and f_0 , are parameters controlling scale, characteristic obsolescence time,

Figure 5.2: The value of patent citations, $c_i(T_i; T_0, \Delta T)$, for different journals published by the APS over the course of the 14-year period during which citation accrual is observed for the 1999–2001 cohort, calculated as described in Eq. (5.1). Not shown are the smaller journals published by the APS – citations from these journals are still counted if they cite an article in one of the cohorts described in Section 5.2.3, however, the six journals shown here represent 98.3% of the citations to the 1999-2001 cohort and an even larger fraction of the earlier cohorts.

preferential attachment strength, and initial attractive respectively. Examples of the fits to these forms can be seen in Fig. 5.3. The use of these particular functional forms, rather than the alternatives described in Section 4.5.4, allows the direct comparison of citation dynamics between patents and scientific articles.

The fitting of these parameters is done in the same way as described in Section 4.4 (two-step weighted regression with logarithmic binning), with one important difference. The number of articles in each journal’s cohort is an order of magnitude smaller than the number of patents in their respective cohorts. To make the most of the data available, therefore, it is possible to change the bin sizes in both the time and k dimensions of the analysis. This means that when fitting $A(t)$ for the groups of articles in particular fixed- k bins, we can more accurately represent the data by slightly increasing the size of the k bins such that more articles are included in each individual fitting procedure for $\bar{\lambda}(k, t)$ at fixed k , while at the same time decreasing Δt for greater time resolution. Due to the finite range of empirically available k values, this means there are fewer fixed- k fitting procedures and thus fewer measurements of τ and A_0 ; however, we do not require

Figure 5.3: Fitting the time and citation-number dependencies of the empirical citation rate for the journal PRD. **Left:** Mean number of additional citations, $\Delta k(t + \Delta t)$, as a function of time t for fixed citation bins with log-midpoints at $k = 4, 16$, and 59 . Hollow symbols are derived from the citation data, while lines represent the fit to the form of $A(t)$ given by Eq. (5.2a). Deviations from exponential-ageing behaviour occur only at short times, and are greater than those observed for patents (see Section 4.5). **Right:** Mean number of additional citations, $\Delta k(t + \Delta t)$, as a function of cumulative number of citations k for various fixed times $t = 4, 8$, and 12 years. Hollow symbols are derived from the citation data, while lines represent the fit to the form of $f(k)$ given by Eq. (5.2b).

a large number of measurements to detect any systematic change in these parameters with k . The opposite is true when fitting $f(k)$ – smaller k bins can be compensated for by a longer time period (Δt) over which the effective citation rate is measured. Based on such considerations, we have chosen $\Delta t = 6$ months when fitting $A(t)$ and $\Delta t = 1$ year when fitting $f(k)$.

Also worth noting is that due to the lack of a non-trivial initial citation distribution, it is not possible to model the preferential attachment mechanism at time $t = 0$, as there is only one possible value of k at this time.

5.5 Results & Discussion

In the analysis of patent citations in Chapter 4, the significant changes in patenting behaviour spurred by legislative and procedural changes external to the inventive process [71] meant it would be difficult to directly compare patents separated by long times. This is not, however, as much of a problem for scientific articles, and particularly so for

such a select sample from high quality journals. It is possible, therefore, to investigate systematically the quality of separability and determine the values for parameters in the ageing function and the preferential-attachment-growth kernel characterizing the citation dynamics of articles published in different years. While the results for the 1999-2001 cohort remain the most statistically robust and are therefore the most appropriate to compare with the results for patents in Chapter 4, results from earlier cohorts are also reported in this chapter. The cohorts chosen for these analyses are defined by a longer time period of 5 years, and are roughly ten and twenty years prior to the focal cohort of 1999-2001: 1988-1992 and 1978-1982. These cohorts are analysed for comparison with the focal cohort, and while they are not ideal for this analysis due to the reasons discussed in Section 5.2.3, there may be some useful information to gain through these comparisons.

Most of the figures found in the main body of this chapter show the results for the journal PRD, a journal publishing research in the fields of particles, fields, gravitation, and cosmology. Research in these fields is unlikely to have a large impact in domains other than physics, and thus this journal represents a relatively homogeneous dataset with regard to citation dynamics. Citations to articles published in PRB are also briefly examined, as these articles represent a much more field-heterogeneous set than those published in PRD. Results for all other journals in the 1999-2001 cohort can be found in Figs. 5.6 & 5.7.

5.5.1 Preferential Attachment

As displayed in Table 5.3, the fits of the PA-mechanism growth kernel given in Eq. (5.2b) to the data yield values of the exponent α that vary moderately across the different APS-journal cohorts and are generally consistent with a superlinear dependence on k ($1.0 < \alpha \lesssim 1.2$) – broadly the same as the measured values of α for the patent cohorts. Additionally, the estimated values of the parameter related to initial attractiveness, f_0 , are around unity, again the same as for the patent cohorts. These similarities are somewhat surprising, given the vast differences in the meaning and process of citation between scientific articles and patents, this is discussed in more depth in Section 5.6.1.

In comparing the three cohorts, the preferential attachment strength appears to be shrinking with time; however, further research is required to estimate the statistical sig-

Year	Journal	α	f_0	τ [years]	A_0 [years ⁻¹]
1999-2001	PRA	1.17 ± 0.05	1.25 ± 0.58	4.63 ± 0.72	0.27 ± 0.08
	PRB	1.13 ± 0.10^a	1.11 ± 0.42	4.44 ± 0.94^b	0.36 ± 0.09
	PRC	1.11 ± 0.10	1.23 ± 0.91	5.74 ± 1.51	0.23 ± 0.05
	PRD	1.11 ± 0.06	1.11 ± 0.85	5.00 ± 0.90	0.21 ± 0.04
	PRE	1.03 ± 0.16^a	0.77 ± 0.41	4.70 ± 1.60	0.37 ± 0.16
	PRL	1.06 ± 0.11^a	1.07 ± 0.55	4.71 ± 0.98^b	0.41 ± 0.15
1988-92	PRA	1.22 ± 0.07	0.86 ± 0.29	5.22 ± 0.53	0.17 ± 0.03
	PRB	1.18 ± 0.06	0.82 ± 0.70	5.74 ± 0.73	0.16 ± 0.01
	PRC	1.19 ± 0.10	1.07 ± 1.48	6.03 ± 1.10	0.13 ± 0.03
	PRD	1.21 ± 0.10	0.71 ± 0.43	6.34 ± 1.08	0.11 ± 0.02
	PRL	1.11 ± 0.08	0.66 ± 0.34	5.53 ± 0.71	0.20 ± 0.02
1978-82	PRA	1.20 ± 0.16	0.95 ± 0.61	7.56 ± 1.28	0.10 ± 0.01
	PRB	1.25 ± 0.07	1.01 ± 0.42	5.44 ± 0.75	0.13 ± 0.01
	PRC	1.28 ± 0.13	1.39 ± 1.44	5.25 ± 0.81	0.13 ± 0.02
	PRD	1.26 ± 0.14	0.85 ± 0.48	6.38 ± 2.80	0.09 ± 0.02
	PRL	1.10 ± 0.10	0.76 ± 0.38	5.76 ± 1.54	0.16 ± 0.02

^a Here α showed a moderate residual time dependence.

^b Here τ showed a moderate residual dependence on the cumulative number of citations.

Table 5.3: Measured values for the parameters characterising preferential attachment (α and f_0) and obsolescence (τ and A_0), extracted by analysing citations to articles published in APS journals in the years 1999–2001. For comparison, results obtained from performing the same analysis on citations to articles published in these journals during the five-year periods 1988–1992 and 1978–1982 are also given. Uncertainties represent 95% confidence intervals.

nificance of this trend for any particular journal as the methodology used here is not designed to answer this question. Additionally, the weak residual dependencies on time observed for some journals in the 1999-2001 cohort are not present for the older cohorts. This is consistent with the view that citation behaviour may be changing considerably with time. The parameter α reflects the fact that the probability of gaining a new citation increases with an increasing number of previous citations. During periods when it was difficult to know how many citations a particular article had, the preferential attachment effect was most likely driven purely by local network growth rules, as the decision to cite a particular article could not have been driven by global information. Whether the decrease in preferential attachment strength was driven by changes in ci-

tation strategy, knowledge diffusion, or ease of search is an avenue for future research. These effects will be difficult to untangle and lead to different conclusions about the meaning of citations. Perhaps science is becoming more meritocratic in its recognition of quality work, or maybe increasing numbers of publicity channels may be leading to greater speed and extent of knowledge diffusion. It is even possible that keyword search could be leading to citation mechanisms that are starting to weight similarity of content over quality of content. Note, interestingly, citations to articles in PRL do not display a change in preferential attachment strength, and therefore may be used as a reference point for any hypothesis put forward with regard to these changes in citing behaviour.

5.5.2 Obsolescence

The obsolescence-related ageing function is observed to be an exponential function of time since publication in the long term, as expressed in Eq. (5.2a) and shown in Fig. 5.3. This observation broadly agrees with recent studies of larger scientific-article cohorts [206,243], although a slightly shorter value for the obsolescence time scale τ is found in our research-field-specific analysis and some variation of this parameter between the different fields is also observed (see Table 5.3). In particular, the subfields associated with PRC and PRD appear to be more slowly changing than those covered by the other three specialised journals that turn out to have a similarly short obsolescence time as the multidisciplinary journal PRL. While not fully conclusive because of the overall scale of uncertainties in the extracted τ values, this observation is consistent with expectations based on known characteristics of the research fields covered by PRC and PRD, especially a dependence on the long-term development of large-scale equipment run by very large consortia of researchers.

Deviations from exponential ageing occur at short times $t \lesssim 3$ years, with more citations being accumulated per unit time than expected from an extrapolation of the long-term ageing trend, similar to the excess of citations arriving in a short time period after grant for some patents (see, for example, Fig. 4.6). The ubiquity of this short-term enhancement suggests the existence of a common origin related to knowledge-flow dynamics, which should be clarified by systematic further studies of the citation-number dependence of the excess number of citations, as well as possible systematic variations across different research fields or technology categories. Interestingly, the relative magnitude of the short-term deviation from exponential ageing exhibited by low to medium-

cited scientific-article cohorts is consistently observed to be larger, by at least a factor of two, than for the comparable cohorts of patents.

The two older cohorts display much larger characteristic obsolescence times than the 1999–2001 cohort, and don’t exhibit any residual dependence on k . Additionally, both older cohorts have approximately the same estimated τ values across most journals (with the exception of PRA). This observation would appear to imply that, after correcting for the variable numbers of published papers, the most significant changes in ageing dynamics took place in the 1990’s – the decade that the current modes of information transmission (e.g. the World Wide Web) took shape.

In addition to the changes in obsolescence rate, small changes in the form of obsolescence may also be occurring over the time periods examined. Due to deviations from exponential decay at short times, the fitting of this decay requires a choice of starting point $t_0 > 0$. For the 1999–2001 cohort of articles, a good fit is found for $t_0 = 3$ (see Fig. 5.3), however for the earlier cohorts this is increased to $t_0 = 4$ to accommodate an increased persistence of the deviation. This may be, however, simply a reflection of the more slowly decaying obsolescence function found for these cohorts, as these two time scales are generally found to be linked. This is also consistent with the hypothesis that the nature of knowledge diffusion changes over time, and that the mechanisms that characterise knowledge diffusion shortly after publication may persist for a longer time in the earlier cohorts. Some journals in the 1978–82 cohort additionally appear to exhibit greater than expected citation rates for long times as well as short times, suggesting that exponential obsolescence may not be the best functional form for this time period. Due to limited data for this cohort, less parsimonious models are not likely to provide a better fit (even with a large amount of data, adding one extra parameter increases parameter uncertainties drastically, see Fig. 4.9), however this may be worth investigating in more depth in future work.

5.5.3 Evidence for Rate Separation in Citations to Scientific Articles

The present approach is designed to carefully disentangle the mechanisms leading to PA-driven citation accumulation from the effects of ageing and inflation – a strong assumption empirically untested in the scientometric literature until this point. As previously described in Section 4.4, the litmus test for having achieved this goal is provided

by the absence of residual time dependences in PA-related parameters that have been extracted from fits of the empirical citation rate. Furthermore, the parameters governing obsolescence-related ageing should be found to be independent of the number of citations, as is indeed the case for most journals at most times. The results for PRD can be seen in Fig. 5.4 for all three article ages, results for the 1999-2001 PRB cohort can be seen in Fig. 5.5, and results for the 1999-2001 cohorts published in all other journals can be seen in Figs. 5.6 & 5.7.

To be able to demonstrate the clear separation observed for most cohorts, we needed to focus our analysis on specific research fields in physics, as defined by the scope of individual APS journals, and account for citation inflation by a journal-specific scaling factor. In contrast, previous studies that did not separate articles by research fields and did not account for citation inflation found significant increases over time in the extracted values of α [38]. We observe some deviations from full separability of the empirical citation rate into independent ageing and PA parts for the cases of PRB (see results presented below and Fig. 5.5), PRL, and PRE, which are all journals that publish articles from a much broader and more heterogeneous range of physics subfields, and even from neighboring disciplines such as chemistry and mathematics, than the other journals. Additionally, these deviations are not observed for the earlier cohorts, which suggests that these deviations from separability may be due to relatively recent changes in citation behaviour. Our results suggest, therefore, that the dynamics of knowledge diffusion and intrinsic obsolescence of knowledge is not only research-field-specific, but may change significantly between cohorts published at different times.

The most drastic deviation from separability occurs for PRB, the APS journal for condensed matter and materials physics, for which both α and τ showed residual dependencies on t and k , respectively, as shown in Fig. 5.5. The observed trend of increasing α as a function of t is slower than, but still of roughly the same order of magnitude as, in studies where articles were not disaggregated by research field and inflation was not accounted for [38]. The increasing tendency of more highly cited articles to attract further citations at a higher rate could reflect the greater importance of autocorrelations in the citation dynamics of the journals displaying this characteristic. For example, ‘sleeping beauties’ (articles whose importance is not immediately recognised) have been found to occur at different rates in different fields [221, 244], and are examples of situations when accounting for autocorrelations become important for the success of a citation rate model [38]. Alternatively, a greater heterogeneity in terms of research field and stronger

Figure 5.4: Disentangling preferential-attachment-type growth from ageing in the citation rate of articles published in PRD in different three time periods. Rows represent cohorts while columns represent the parameter measurements.

Rows: Parameter estimates for the cohorts of articles published in **Top:** 1999–2001, **Centre:** 1988–1992, **Bottom:** 1978–1982.

Columns: **Left:** Obsolescence time τ extracted for each bin of articles with fixed number of citations k from fitting the citation rate to the form of Eq. (5.2a) for times $t \geq t_0 = 3$ years for the 1999–2001 cohort and $t \geq t_0 = 4$ years for the two earlier cohorts. **Centre:** Preferential-attachment exponent α , extracted for fixed times t from fits of the citation rate to the form of Eq. (5.2b). **Right:** Constant contribution f_0 to the preferential-attachment part of the citation rate, extracted for fixed times t from fits to Eq. (5.2b). Circles are fit-parameter values, solid lines their weighted averages, and black dashed (red dotted) curves show 95% confidence intervals for fit-parameter values (weighted averages).

multidisciplinary influences from fields outside physics that characterize PRB, PRE, and PRL could be the cause. Support for this conclusion is also provided by the results in Section 4.5 where deviations from separability of the empirical citation rate also occurred for the more heterogeneous technology categories. PRB, for example, publishes

Figure 5.5: Attempting to disentangle preferential-attachment-type growth from ageing in the citation rate of articles published in PRB in the years 1999–2001. **Left:** Obsolescence time τ extracted for each bin of articles with fixed number of citations k from fitting the citation rate to the form of Eq. (5.2a) for times $t \geq t_0 = 3$ years. **Centre:** Preferential-attachment exponent α , extracted for fixed times t from fits of the citation rate to the form of Eq. (5.2b). **Right:** Constant contribution f_0 to the preferential-attachment part of the citation rate, extracted for fixed times t from fits to Eq. (5.2b). Circles are fit-parameter values, solid lines their weighted averages, and black dashed (red dotted) curves show 95% confidence intervals for fit-parameter values (weighted averages).

large numbers of articles in both experimental and theoretical condensed matter physics; these two groups may not cite each other often, and may additionally display different citation dynamics that are unable to be untangled with the available data.

5.6 Conclusions & Outlook

On a qualitative level, it can be surmised that diffusion of knowledge drives the accumulation of citations by articles or patents that report useful new information. The ability to accurately model the dynamics of knowledge diffusion should be facilitated by the increased availability of high-quality citation data, if only the additional mechanisms affecting real-world citing behaviour could be reliably identified and accounted for.

Although the obsolescence-induced ageing is accurately described by an exponential function for intermediate to long times after publication, an excess of citations above the extrapolated exponential behavior occurs within the first 2–3 (4–5) years after publication for the articles in the 1999–2001 (1978–82 and 1988–92) cohort(s), with stronger deviations occurring for articles with smaller number of accrued citations. More systematic study is needed to determine the origin of these deviations, but their particular

features may point to the existence of a special knowledge-propagation mechanism or set of mechanisms that are effective at short times (such as self-citations or co-author citations), or perhaps latent quality that is only well-captured by citations after a few years for most articles (see Chapter 6 for an investigation of this latter phenomenon in the case of patents).

Preferential attachment in the form of Eq. (5.2b) fits the observations very well, and articles published in most journals at most times display citation growth parameters that are unchanging with time. The preferential attachment exponent α is consistently superlinear for all journals at all times, and tends to be larger for earlier cohorts, suggesting that at these earlier times citations were even more important in helping other researchers find (and subsequently cite) relevant research than they are today. A few journals in the 1999-2001 cohort (PRB, PRE, and PRL) showed slowly increasing values of α . The reasons for these systematic variations is not known, however it is noted that these journals in particular are much more likely to have influence in disciplines that are not represented by APS publications and, therefore, the internal APS citation data on which these analyses are conducted are unlikely to capture the true influence of papers published in particular sub-fields within these journals – this may lead to distorted citation counts.

As for the previous chapter, many of the methods detailed in this chapter were first published by the author of this thesis in Ref. [245]. That research examined smaller cohorts and did not contain any results from the 1978-82 period. The results here remain qualitatively similar to those described in Ref. [245].

5.6.1 Striking Similarities

Even though the mechanisms and motivations determining citing behavior of academics and inventors have been identified to be quite different, the results of our present study turn out to be strikingly similar, both qualitatively and quantitatively, to those found previously in Chapter 4. In particular, the values of the exponent α characterizing preferential-attachment-type growth vary over the same basic range of magnitudes between the different article and patent cohorts published/granted over the same period. The rate of obsolescence is observed to be slower for the patents compared to that of the APS-journal articles, whereas the magnitude of the citation-rate excess over the ex-

trapolated exponential-ageing behavior in the short term is systematically larger for the cohorts of articles than for the comparable patent cohorts. This last point is consistent with other recent results; for example, Mariani et al. [246] found that “patents’ citation dynamics is significantly slower than that of papers”, and that this makes prediction-related tasks much more challenging for patents than for scientific articles.

Are these similarities surprising? After all, there are major differences in the publication and citation processes that one would expect to introduce significant differences between their network growth dynamics. For the moment it will be assumed that, APS journals being of generally high quality, the articles represented in this chapter have been judged to meet some form of novelty, usefulness, and/or non-obviousness requirements similar to those defined for patents in Section 2.1.1, justifying the existence of these articles in the dataset and allowing for attention to be turned to the differences in citation practices (as opposed to publication practices).

The primary difference between the citation of articles and patents is author/inventor autonomy with regard to these citations. The inventor cites the prior art that they are aware of but these citations are culled and added to by an examiner in accordance with the policies of the patent office. The academic, on the other hand, has much more freedom in the works they cite, and this results in particular behaviours and norms that may be expected to leave a kind of signature on the resulting citation network that differentiates it from not only the patent citation network, but also citation networks in different fields.

Recent work on the transitive reduction of citation networks [95] (elaborated on in Section 3.3.1) has shown that these differences certainly exist on the mesoscopic scale – that is, on a scale where phenomena such as clustering need to be taken into account. However, the results in this chapter suggest that while these mesoscopic differences are significant, the global attachment mechanisms appear to be relatively unaffected. The mechanisms that lead to the preferential attachment effect in the scientific and patent citation networks are clearly very different (due to the differences in citation practice), however the parameters that describe the sum of these effects happen to be the same. There are moderating mechanisms common to both systems that prevent the strength of these mechanisms from increasing to the point of winner-takes-all, such as knowledge obsolescence and the distinct and finite domains of research outside which any work is unlikely to have influence. Both of these mechanisms act to limit the influence of a

particular piece of knowledge both in time and academic/technological scope. There are also mechanisms common to both systems that lead to important works generally being recognised as such, whether or not citations are accrued in proportion to this intrinsic value. These facts considered, it is perhaps less surprising that global network growth parameters are similar in both networks: their values likely sit at a kind of equilibrium point, where the underlying forces that lead to this equilibrium are common to both networks.

Yet to be discussed is the concept of citations as an indicator of the quality of the cited document. In the case of scientific citations, much work has been done to determine what citations are actually indicators of (see, e.g., Refs. [247, 248] for overviews of this field). Due to citation norms in science, it is difficult to treat citations in isolation as indicators for measuring the intrinsic quality of scientific work [249], because the motivations behind citations are not particularly clear for any individual citation. The meaning of patent citations, however, is much better defined, and this means a notion of quality can be defined in relation to them. Whether or not one agrees that citations are a good indicator of quality (however this is defined), it is difficult to argue that citations do not reflect, however noisily, technological importance (as discussed in Section 2.3.3). The next chapter will attempt to construct a flexible metric to capture the intrinsic importance of patents, and then examine the role of this so-called ‘fitness’ in the dynamics of citation network growth.

5.7 Additional Figures for Chapter 5

Figure 5.6: Disentangling preferential-attachment-type growth from ageing in the citation rate of articles published by the APS from 1999-2001. Rows represent journals (excluding PRB and PRD, for which the parameter estimates can be seen in Figs. 5.5 and 5.4 respectively) while columns represent the parameter measurements.

Rows: Parameter estimates for the cohorts of articles published in journals PRA (top) & PRC (bottom).

Columns: **Left:** Obsolescence time τ extracted for each bin of articles with fixed number of citations k from fitting the citation rate to the form of Eq. (5.2a) for times $t \geq t_0 = 3$ years. **Centre:** Preferential-attachment exponent α , extracted for fixed times t from fits of the citation rate to the form of Eq. (5.2b). **Right:** Constant contribution f_0 to the preferential-attachment part of the citation rate, extracted for fixed times t from fits to Eq. (5.2b). Circles are fit-parameter values, solid lines their weighted averages, and black dashed (red dotted) curves show 95% confidence intervals for fit-parameter values (weighted averages).

Figure 5.7: Disentangling preferential-attachment-type growth from ageing in the citation rate of articles published by the APS from 1999-2001. Rows represent journals (excluding PRB and PRD, for which the parameter estimates can be seen in Figs. 5.5 and 5.4 respectively) while columns represent the parameter measurements.

Rows: Parameter estimates for the cohorts of articles published in journals PRE (top) & PRL (bottom).

Columns: **Left:** Obsolescence time τ extracted for each bin of articles with fixed number of citations k from fitting the citation rate to the form of Eq. (5.2a) for times $t \geq t_0 = 3$ years. **Centre:** Preferential-attachment exponent α , extracted for fixed times t from fits of the citation rate to the form of Eq. (5.2b). **Right:** Constant contribution f_0 to the preferential-attachment part of the citation rate, extracted for fixed times t from fits to Eq. (5.2b). Circles are fit-parameter values, solid lines their weighted averages, and black dashed (red dotted) curves show 95% confidence intervals for fit-parameter values (weighted averages).

Chapter 6

Patent Quality and Network Growth

6.1 Patent Quality and Fitness in Networks

Significant attention has been given to trying to distinguish the extent to which the apparently self-reinforcing behavior of popularity is purely a property of the dynamic system, versus being generated by intrinsic heterogeneity that allows inherently better agents or products to persistently succeed [22,26,123,125,187,246,250–253]. This is not a trivial task, however, because oftentimes popularity is the only objective way to measure how something is valued by society. For example, it would be exceedingly difficult to judge how popular an anonymous person joining a social networking service will become, at the time they join. This is mostly to do with a lack of information – generally people are not subjected to, for example, a personality test as a condition of joining the network.

Patents, on the other hand, *are* subjected to a kind of personality test before they are granted: first through the invention and drafting process, then further clarified and refined through the examination process. By the time it is granted, a patent will contain an abundance of information about the invention it protects – before it joins the citation network, a detailed, vetted profile of the patent must be written and subsequently made public. Furthermore, the special purpose and associated legal ramifications of patent citations [15,95,185] make them particularly suitable for investigating the possible relationship between popularity and any particular qualities that can be extracted from the associated patent’s profile. This provides an opportunity to study the role of heterogeneity in quality or *node fitness* in growing networks in a way that is simply not

possible for most other real-world networks.

On a phenomenological level, purely preferential-attachment-based models seem to be able to successfully describe the dynamics of how patents receive forward citations. However, they are at odds with the general expectation [152, 166] that patents are intrinsically heterogeneous in quality, and that citing behavior will, at least in part, be influenced by this intrinsic heterogeneity across the patent population. In this chapter, ex-ante measurements of various dimensions of patent quality are extracted from patents at the time that they join the citation network, and this information is then integrated into the growing network model described in detail in Chapter 4 in order to understand the role of intrinsic patent heterogeneity and its interplay with preferential-attachment-like growth.

6.1.1 Measuring Patent Quality and Value

First, it is important to define what is meant by the terms *quality* and *value* in the context of this work. In this chapter, quality is viewed as a multidimensional description of the inherent characteristics of an invention, while value is something that can in principle be measured monetarily and may be associated with many different, technology-specific, aspects of patent quality. Value is mentioned here because this is often what economists are attempting to capture – a single figure that sums up the utility and demand for a particular invention that can be easily integrated into economic models. The attempt to obtain a single figure associated with each patent (to become known as *fitness*) is also the goal of this chapter, but with the above definition of quality in mind rather than monetary value.

Measurements of patent quality and value are of great interest to the economics community for a variety of reasons, and it is therefore important to give some context to this existing body of literature and justify the importance of this research today through two issues that currently motivate work in this field. First, the number of patent applications (and subsequent grants) has ‘exploded’ [72] over the past 30 years or so, and it is important to know whether any fraction of this rise is due to a genuine increase in research activity. After all, the same effect may result from a drop in the size of the inventive step required for patentability. If this latter scenario were true, it would place unnecessary financial burden on society due to the high costs of both the operation

of a government agency and deadweight costs due to inefficient allocation of monopoly rights that patents carry. Being able to objectively measure patent quality will provide insight into the truth of these scenarios. Second, patents are often used as a metric by which to measure innovative output, particularly on the scale of industries, cities, or countries [254]. It is well known, however, that patents are very heterogeneous in both value and quality regardless of how these quantities are measured [255, 256]. Therefore, simple patent counts will likely not adequately capture either the true technological output represented by any given set of patents, or the economic impact these technological outputs have¹.

The implications of variable patent quality and value has been studied for many decades [257], and the first publicly available, quantitative proxies for quality came in the form of patent citations [61, 89]. Citations have been very well tested as proxies for the technological impact [30, 89], societal impact [86], and research expenditure [67, 78] associated with particular inventions. Due to this history, and to their other useful traits such as availability, well-defined legal meaning, and quantitative nature, citations often play the role of a yardstick with which to measure the effectiveness of other indicators. In this chapter, citations constitute the links in the network being explored, and thus all measurements of quality will eventually be related back to their ability to explain the citation behaviour observed. The exact combination of quality dimensions that determine patent fitness in this context is, therefore, determined by the aspects of quality represented by forward citations.

Many of the indicators in the economics literature are *ex-post*, that is, they are determined after the grant of the patent they are associated with. These indicators are not useful to this work, for reasons discussed below in Sections 6.1.2 & 6.1.3, however there is good reason to associate them with quality generally. For example, a patent's quality may not be apparent for a number of years after its grant, and if simply obtaining a measurement of aspects of patent quality at a particular point in time after grant is the goal, then ex-post indicators can be very useful. However, for the case of static intrinsic quality as considered in this chapter, it is important to separate the information known at the time of grant with any new information that arrives after grant; this is discussed

¹Heterogeneity in units of observation is a problem encountered in many fields, particularly in the social sciences where measured outcomes (e.g. unemployment, productivity, or income) are often attributed to characteristics of individuals that may be very difficult to quantify (e.g. socio-economic background or health).

further in 6.1.2 & 6.1.3.

Many, if not most, works on the measurement of patent quality pertain to the dimensions that reflect the monetary value of a patent. The most common indicators, therefore, focus on explicit or implicit signals of patent value from either the firm that the patent is assigned to (or its shareholders), or competitors of this assignee. Value indicators originating from the patent owners include any activity that involves significant investment on the owner's part, such as patenting in multiple jurisdictions (family size) [84,258,259], paying maintenance fees to keep the patent in force (renewal) [260–262], paying extra fees for accelerated examination [263–265], and market-based approaches [67, 78, 266–268]. Among value indicators originating from competitors, the most prevalent are related to opposition (a post-grant patent validity dispute) [84,269] or simply litigation [79,270,271] – this is an expensive process for all parties involved and is not worth pursuing over a worthless patent.

Some research (including many of the works cited above) examines multiple dimensions of patent quality by making use of a number of different indicators [259,272–275]. Eventually, however, these indicators (or combination of indicators) must be validated against some well-established reference indicator(s) that depends on the quality dimension of interest, such as monetary value [276] or technological impact [277], or something more abstract, such as inventive step size [278]. This literature often includes a number of bibliometric indicators, including some of those listed in Section 6.2, and generally consists of comparisons of the performance of these indicators in explaining the variance in the reference indicator(s). This work is extended conceptually in this chapter and tailored to shine light on the specific network-growth dynamics being investigated in this thesis. A number of excellent reviews of patent quality measurements are available elsewhere [272, 279, 280], and the literature relevant to the indicators utilised in this chapter is included in Section 6.2.

6.1.2 Node Fitness in Growing Networks

Node fitness, in the network-science context, is usually a variable used to capture differences in nodes' intrinsic ability to attract links and is generally not influenced by dynamic properties of the node, such as degree. In principle, a fitness variable can be constructed from any quantifiable heterogeneous property, or a number of such properties, exhibited by individual nodes [281], and it can even be designed to depend on

the properties of the linking node [123, 281]. In theoretical models of network growth, node heterogeneity has been introduced by a fitness, or attractiveness, variable η with distribution $\rho(\eta)$ so that an individual node i having k_i links to other nodes at time t after its creation gains new links with a rate [191]

$$\lambda_i = \eta_i \tilde{A}(t) \tilde{f}(k_i) \quad , \quad (6.1)$$

where $\tilde{A}(t)$ and $\tilde{f}(k_i)$ are the ageing and preferential attachment functions observed once fitness is accounted for.

A rate of the form given in Eq. (6.1) also describes the situations where growth is only driven by preferential attachment [1] or purely by fitness [19] as limiting cases. In many real-world systems, however, it is likely that both preferential attachment and fitness are working simultaneously to produce the observed distribution of node degrees. This interplay has been studied in detail theoretically [123, 124, 282], and has been observed endogenously in real-world networks [125, 187, 282–284]. However, in the context of a growing network, an empirical determination of node fitness that is exogenous to the network itself has yet to be done – this would require both a large amount of information about the nodes at the time that they join the network, as well as a methodology that made it possible to associate this information with a measure of network growth in a way that does not implicate preferential attachment effects. Here we provide such a framework.

Another approach that has been suggested in the theoretical literature is that of additive fitness [285]. In Eq. (6.1), fitness acts multiplicatively on the preferential attachment function such that it does not lose relevance as citations increase. On the other hand, a rate function with an additive fitness of the form

$$\lambda_i = \tilde{A}(t) [p\tilde{f}(k_i) + q\eta_i] \quad , \quad (6.2)$$

would be able weight the relative importance of preferential attachment and fitness according to the parameters p and q (which may be related if model parsimony is desirable). One attractive feature of such a model is that it limits the effect of fitness as citations accrue – this may be appropriate in a system where pure preferential attachment is the dominant driver of citation growth for documents with many citations, the scientific citation network may exhibit such a property, for example. The problem with this kind of model, and with additive models generally, is that η does not have a predetermined

scale and so it is difficult to interpret the parameters empirically obtained in such a model other than the limit on the effect of fitness for highly cited documents. In Section 6.3.3, where a multiplicative model is used, it is possible to define η such that its value for each patent is relative to the value for the average patent; no such sensible definition is possible for additive models.

A multiplicative analogue of the above approach that is not considered here is one in which a exponential weighting parameter, γ , is introduced, as in

$$\lambda_i = \tilde{A}(t) \eta_i^\gamma \tilde{f}(k_i)^{(1-\gamma)} \quad . \quad (6.3)$$

This formulation would also have pure preferential attachment ($\gamma = 0$) and pure fitness ($\gamma = 1$) as limiting citation growth cases, at the expense of an additional parameter. However, this parameter will either interact with the strength of the preferential attachment effect (represented by α in previous chapters), or, in the simple case of $f(k) = k + f_0$, effectively replace it. When both fitness and preferential attachment only describe the average behaviour of a very noisy system, as is the case in this chapter, it is prudent to restrict the number of assumptions in, and complexity of, the citation growth model to prevent over-fitting and misleading results. This has the additional benefit of simplifying any analytical derivations, such as those presented in Section 6.3.3. For these reasons, Eq. (6.1) provides a satisfactory starting point for the following analysis.

6.1.3 Patent Quality as Node Fitness

A number of special properties of patents make them ideal candidates to study the effect of node heterogeneities on network growth dynamics. First, at the time of grant, a lot of information about the patent is available. Disclosure, after all, is one of the bedrock principles on the basis of which the modern patent system can function. As this information is determined before any fact about the patent are made public,² and thus falls into the category of *ex-ante* (literally “before the event”) information. This is a very important property, as intrinsic properties (as defined in this chapter) are not dynamic and are, therefore, defined at grant.

²However, all patent applications filed after November 2000 are made public after 18 months. This does not affect our cohort except in cases where early pre-grant publication was approved and published, and the patent subsequently examined and granted, within 12 months of application – likely a very rare event given an average grant lag of about 2-3 years.

Second, there exists a substantial lag between the application for a patent and its eventual grant. When the citation lag is defined as in Section 4.3.1 (the time between the grant of the cited patent and the application of the citing patent), citations can accumulate before the grant of the cited patent. This occurs when a citing patent is applied for between the application and grant of the cited patent. The citations obtained during this period are, therefore, almost exclusively self-citations and examiner citations, because the cited patent was not public knowledge at the time that the citing patent was applied for. This allows the use of these citations as indicators of quality that are mostly unrelated to knowledge diffusion processes (and therefore the pure attention-based preferential attachment processes) that occur once an invention is public knowledge, after which time inventions receive the vast majority of their citations. Pre-grant citations can, therefore, be used to judge the relative importance of the ex-ante information listed on the patent document in determining the effect of intrinsic patent qualities on knowledge diffusion.³ In turn, fitness as constructed in this chapter reflects the properties of patents that lead to pre-grant citations and, therefore, the aspects of quality that these properties capture — the question of whether fitness reflects commercial value, technological quality, social value, or some combination of these is not assessed in detail here and is an avenue for future work.

Of course, there are various potential dimensions of usefulness for an invention (including, but not limited to, technological, economic, and strategic/legal dimensions), and the number of suggested plausible quality-indicator variables has proliferated in the economics literature, as illustrated in Section 6.1.1. It is first necessary, therefore, to choose appropriate variables for the purpose of measuring different aspects of patent quality, before sensibly combining these variables to obtain a fitness parameter that can be integrated into a growing network model.

³One limitation of using citations at grant is that, in practice, patents do not all start gaining citations at the same time, at least in our framework – this effectively introduces a head-start for patents with longer pendency times, as these patents have a longer period during which to gain pre-grant citations.

6.2 Ex-Ante Determinants of Patent Quality

To obtain intrinsic (and static) patent fitness values, all variables considered here are based on ex-ante information – all of the following variables for a particular patent i are determined at the time of grant and become part of its retrievable official record. The raw variables on which all of the following are based are available for all patents,⁴ not just a subset such as those assigned to publicly traded companies [78], those that have been licensed for a known price [286], or those that have been publicly auctioned [268]. The discussions below will cover the ways in which each of the variables described may contribute to our knowledge of the intrinsic quality of any particular patent. At the end of this Section, Table 6.1 lists all variables described here and their abbreviations for easy reference.

Ex-post variables have been excluded due to their potentially dynamic nature – the value of a patent may change depending on events occurring or environments changing *after* grant – which results in a chicken-and-egg problem with regard to citations. For example, it is difficult to know whether there is a causal linkage between the litigation of a patent (indicating commercial value [79, 262, 271]) and the patenting of similar inventions (leading to citations) or whether some third factor suddenly renders a particular invention much more valuable, leading to both litigation and citations. By excluding all ex-post information, it is possible to avoid these endogeneity problems caused by simultaneity (the existence of feedback loops that may prevent us from identifying time-invariant intrinsic qualities).

Variables Related to Backward Citations

Backward citations are the citations a patent makes to related earlier patents and other sources of knowledge, and therefore contain not only information about the citing patent for which a fitness value is desired, but also information on the inventions or ideas on which the patent is dependent on or related to. (See Chapter 2.3.3 for a discussion on the meaning and utility of backward citations generally). The discussions below will focus on the many ways in which backward citations, and the information contained therein,

⁴Some of the constructed variables, however, are not defined for all patents. For example, the average age of a patent's backward citations is not defined for a patent with no backward citations.

may inform us about the quality of the patent making the citations.

From the perspective of the patent owners, the number of backward citations should be minimized for strategic reasons [15], as this would maximize the technological space this patent allows them to exploit (see Section 2.3.2). Therefore, one could expect a large number of backward citations to hurt the invention's prospects for economic-value generation. On the other hand, a large number of backward citations could also be an indicator for a technologically complex invention with large potential value, or equally plausibly signify a merely incremental inventive step. Empirically, there has generally been a positive correlation between backward citations and the attractiveness for gaining forward citations [279]. To be able to better differentiate between the various possible relationships between backward citations and value estimates, we consider certain subsets and properties of backward citations separately in this work. The following variables are derived from information about backward citations and attempt to capture many of these distinct subsets and properties.

Backward citations to patents (BPA) are non-inclusive of self-citations, which constitute their own variable (see below). As only cohorts of USPTO patents are considered in this work, the variable BPA counts backward citations to patents that were also granted by the USPTO. The patents included in the analyses conducted in this chapter cite at least one USPTO patent, which ensures that all variables related to backward citations have definite values.

Backward self-citations (BSC) are citations to previously granted USPTO patents owned by the same assignee. These are often interpreted as the result of cumulative research [87, 287], which in turn is an indicator of a firm's willingness to invest in R&D to strengthen their competitive position in a particular technology space [67, 288]. It can be expected that firms do not undertake this kind of research without the expectation that the technology, or the particular invention, will be valuable in some way. Evidence for this includes the fact that the number of self-citations is often significantly correlated with inventor team size – another indicator of firm investment. Inventions that build on previous work by the same firm may therefore have a larger impact than those that do not. Conversely, self-citations could also be indicative of an incremental improvement rather than a technological breakthrough, which may hinder impact [289].

Backward citations to non-patent literature (BNP) are generally to scientific articles.

The vast majority of these citations originate with the applicant, as part of their duty of candor, rather than with the examiner [44]. Inventions with a strong scientific grounding, or those that otherwise build on what could be considered more fundamental research, may be more likely to be among the first patents granted in a new technological avenue. This property may lead to an increased likelihood of obtaining forward citations [290].

Backward citations to foreign patents (BFP) cover knowledge disclosed in patents granted outside the USPTO's jurisdiction. The vast majority of these citations are added by the applicant rather than the examiner [44].

The *average age of backward citations (BAG)* is calculated as the arithmetic mean of the ages of all the patents cited by a new patent i (including self-citations) at the latter's time of application $T_i^{(a)}$, with (a) denoting application. This variable can be indicative of the proximity of the citing patent to the technological frontier [291] because the fact that a patent mostly cites very recent inventions would generally indicate that the patented material is cutting-edge. Conversely, a patent that cites mostly older inventions may be either a late addition to an increasingly obsolete technological field or part of a generally slow-moving technological field. In any case, it is expected that technologically influential patents are more likely to cite newer patents [292]. As in Section 4.3.1, the age of a citation is defined as the time between the grant of the cited patent and the application of the citing patent: $T_i^{(a)} - T_j^{(g)}$, where (g) denotes grant. To avoid mathematical difficulties, we exclude the relatively small number of patents for which BAG itself turns out to be negative (1-5% depending on the technology category).

Backward-citations' pedigree (BCP) is a metric defined for the first time in this work that is calculated based on the normalized geometric mean of the number of forward citations accrued by the cited patents listed on a particular patent document at the latter's time of application. More specifically, the backward-citations' pedigree of patent i that has been applied for at time $T_i^{(a)}$ is defined as

$$\text{BCP}_i = \left[\prod_{j=1}^{N_i} \frac{k_j(T_i^{(a)})}{\langle k(T_i^{(a)}) \rangle_{S_j}} \right]^{\frac{1}{N_i}}, \quad (6.4)$$

where N_i is the number of patents cited by patent i , and $k_j(T_i^{(a)})$ denotes the number of citations the cited patents j has at $T_i^{(a)}$. The citation counts of cited patents j are normalized by the average number $\langle k(T_i^{(a)}) \rangle_{S_j}$ of citations of patents from the same

CPC Section S_j of patent j granted within the same three-month period. The geometric, rather than the arithmetic, mean has been adopted because of the high skewness of citation distributions.

Originality (ORI) is a metric that quantifies the diversity of technologies that a patent cites. The originality of patent i is defined as [88]

$$\text{ORI}_i = 1 - \sum_C \left(P_C^{(i)} \right)^2, \quad (6.5)$$

where $P_C^{(i)}$ is the proportion of the N_i backward citations from patent i to other patents going into a given Class C from the CPC system.⁵ Here the information about all Classes a cited patent j is categorised into is utilised, rather than just the first listed as was done in Ref. [88]. Specifically, each cited patent j contributes to $P_C^{(i)}$ for each one of the $N_j^{(\text{cl})}$ classes it is listed under with a value $(1/N_j^{(\text{cl})}) \cdot (1/N_i)$. This construction ensures the relation $\sum_C P_C^{(i)} = 1$ and, thus, the reasonability of the definition (6.5) – ORI is limited to the half-open interval $[0, 1)$. In particular, a patent with $\text{ORI}_i = 0$ will have all its backward citations to patents from one particular class. In contrast, ORI_i will be large when the citations of a given patent are distributed over many CPC Classes. In contrast to Ref. [88], we use the CPC Classes rather than the technology classes defined in the NBER patent database [185].

Two idiosyncrasies inherent in the originality variable are, firstly, that it is sensitive to the classification system being used (a coarse system would skew the originality distribution towards zero, and vice versa for a highly specific system), and, secondly, that it is correlated with the number of backward citations (greater number of citations creates more opportunity for technological diversity in those citations). The CPC system makes adjusting for the first problem quite straightforward; it is possible to simply tune the originality distribution by moving up or down the classification hierarchy until a distribution with reasonable spread is obtained. The second issue is addressed by the factor analysis (see Section 6.3.1). The remaining variables are unrelated to backward citations.

⁵ *Classes* are the second-highest hierarchy level in the CPC, just below *Sections*.

Variables Unrelated to Backward Citations

Number of Claims (CDE & CIN) is a common indicator of both technological significance and economic value [279], for a number of reasons. First, this metric appears to be a robust correlate of many measurable dimensions of patent value across all technology categories (e.g. economic value in surveys [276], probability of litigation [79], and forward-citation counts [43]). Second, the number of claims varies quite widely, so bias or error due to discretisation of count data is smaller than for most other available indicators. Third, the relationships between the number of claims and patent value is usually seen to be quite transparent — the more claims on a patent, the more legally and technically broad the patent is, and the more economic and technological impact the patent could potentially have.

However, it is also possible to restrict patent rights with increasing numbers of claims, as additional claims may reflect the narrowing of the attributes unique to the invention in question. For this reason, this metric is split into two variables — *independent claims (CIN)* and *dependent claims (CDE)* [293]. Independent claims are written in broad terms and serve to delineate the distinct components of an invention and stand on their own. Dependent claims, on the other hand, generally specify the precise embodiment of the independent claims such that the scope of each associated independent claim is well-defined. The separation of claims into these two categories is an attempt to roughly distinguish the claims that narrow technological breadth from those that broaden it, and as the total number of claims is simply the sum of the independent and dependent claims, we can only gain information via this separation. The effect of this separation turns out to be minimal, with the only consistent difference between the two variables being that CIN tends to be much more positively correlated with variables that indicate technological complexity such as the number of figures or the grant lag (both defined below). Patents must contain at least one independent claim; hence the need to subtract 1 from the total number of independent claims to get CIN.

Class Membership (NCL), which has also been called *patent scope* [280,294], is simply a count of how many technology categories a patent has been classified under. It constitutes a measure for technological breadth and has been linked to patent value [294]. For this work, we are using the CPC Classes [295] as the categories, as they represent a reasonable level of specification. If the categorization levels used for this variable are too specific (e.g., using CPC Sub-Classes), we may expect the number of classifications

a patent has to become technology-dependent, and so the explanatory power would weaken when attempting to detect within-category variation in quality. On the other hand, if the categories were too broad (e.g., when using the CPC Sections as categories), very few patents will belong to more than one category and therefore the information content of this variable would be very limited. As it is a requirement to specify at least one class for any patent, this variable is defined as $NCL_i = N_i^{(cl)} - 1$, where $N_i^{(cl)}$ is the total number of classes a patent i is listed under.

Grant Lag (LAG) is the length of time between the initial application date and the eventual grant date of a patent. While a very simple metric, there are many ways in which this time lag may be relevant to technological influence. Recent work [265,280,296] suggests that the grant lag is inversely proportional to patent quality, primarily due to two factors. Firstly, the claims on a patent that is marginal with respect to patentability and therefore of lower quality may take longer to negotiate and examine. Secondly, in some jurisdictions, companies have the ability to pay for accelerated examination of their most valuable patents. This latter phenomenon does not affect the patents in our cohort as the accelerated-examination programme at the USPTO was instituted only after these were granted, however it may affect the aforementioned recent work [265,280,296]. The former point may be relevant for us, but there is also evidence that the number of claims and number of backward citations are both positively correlated with forward citations [259], and one would expect that a longer negotiation process would result in increases in both of these variables.⁶ Additionally, a longer examination process requires a greater financial investment on the part of the inventor/assignee because the attorney spends a longer time negotiating with the examiner during this process. Thus, one would expect that an extended examination period would only be tolerated if there was some prospect of return on this investment. A long grant lag may, therefore, reflect a firm's confidence in the economic value of an invention, even if the invention is only marginally patentable. Finally, the complexity of an invention may also affect the length of the examination period, as the more components an invention has, or knowledge required to understand its function, the longer it may take to determine its patentability.

The number of inventors (INV) listed on a patent is often used as an indicator of

⁶Note that even if grant lag correlated with both claims and backward citations (and these both correlated with forward citations), this does not automatically imply a correlation between grant lag and forward citations.

Variable	Abbreviation
backward citations to patents	BPA
independent claims – 1	CIN
backward self-citations	BSC
dependent claims	CDE
backward citations to foreign patents	BFP
inventor team size – 1	INV
backward citations to non-patent lit.	BNP
class membership – 1	NCL
average age of backward citations	BAG
originality	ORI
backward-citations' pedigree	BCP
grant lag	LAG
number of figures	FIG

Table 6.1: List of the variables used this chapter as raw measures of intrinsic patent quality and their abbreviations. The notation $v^{(a)}$ with $a = 1, 2, \dots, N_v$ is used to refer to these variables generally in mathematical formulae in Sections 6.3 onward.

both the technological and economic value of a patent [297]. This is because a large team of inventors may not only be indicative of the large breadth of experience and technical expertise that was required to successfully produce the invention, but also reflect the amount of human resources (and therefore financial resources) the company dedicated to the project. As at least one inventor must be listed on any patent, we subtract 1 from the total number of inventors to get the value of INV.

The number of figures (FIG) in a patent could be expected to measure an aspect of the technological complexity of an invention. This metric does not appear to have been used in previous research on patent quality.

6.3 Constructing a Fitness Metric

To integrate the quality-indicative variables described in Section 6.2 into a growing network model as described in Eq. (6.1), it is necessary to condense these variables into a single fitness value for each patent i , η_i . The first step towards this goal is to minimise

cross-correlations between these variables by reducing them to abstract dimensions of patent quality such that variables that are indicative of the same kind of quality are combined. This is primarily to reduce the over-representation of particular qualities. For example, CIN and CDE represent similar information and so rather than including both of these variables in the fitness parameter directly, it would be preferable to appropriately combine them into a single variable representing some notion of legal strength, which more accurately represents some distinct aspect of patent fitness. More complex, however, are variables that may indicate multiple dimensions of patent quality. A large inventor team, for example, may be indicative of a highly complex invention that requires the input of knowledge from many disparate domains, while also indicating a large investment on the part of the assignee. Technological complexity and firm valuation are not necessarily particularly related to one another, and yet there exists a variable which may capture both of these dimensions, each of which may independently contribute to patent fitness.

Factor analysis [298] facilitates this kind of dimensionality reduction. This methodology will allow the reasonable combination of the variables listed in Section 6.2 (see Table 6.1) in order to represent their variance using a small number of key quality dimensions, referred to henceforth as *factors*.

6.3.1 Dimensionality Reduction

After impact-relevant variables have been identified, an exploratory factor analysis [298] is conducted. The methodology described here involves the extraction of factors based on a principal factor analysis (also known as principal axis factoring) [299]. Unlike a principal component analysis (PCA) [300], extracted latent factors are not assumed to explain all variance quantified by the covariance matrix,⁷ only the so-called *common variance* (defined in Section 6.3.1). In this work, this assumption implies that the portion of variance that is unique to any individual variable has minimal influence on forward citations. The choice of the variables listed in Section 6.2 ensure there are no patent quality dimensions that are obviously captured primarily by a single variable that is relatively uncorrelated to all others.

⁷It was found that PCA tends to overload particular variables in the extracted components due to this assumption.

For these reasons, factor analysis is both a more intuitive framework than PCA for the purposes of capturing the abstract, hidden influences on technological impact, and a much more flexible framework that can be used as a bedrock on which to build more complex models of patent value and quality measures, such as confirmatory factor analysis [301] or structural-equation-based approaches [302]. These methodologies may provide more insight into particular aspects of patent quality and how they are related; however, they are not well-suited to obtaining a single fitness value for each patent due to the inherent inter-relatedness of the resultant quality dimensions.

Sampling Adequacy

In typical factor analysis studies, the commonly reported Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy [303] is given as an indicator of *factorial simplicity* of a particular variable or set of variables. For any individual variable, this measure is maximized when it has a non-zero loading on only one factor. The KMO is then averaged over all variables as a way of measuring this simplicity. Hence, it is a way to quantify how useful particular variables are in determining the meaning of extracted factors, and therefore how useful a factor analysis may be in unravelling the correlational structure of a set of variables – if the goal is to interpret the resultant factors.

The KMO values for the set of variables listed in Section 6.2 is between 0.62 and 0.68 for each of the patent cohorts. KMO values of this magnitude indicate that only large or consistent differences between factors and technology categories should be taken seriously for any statistically relevant interpretation [303]. The KMO values are quoted here only for general information, as finding and interpreting the latent factors influencing the chosen variables is not the primary goal of this work. Instead, factor analysis is used only as a tool to reduce cross-correlations between variables that determine the fitness variable for citation-network growth. Thus, for our purposes, the KMO values are actually not relevant.

Transformation of Variables

Before factor analysis, most variables are log-transformed and normalised. Applying the log-transformation is motivated by the observation that the variables chosen in Section 6.2 are generally close to being log-normally distributed like many performance-indicator-type quantities from various fields [138, 187, 304, 305]. This reflects the de-

creasing marginal relevance of changes in these variables with respect to technological impact; e.g., an increase from 1 independent claim to 2 independent claims will generally represent a larger change in impact than an increase from 10 to 11 independent claims. The log-transformation also reduces the effect of extreme values in these variables.

The variables CIN, INV, and NCL are defined by subtracting 1 from the patent characteristics they represent, as they all have a minimum value of 1. This means all variables listed in Table 6.1 have a minimum value of 0. After the log-transformation, all variables are then shifted such that their mean is zero, then scaled such that their standard deviation is unity. This removes from the analysis any bias introduced by differences in scale between these variables. The variable ORI is already approximately normally distributed and therefore does not require a log-transformation, however it is still shifted and scaled as usual. The variables on which the factor analysis is conducted are therefore defined

$$\bar{v}^{(a)} = \begin{cases} \left[\ln(1 + v^{(a)}) - \mu_{\ln(1+v^{(a)})} \right] / \sigma_{\ln(1+v^{(a)})} & \text{for all variables except ORI,} \\ \left[v^{(a)} - \mu_{v^{(a)}} \right] / \sigma_{v^{(a)}} & \text{for ORI,} \end{cases} \quad (6.6)$$

where μ_x and σ_x represent the mean and standard deviation of x respectively.

Factor Extraction

Factor analysis is underpinned by the assumption that the $N_v = 13$ variables $v^{(a)}$ chosen and transformed as described above are comprised of two distinct components: one common component that originates from N_u latent variables, or factors, $u^{(b)}$ that can influence multiple observable variables; and a unique component which is due to external factors affecting each component independently. That is, we can express our observed variables as $\bar{\mathbf{v}}$, an $N_v \times 1$ column vector, as the linear combination

$$\bar{\mathbf{v}} = \Lambda \mathbf{u} + \mathbf{q} \quad , \quad (6.7)$$

where Λ is an $N_v \times N_u$ matrix of weights that specifies the composition of the common component of the observed variables in terms of N_u factors contained in vector $\mathbf{u}$, while the length- N_v vector $\mathbf{q}$ specifies the unique component of $\bar{\mathbf{v}}$. Eq. (6.7) is sometimes referred to as the ‘Fundamental Equation of Factor Analysis’ [298].

We can now express the covariance structure of the observed variables in terms of the proposed composition in Equation 6.7 as

$$\begin{aligned} E(\bar{\mathbf{v}}\bar{\mathbf{v}}^T) &= E[(\Lambda\mathbf{u} + \mathbf{q})(\Lambda\mathbf{u} + \mathbf{q})^T] \\ &= \Lambda E(\mathbf{u}\mathbf{u}^T)\Lambda^T + \Lambda E(\mathbf{u}\mathbf{q}^T) + E(\mathbf{q}\mathbf{u}^T)\Lambda^T + E(\mathbf{q}\mathbf{q}^T) \end{aligned} \quad (6.8)$$

At this point, correlational assumptions need to be specified. The first is that the common factors are mutually uncorrelated and have unit variance such that $E(\mathbf{u}\mathbf{u}^T) = I_{N_u}$, where I_{N_u} is the $N_u \times N_u$ identity matrix. The second is that the common factors are uncorrelated with the unique factors, $E(\mathbf{u}\mathbf{q}^T) = E([\mathbf{q}\mathbf{u}^T]^T) = \mathbf{0}$. Lastly, the unique factors are mutually uncorrelated such that we can express $E(\mathbf{q}\mathbf{q}^T) = \Psi$ as a diagonal matrix with non-zero entries ψ_{aa}^2 representing the unique variance associated with observed variable $\bar{v}^{(a)}$. We can now simplify Eq. (6.8) to obtain

$$E(\bar{\mathbf{v}}\bar{\mathbf{v}}^T) = \Sigma = \Lambda\Lambda^T + \Psi \quad , \quad (6.9)$$

where Σ is the covariance matrix of the set of variables $\bar{\mathbf{v}}$. Eq. (6.9) is sometimes referred to as the ‘Fundamental Theorem of Factor Analysis’ [298].

From this point, the factor extraction process is an eigendecomposition of the covariance matrix representing the common components, or communalities, of the observed variables, $\Lambda\Lambda^T$. However, this calculation first requires an estimate of Ψ , denoted $\hat{\Psi}$. It can be shown that the multiple squared correlation coefficient for variable $\bar{v}^{(a)}$ with all other $N_v - 1$ observed variables gives a lower bound on the true value [306, 307], and so this is used as a conservative estimate for ψ_{aa}^2 . This value is equal to

$$\hat{\psi}_{aa}^2 = \mathbf{c}_a^T R^{-1} \mathbf{c}_a \quad (6.10)$$

where $\mathbf{c}_i$ is the column vector of correlations between $\bar{v}^{(a)}$ and each of the other $N_v - 1$ observed variables, and R is the correlation matrix for the $N_v - 1$ variables excluding $\bar{v}^{(a)}$. The eigendecomposition is then conducted on the modified covariance matrix $\Sigma - \hat{\Psi}$, whereupon N_v eigenvalue-eigenvector pairs are extracted.

Selection of Factors

Factor analysis involves the eigendecomposition of a modified covariance matrix of the input variables, and therefore produces the same number of factors as variables (assuming there are no repeated eigenvalues – an exceedingly unlikely event when dealing with real-world data). It is necessary, therefore, to discard factors corresponding to dimensions that explain very little of the variance in the data, according to some criterion. It is generally advised that the selection of the principal factors should be done by Horn’s parallel analysis [308]. With a finite number of observations, the largest extracted eigenvalue from a factor analysis on a random matrix will always be greater than zero, and the smallest will always be less than zero – if one were to select an eigenvalue of zero as the threshold for choosing which factors to keep, there will always be some factors that would be deemed to contain sufficient information to retain, even if the data is completely random. Horn’s parallel analysis [308] method accounts for this by running a Monte-Carlo simulation to find the expected eigenvalues for a randomly generated dataset with the same number of variables and observations, after which a factor is retained if its eigenvalue is larger than that of the same rank found for the random data. However, for a large amount of data such as the patent cohorts analysed here, Horn’s parallel analysis produces eigenvalues which are all very close to zero, such that it is not distinguishable from a threshold of zero, commonly known as the Kaiser criterion [309]. Hence, in practice, the Kaiser criterion is used to reduce computation time but the results are identical to those obtained using Horn’s parallel analysis.

After discarding eigenvectors with negative eigenvalues, the N_u retained eigenvectors are arranged into columns by descending order of their associated eigenvalues to yield the $N_v \times N_u$ matrix Λ . These vectors are indexed in this order and define the length of the column vector of factors $\mathbf{u}$ as well their association with each of the columns in Λ .

Factor Rotation

After factors with small eigenvalues have been excluded, it is possible to *rotate* the remaining factors in this lower-dimensional factor space to make them more interpretable. The goal of the analysis conducted in this chapter is not to produce interpretable factor loadings, but applying a transformation that rotates the factors while retaining orthogonality does not reduce the information content of the factors and may better inform us about the nature of the dimensions they represent. This process will provide a glimpse

into the kind of insights that are possible should the goal of such research be obtaining (potentially interacting) patent quality dimensions that reflect reality as closely as possible.

Generally, an oblique rotation (one that does not require the factors to remain uncorrelated) is advised for factor rotation as it is likely that the real-world latent factors are, in fact, correlated, and enforcing orthogonality may introduce unwanted artefacts [298]. The reason orthogonality is enforced is because interpreting the resulting factors is not our primary goal. Instead, factor analysis is used to minimise collinearity between our quality-indicator variables.

As there are fewer factors than observed variables (when $N_u < N_v$) the linear model in Eq. (6.7) is not identifiable; that is, there are an infinite number of matrices Λ that will satisfy Eq. (6.7). We may select an orthonormal transformation matrix T such that

$$\Lambda \mathbf{u} = \Lambda(TT^T)\mathbf{u} = (\Lambda T)(T^T\mathbf{u}) \quad (6.11)$$

also satisfies Eq. (6.7). Therefore, we can construct a matrix T to optimise the interpretability of the factor loading matrix using some constraint that captures whatever notion of interpretability we require. The most common constraint in the orthonormal case is called the Varimax Criterion, which aims to maximise the sum of the variances of the squares of entries in each column of ΛT , and in practice acts to reduce the number of high factor loadings while increasing the number of near-zero loadings – thereby simplifying the factor loadings and facilitating their interpretation. Therefore, we must find the transformation T that maximises the objective function

$$\phi = \sum_b \frac{N_v \sum_a (\lambda_{ab}^2/h_a^2)^2 - (\sum_a (\lambda_{ab}^2/h_a^2))^2}{N_v^2}, \quad (6.12)$$

where a and b correspond to the rows and columns of Λ respectively, λ_{ab} corresponds to the entry of Λ in row a and column b , and $h_a^2 = 1 - \psi_{aa}^2$ is the communality (the non-unique component) of $\bar{v}^{(a)}$. Weighting this optimisation procedure with communalities removes bias towards variables with large communalities (or low uniqueness) [310], which is important as communalities largely depend on the choice of variables on which we conduct the factor analysis. In practice, this weighting is done before rotation and reweighting done after rotation, such that there is no permanent changes in vector lengths. This transformation yields the new factor loadings $\tilde{\Lambda} = \Lambda T$ and factors $\tilde{\mathbf{u}} = T^T\mathbf{u}$.

Factor Scoring

In order to construct a fitness metric for each patent, we need to calculate a factor score for each factor and every patent. For example, if $u^{(1)}$ has a large loading on independent claims, then we may expect a patent with many independent claims to have a relatively large $u^{(1)}$ score. While we have estimated the variances of the unique components $\mathbf{q}$, we do not know what fraction of the observed variables the unique component makes up for any individual patent. Therefore, we need to obtain the value of N_u factor scores using N_v variables, however those N_v variables are the linear combination of N_u latent factors and N_v unique factors (see Eq. (6.7)). This is the so-called indeterminacy problem of factor scoring. As there is no unique way to separate N_v variables into $N_v + N_u$ factors given only their variances, we are only able to estimate the linear combinations $\tilde{\Lambda}\tilde{\mathbf{u}}$ for each patent. The most common way to do this is to minimise the squared expected prediction error between the estimated scores and the extracted factors, given the empirical data – a regression.

First, we define the b th estimated factor $u^{(b)*}$ to be a linear combination of the observed variables with some to-be-determined $N_v \times 1$ weight vector $\mathbf{d}_b$, $u^{(b)*} = \tilde{\mathbf{v}}^T \mathbf{d}_b = \mathbf{d}_b^T \tilde{\mathbf{v}}$. Then, we minimise the objective function $E([u^{(b)*} - \tilde{u}^{(b)}]^2)$ with respect to $\mathbf{d}_b$, calculating

$$\begin{aligned} \nabla_{\mathbf{d}_b} E([u^{(b)*} - \tilde{u}^{(b)}]^2) &= \nabla_{\mathbf{d}_b} E([\tilde{\mathbf{v}}^T \mathbf{d}_b - \tilde{u}^{(b)}]^2) \\ &= 2[E(\tilde{\mathbf{v}}\tilde{\mathbf{v}}^T)\mathbf{d}_b - E(\tilde{u}^{(b)}\tilde{\mathbf{v}})] \quad . \end{aligned} \quad (6.13)$$

We recognise $E(\tilde{\mathbf{v}}\tilde{\mathbf{v}}^T) = \Sigma$ (the covariance matrix of the input variables) and $E(u^{(b)}\tilde{\mathbf{v}}) = \tilde{\Lambda}_b$ where $\tilde{\Lambda}_b$ is the b th column of $\tilde{\Lambda}$. Setting Eq. (6.13) equal to zero and rearranging, we get

$$\mathbf{d}_b = \Sigma^{-1} \tilde{\Lambda}_b \quad (6.14)$$

which leads to

$$u^{(b)*} = (\Sigma^{-1} \tilde{\Lambda}_b)^T \tilde{\mathbf{v}} = \tilde{\Lambda}_b^T \Sigma^{-1} \tilde{\mathbf{v}} \quad . \quad (6.15)$$

This can then be generalised to the $N_u \times 1$ vector of factors as

$$\mathbf{u}^* = \tilde{\Lambda}^T \Sigma^{-1} \tilde{\mathbf{v}} = L_a^{(b)} \tilde{\mathbf{v}} \quad , \quad (6.16)$$

which gives us the *regression* or *Thomson estimates* for the individual factor scores [311, 312], from which we can construct a fitness metric as described in the following Sections. The loading matrices $L_a^{(b)}$ for each of the CPC technology categories examined in this chapter are shown in graphic form in Figs. 6.1-6.4. Note, categories D (Textiles) and E (Fixed Constructions) are not analysed here due to an insufficient number of patents available to produce interpretable parameter estimates once fitness is integrated into the dynamic citation accrual model.

6.3.2 Fitness Metric Composition

Performing the factor analysis described in the Section 6.3.1 has yielded the predicted factors $u^{(b)*}$ as the minimally correlated superpositions of the normalized patent-quality indicator variables $\bar{v}^{(a)}$. These quality factors will now be used to gain a proxy measure for the unobserved fitness variable η that quantifies the intrinsic attractiveness for patents to be cited.

To this end, we determine the weights w_b of these factors in their combination that is most highly correlated with citation growth at early times by running a forward-stepwise ordinary least squares regression on a randomly selected small training dataset with the Bayes Information Criterion (BIC) [313] as our test statistic and the log of the (mean-scaled) citations at grant as the dependent variable.⁸ This choice of dependent variable is discussed in Section 6.1.3. In order to put all factors on equal footing in this regression, they are rescaled to have unit standard deviation, $\bar{u}^{(b)} = u^{(b)*}/\sigma_b$, where σ_b is the standard deviation of $u^{(b)*}$.

Patents in the training dataset are randomly selected and comprise about 10% of each technology-specific cohort (8,000-10,000 patents). The stepwise regression then yields coefficients that comprise the weights w_b measuring the variances in the training dataset associated with each of the retained factors $\bar{u}^{(b)}$. The specific way in which we

⁸Note that, for our purposes, the ability to plausibly exclude citation-mediated knowledge diffusion by choosing to use citations at grant justifies adopting this procedure even though it may bias the model towards patent qualities influencing the citing behavior of examiners and self-citations by inventors.

combine these to construct a single fitness value for each patent is detailed in Section 6.3.3. As the training dataset comprises only 10% of patents from any given cohort, and the dependent variable is determined by the citation count at only one point in time, this procedure does not preordain the functioning of our approximation for the citation-related fitness parameter across the entire cohort or for the entire ten-year time period for which forward-citation accrual is analyzed. However, using the log of citation counts at a later time as the dependent variable in the weight-determining regression for the given set of chosen factors does not yield significantly different results (it does, however, result in slightly different weights w_b). To check the information loss due to the factor analysis process (lost first when factors with eigenvalues less than zero were discarded, and second in the stepwise regression), Table 6.2 lists the fraction r^2 of variability in the dependent variable (the log of citation counts for patents from a given cohort at a given time t after their time of grant) explained by all of the normalized quality-indicator variables $\bar{v}^{(a)}$, alongside the proportion r_u^2/r^2 of the explained variability retained in the given set of selected factors.⁹ The factors are able to adequately represent the total variability in citation counts associated with the quality-indicator variables known at time of grant, with the goodness of fit r_u^2/r^2 generally decreasing only slightly for later t . In the cases where r^2 is seen to increase with t , this most likely reflects a reduction of granularity in the data by the increased spread of citation counts among patents in these cohorts.

Figs. 6.1-6.4 show both the results of the factor analysis and the factor weighting from the stepwise regression, obtained for all technology-specific cohorts of patents granted during 1999-2001. Note that, due to this stepwise nature of the regression, some factors never add enough information (as measured by the BIC) to be included in the fitness metric. We furthermore adopt the convention where $w_b \geq 0$ by absorbing the sign of the regression coefficients into the definition of the factors $\bar{u}^{(b)}$ in terms of the variables $\bar{v}^{(a)}$. This fixes the sign of the loadings $L_a^{(b)}$ shown in Figs. 6.1-6.4. While interpreting these results is not the goal of this analysis, the relative independence of each factor in terms of their high-loading variables, as well as the significant reduction in dimensionality, indicates that the factor analysis has produced sensible composite variables: claims of both types (CIN and CDE) consistently load on the same factor, as do indicators of technological scope (NCL and ORI), reliance on established technology (BPA and

⁹The value r^2 represents the maximum possible multiple correlation coefficient obtainable using the information contained in these variable, and, therefore, is an upper limit on the goodness-of-fit that is possible when only using the retained factors, r_u^2 .

t (yr)	Section	A	B	C	F	G	H	Y
0	r^2	0.17	0.14	0.13	0.08	0.27	0.27	0.22
	r_u^2/r^2	0.95	0.92	0.87	0.93	0.91	0.93	0.96
5	r^2	0.26	0.20	0.19	0.15	0.29	0.26	0.26
	r_u^2/r^2	0.94	0.87	0.87	0.86	0.84	0.86	0.94
10	r^2	0.28	0.20	0.19	0.16	0.29	0.25	0.26
	r_u^2/r^2	0.94	0.87	0.87	0.84	0.82	0.85	0.93

Table 6.2: Fraction r^2 of variability explained by the variables $\bar{v}^{(a)}$ in the log of citation counts accrued by time t after their time of grant by USPTO patents from indicated CPC Sections granted during 1999-2001, and the proportion r_u^2/r^2 of this explained variability that is able to be captured by the set of factors $\bar{u}^{(b)}$ chosen via the stepwise regression.

BAG), and firm investment (BFP, BSC, and INV).

Figure 6.2: Factor loadings $L_a^{(b)}$ obtained from the exploratory factor analyses (left) and fitness composition weights w_b obtained from the stepwise regression (right) performed on patent-quality indicator variables for cohorts of patents granted during 1999-2001 within the specified CPC Sections C & F. The order of factors from left to right in each panel corresponds to the order of decreasing eigenvalues associated with each of the factors. Note that fewer factors appear in the right panels than the left panels, as some factors are dropped in the stepwise regression process.

Figure 6.3: Factor loadings $L_a^{(b)}$ obtained from the exploratory factor analyses (left) and fitness composition weights w_b obtained from the stepwise regression (right) performed on patent-quality indicator variables for cohorts of patents granted during 1999-2001 within the specified CPC Sections G & H. The order of factors from left to right in each panel corresponds to the order of decreasing eigenvalues associated with each of the factors. Note that fewer factors appear in the right panels than the left panels, as some factors are dropped in the stepwise regression process.

Figure 6.4: Factor loadings $L_a^{(b)}$ obtained from the exploratory factor analyses (left) and fitness composition weights w_b obtained from the stepwise regression (right) performed on patent-quality indicator variables for the cohort of patents granted during 1999-2001 within CPC Section Y. The order of factors from left to right in each panel corresponds to the order of decreasing eigenvalues associated with each of the factors. Note that fewer factors appear in the right panels than the left panels, as some factors are dropped in the stepwise regression process.

6.3.3 Calibration of Fitness Distribution

Once the composition of the fitness metric η is determined through the process detailed in Section 6.3.2, the shape and scale of the final fitness distribution $\rho(\eta)$ must be determined – these properties are not trivially determined, as there exist model restrictions and real-world considerations that need to be built into the final fitness values. For example, fitnesses are strictly positive, as they enter the rate function (Eq. (6.1)) multiplicatively in our model. Therefore, a normal distribution centred on zero is not feasible – patents cannot lose citations. Additionally, there exists a question about the width of the fitness distribution: how do we determine the size of the fitness differences between poor, average, and high quality patents? What follows is a novel methodology that aims to answer this question.

By construction, the linear superposition of relevant factors $\bar{u}^{(b)}$ with weights w_b as coefficients will be a normally distributed quantity having zero mean and variance given by $\sum_{b=1}^{N_u} w_b^2$. However, motivated by empirical observations that have generally found citations and rates of citation to be log-normally distributed [193], the distribution $\rho(\eta)$

of fitnesses will be assumed to be log-normal also. This choice allows for a skewed fitness distribution with strictly positive support (the distribution is defined for $\eta > 0$), and is a simple transformation from the normally distributed factors (or any linear combination of factors). Additionally, for conceptual simplicity, it is possible to fix the mean value of the fitness, $\mu_\eta \equiv 1$, which implies $\mu_{\ln(\eta)} = -\sigma_{\ln(\eta)}^2/2$. Thus the only free parameter characterizing the log-normal distribution of fitnesses is the standard deviation $\sigma_{\ln(\eta)}$. This simplification has the additional benefit of having the limit $\rho(\eta) \rightarrow \delta(\eta - 1)$ as $\sigma_{\ln(\eta)} \rightarrow 0$, where $\delta(\cdot)$ is the Dirac- δ function, which corresponds to network growth with no fitness as explored in Chapter 4.

To determine $\sigma_{\ln(\eta)}$, and thus characterise the shape of the entire fitness distribution, it is possible to examine the number of uncited patents. Intuitively, if all patents with zero accrued citations (at some time t) had the same chance of being cited, then for small times the proportion of patents with zero citations, $n(0, t)$, would decay exponentially (see Eqs. (3.20) & (3.21) in Chapter 3). If fitness were introduced, those patents with large fitnesses would likely gain their first citation very quickly, while most patents would have small fitnesses and exhibit a much slower rate of first citations. The exact shape of $n(0, t)$ can, therefore, inform the width of the fitness distribution by measurement of the deviation from the exponential decay at small times associated with no fitness.

The fraction $n_\eta(0, t)$ of uncited patents with fixed fitness η in some cohort satisfies the Master equation [157]

$$\frac{dn_\eta(0, t)}{dt} = -\lambda(0, t) n_\eta(0, t) \quad . \quad (6.17)$$

Assuming the citation rate to be of the form $\lambda(k, t) = \eta \tilde{A}(t) \tilde{f}(k)$, we find [157]

$$n_\eta(0, t) = \exp \left\{ -\eta \tilde{f}(0) \int_0^t dt' \tilde{A}(t') \right\} \quad . \quad (6.18)$$

For a cohort where patents with fitness η are distributed according to the distribution $\rho(\eta)$,

$$\begin{aligned} n(0, t) &\equiv \int d\eta \rho(\eta) n_\eta(0, t) \\ &= \int d\eta \rho(\eta) \exp \left\{ -\eta \tilde{f}(0) \int_0^t dt' \tilde{A}(t') \right\} \quad . \end{aligned} \quad (6.19)$$

Expanding the exponential on the right hand side of this expression yields

$$n(0, t) = \sum_{m=0}^{\infty} \frac{E[\eta^m]}{m!} \left\{ -\tilde{f}(0) \int_0^t dt' \tilde{A}(t') \right\}^m \quad (6.20)$$

in terms of the moments $E[\eta^m]$ of the fitness distribution $\rho(\eta)$. For our particular case of interest, we have $\rho(\eta) = \text{Lognormal}(-\sigma_{\ln(\eta)}^2/2, \sigma_{\ln(\eta)}^2)$ and thus obtain

$$n(0, t) = \sum_{m=0}^{\infty} \frac{\exp\left\{\frac{m(m-1)}{2} \sigma_{\ln(\eta)}^2\right\}}{m!} \left[-\tilde{f}(0) \int_0^t dt' \tilde{A}(t') \right]^m. \quad (6.21)$$

Focusing on the time evolution of uncited patents has the advantage that it is insensitive to the actual k dependence of $\tilde{f}(k)$. For any sufficiently well-behaved ageing function $\tilde{A}(t)$, Eq. (6.21) can be calculated and compared with empirical data. However, for our purposes, we do not want to prejudice the estimate for the fitness variable by any *ad-hoc* assumption about ageing. We therefore consider the short-time limit where $\tilde{A}(t) \approx \tilde{A}(0)$ and, hence, Eq. (6.21) provides the Taylor expansion for $n(0, t)$ in the variable t with only two independent parameters: $\tilde{A}(0)\tilde{f}(0)$ and $\sigma_{\ln(\eta)}^2$. After taking the logarithm, this expansion yields, up to the second order,

$$\ln n(0, t) \approx -\tilde{A}(0)\tilde{f}(0)t + \frac{1}{2} \left(\exp\left\{\sigma_{\ln(\eta)}^2\right\} - 1 \right) \left[\tilde{A}(0)\tilde{f}(0)t \right]^2, \quad (6.22)$$

which could ideally be fitted to the citation data and yield empirical values for both $\sigma_{\ln(\eta)}$ and the product $\tilde{A}(0)\tilde{f}(0)$ for each of the patent cohorts considered here. The former is exactly the value required to calibrate the width of the fitness distribution, while the latter can be used to cross-check the adequacy of the functional forms for $\tilde{A}(t)$ and $\tilde{f}(k)$ obtained in Section 6.4 from the empirically observed citation rate.

In practice, the fact that patents generally already acquire examiner citations and self-citations before the time of grant means that $n(0, 0) < 1$ and, hence, a more subtle approach is needed to conduct these measurements. In fact, the $t = 0$ point as used in previous chapters (and below in Section 6.4) is almost arbitrary, apart from being the time at which the patent is made public. It will not suffice as the point at which patents start gaining citations for the above calculation. Fig. 6.5 illustrates the typically observed time evolution of $n(0, t)$. The rate at which patents accrue their first citation, and the rate at which $n(0, t)$ concomitantly decreases, is initially slow. A marked increase in the rate is observed for $t_0 \lesssim t < t_*$, after which the decrease of $n(0, t)$ slows again. Finding the slope T^{-1} at the inflection point t_* that separates the regions of accelerated

Figure 6.5: Log-linear plot of the fraction $n(0, t)$ of uncited USPTO patents assigned to CPC Section A with grant date during 1999-2001 as a function of time t measured from each individual patent's time of grant. The approximation $n_{\text{fit}}(0, t)$ given in Eq. (6.23) obtained from fitting the line shape of $n(0, t)$ around the inflection point at $t = t_*$ is used to extract the standard deviation $\sigma_{\ln(\eta)}$ of the log of fitness. The time t_0 is determined via the extrapolation $n_{\text{fit}}(0, t_0) \equiv 1$.

and later slowed decrease in the observed $n(0, t)$ and fitting the $n(0, t)$ curve to quadratic order in the vicinity of t_* yields the approximation

$$n_{\text{fit}}(0, t) = \exp \left\{ -\frac{t - t_0}{\bar{t}} \Theta(t - t_0) + \frac{\varepsilon}{2} \frac{(t - t_*)^2}{\bar{t}^2} \Theta(t - t_*) \right\} , \quad (6.23)$$

where ε and $\bar{t}$ are fitting parameters.

Motivated by the theoretical result for the short-time limit of $n(0, t)$ given in Eq. (6.22), it is possible to identify $\bar{t} \equiv 1/[\tilde{A}(0)\tilde{f}(0)]$ and $\varepsilon = \exp\{\sigma_{\ln(\eta)}^2\} - 1$. The parameters $\tilde{A}(0)\tilde{f}(0)$ and $\sigma_{\ln(\eta)}$ obtained from these fits are listed for each technology-specific cohort in Table 6.3. The fact that t_* turns out to be within 2–6 months from the date of grant indicates that the thus-determined $\sigma_{\ln(\eta)}$ appropriately reflects the short-time behavior considered in the idealized model leading to Eq. (6.22).

Combining all of the information obtained from the factor analysis, the stepwise regression, and the fitness width calibration, it becomes possible to sensibly estimate

fitness values for every patent (for which all variables $\bar{v}^{(a)}$ are defined) in each of the technology-specific cohorts. For patent i , therefore, the estimated intrinsic fitness is calculated via

$$\eta_i = \exp \left(-\frac{\sigma_{\ln(\eta)}^2}{2} + \frac{\sigma_{\ln(\eta)}}{\sqrt{\sum_{b=1}^{N_u} w_b^2}} \sum_{b=1}^{N_u} w_b \sum_{a=1}^{N_v} L_a^{(b)} \bar{v}_i^{(a)} \right) . \quad (6.24)$$

The first term in the parentheses is to ensure the mean fitness value $\mu_\eta \equiv 1$, while the second term, from right to left, applies the Thomson regression scores from the factor analysis to each patent, applies the weights w_b associated with each retained factor in the stepwise regression, and finally scales the fitness width to $\sigma_{\ln(\eta)}$ as calculated via Eq. (6.23). The resulting (normally distributed) values are then exponentiated to obtain log-normally distributed fitnesses η_i . The next sections will incorporate these fitnesses into the rate-function measurements in order to observe their effect on network growth.

6.4 Integration of Fitness into a Growing Network Model

Now that fitness values have been obtained for all patents in each of the cohorts, these need to be integrated into the network growth measurements. As stated in Eq. (6.1), fitness enters the citation rate multiplicatively. To determine $\tilde{A}(t)$ and $\tilde{f}(k)$, therefore, the change $\Delta k_i(t + \Delta t)$ of inflation-adjusted citations received by patent i over the time interval $(t, t + \Delta t]$ is divided by the fitness η_i estimated for that patent. As for Chapters 4 & 5, the data for $[\Delta k_i(t + \Delta t)/\eta_i]/\Delta t$ (corresponding to the fitness-controlled citation rate λ_i/η_i) are still best fitted when the ageing function $\tilde{A}(t)$ and the preferential-attachment kernel $\tilde{f}(k)$ take the respective forms

$$\tilde{A}(t) = \tilde{A}_0 \exp \left(-\frac{t}{\tilde{\tau}} \right) , \quad (6.25a)$$

$$\tilde{f}(k) = k^{\tilde{\alpha}} + \tilde{f}_0 , \quad (6.25b)$$

see Fig. 6.6. As usual, the dependency on accrued citations k (time t) is fitted to the averages $[\Delta k_i(t + \Delta t)/\eta_i]/\Delta t$ over patents in the fixed- t (fixed- k) bins using a weighted non-linear logarithmic regression with $\Delta t = 3$ months. By dividing the citation rate by the fitness values for each patent, the functional form fit to the data is exactly the same as in Chapter 4. However, fitness could enter the model explicitly and we would

Figure 6.6: Residual preferential attachment and obsolescence-induced ageing in the fitness-controlled citation rate of patents from CPC Section A granted during 1999-2001. Left (Right) panel shows fits of preferential-attachment kernels for $t = 5$ years (exponential-ageing functions for $k = 7$) to both the fitness-controlled citation rate $\lambda_i/\eta_i \equiv [\Delta k_i(t+\Delta t)/\eta_i]/\Delta t$ (data points indicated by triangles) and the uncontrolled-for-fitness average citation rate $\bar{\lambda}_i \equiv \Delta k_i(t+\Delta t)/\Delta t$ (data points indicated by circles). The inclusion of fitness into the model flattens both preferential attachment and obsolescence curves, which for the former indicates a weaker preferential attachment effect, while for the latter indicates a smaller decay rate. This is achieved without sacrificing goodness of fit, and in the case of obsolescence, improves it.

obtain the same results – η_i are determined before any modelling of the rate function is conducted.

After this process, these results can be compared to those obtained with the same set of patents but without the rate correction. That is, we can compare the fitting parameters obtained above with those obtained when we fit the functional forms shown in Eqs. (6.25a) & (6.25b) to the rate averages $[\overline{\Delta k_i(t+\Delta t)}]/\Delta t$, where these are not first divided by patent fitness. This is almost identical to the analysis in Chapter 4, except the patents with undefined fitness and those in the training set have been removed such that it is possible to directly compare these results with the fitness controlled results. Both sets of analyses include the same citation-inflation procedure as for Chapter 4 (this is done before the removal of any patents from the cohorts and thus the citation values illustrated in Fig. 4.3 in Chapter 4 are still accurate).

Figure 6.7: Parameters resulting from the measurement of ageing and preferential-attachment-type growth in the fitness-controlled citation rate of patents from CPC Section A granted during 1999-2001. **Left:** Obsolescence time τ extracted for each bin of patents with fixed number of citations k from fitting the fitness-controlled citation rate to the form of Eq. (6.25a) for times $t \geq t_0 = 2$ years. **Centre:** Preferential-attachment exponent α , extracted for fixed times t from fits of the fitness-controlled citation rate to the form of Eq. (6.25b). **Right:** Constant contribution f_0 to the preferential-attachment part of the fitness-controlled citation rate, extracted for fixed times t from fits to Eq. (6.25b). Circles are fit-parameter values, solid lines their weighted averages, and black dashed (red dotted) curves show 95% confidence intervals for fit-parameter values (weighted averages).

6.4.1 Consequences for the Preferential Attachment Strength

As shown in Table 6.3, the exponent $\tilde{\alpha}_0$ extracted from citations immediately after time of grant is found to be much smaller than when fitness is not controlled for, suggesting that a portion of the observed preferential attachment is due to heterogeneous fitness, as shown in Figures 6.7 & 6.8. Even in the long-time limit, the fitness estimated via quality indicators available at the time of grant still accounts for a sizeable reduction by about 20-30% in the exponent α governing preferential attachment in the uncontrolled-for-fitness average citation rate. The preferential attachment exponent at long times, $t \geq 6.5$ years,¹⁰ will henceforth be referred to as $\tilde{\alpha}_\infty$.

Additionally, the average fitness for patents having k citations at time t after grant is given by $\bar{\eta}(k, t) \equiv \bar{\lambda}(k, t) / [\tilde{A}(t) \tilde{f}(k)]$ and constitutes a direct measure for the fitness-explained fraction of the average citation rate $\bar{\lambda}$. The results of our empirical analysis imply that

$$\bar{\eta}(k, t) \rightarrow \frac{k^\alpha + f_0}{k^{\tilde{\alpha}_\infty} + \tilde{f}_0} \frac{A_0}{\tilde{A}_0} \exp \left\{ \frac{\tau - \tilde{\tau}}{\tau \tilde{\tau}} t \right\} \quad (6.26)$$

¹⁰6.5 years was chosen as this constitutes the last third of the measurements, a period where all technology categories displayed constant values of $\tilde{\alpha}$.

in the long-time limit. Thus $\bar{\eta}(k, t) \propto k^{\alpha - \tilde{\alpha}_\infty}$ for patents having a citation count exceeding the average in the long-time limit. The fact that $\alpha > \tilde{\alpha}_\infty$ implies that the most highly cited patents will be associated with, and therefore detectable by, high values for the ex-ante-determined fitness variable constructed in this chapter.

Figure 6.8: Comparison of preferential-attachment exponents $\tilde{\alpha}$ (top) and obsolescence times $\tilde{\tau}$ (bottom) found for the fitness-controlled citation rate λ_i/η_i (cf. Eq. (6.1) in conjunction with Eqs. 6.25a and 6.25b) with α and τ extracted from fitting the uncontrolled-for-fitness average citation rate $\bar{\lambda}_i \equiv (k_i^\alpha + f_0) A_0 \exp\{-t/\tau\}$ as described in Section 6.4. As the exponent $\tilde{\alpha}$ obtained from fits of Eq. (6.25b) to patent cohorts with fixed age t saturates only at later times (see Fig. 6.7), both the saturation value $\tilde{\alpha}_\infty$ (the weighted average of the latest third among the extracted values for $\tilde{\alpha}$) and the value $\tilde{\alpha}_0$ for $t = 0$ (corresponding to the time of grant) are shown. The time $\tilde{\tau}$ is the weighted average of obsolescence times extracted for the fixed- k bins as shown in Figures 6.7, 6.10, & 6.11. The full results for all parameters estimated over the course of this chapter can be found in Table 6.3.

Figure 6.9: Empirically extracted ageing functions typically show the behavior illustrated in Fig. 6.6. Deviations of the kind shown here occur sporadically, and mostly for the large- k bins where statistical accuracy is limited due to the relatively small numbers of patents in them. Shown here are fits of exponential-ageing functions to both the fitness-controlled citation rate $\lambda_i/\eta_i \equiv [\Delta k_i(t + \Delta t)/\eta_i]/\Delta t$ (data points indicated by triangles) and the uncontrolled-for-fitness average citation rate $\bar{\lambda}_i \equiv \Delta k_i(t + \Delta t)/\Delta t$ (data points indicated by circles) at fixed k . Data shown here pertain to the $k = 27$ bin of USPTO patents from CPC Section A granted during 1999-2001.

6.4.2 Consequences for Obsolescence

The exponential time dependence (Eq. (6.25a)) generally describes ageing exhibited by the fitness-controlled citation rate very well at all times (see Fig. 6.6). A worst-case scenario can be found in Fig. 6.9 – situations like this seem to be isolated to large bins at small times, and so may be related to a lack of data in these bins. In contrast, the dynamics of the uncontrolled-for-fitness average citation rate at short times systematically shows an excess over the extrapolated exponential behavior observed in the long-time limit, as seen in Section 4.5.

In addition to explaining a significant portion of the observed preferential attachment, controlling for citation inflation as well as fitness has enabled us to reveal more clearly the purely obsolescence-induced ageing process governing patent-citation dynamics. The variation of the obsolescence times $\tilde{\tau}$ extracted for patent cohorts from different

CPC Sections can be found in Table 6.3. All parameter measurements are displayed for all CPC Sections in Figs. 6.10 & 6.11, except for Section A, which is shown in Fig. 6.7. Generally, $\tilde{\tau}$ turns out to be larger by between 0.9 and 1.7 years than the time scale τ extracted from the exponential ageing displayed by the uncontrolled-for-fitness citation rate in the long-time limit, as shown in Fig. 6.8. While uncertainties are larger than those observed for the uncontrolled-for-fitness case, and more categories displaying small deviations from separability (manifesting as slow, systematic changes in $\tilde{\tau}$ as k increases), the approach taken in this chapter clearly shows much promise as a general framework for measuring and incorporating intrinsic fitness into a model of citation network growth.

CPC Section	A	B	C	F
$\tilde{\alpha}_0$	0.65 ± 0.09	0.76 ± 0.10	0.73 ± 0.08	0.81 ± 0.20
$\tilde{\alpha}_\infty$	0.82 ± 0.02	0.85 ± 0.04	0.83 ± 0.03	0.92 ± 0.07
α	1.13 ± 0.03	1.10 ± 0.06	1.11 ± 0.04	1.14 ± 0.07
$\tilde{f}_0$	1.14 ± 0.13	1.33 ± 0.18	1.01 ± 0.16	1.87 ± 0.52
f_0	1.32 ± 0.43	1.48 ± 0.62	1.05 ± 0.27	1.64 ± 0.66
$\tilde{\tau}$ (yr)	7.53 ± 2.92	7.12 ± 1.97	6.50 ± 1.33	7.37 ± 2.08
τ (yr)	5.25 ± 0.54	5.47 ± 0.63	5.12 ± 0.48	5.84 ± 0.67
$\tilde{A}_0$ (yr ⁻¹)	0.54 ± 0.08	0.35 ± 0.05	0.54 ± 0.05	0.30 ± 0.03
A_0 (yr ⁻¹)	0.33 ± 0.04	0.26 ± 0.03	0.31 ± 0.04	0.23 ± 0.03
$\sigma_{\ln(\eta)}$	0.74	0.58	0.80	0.69
$\tilde{A}(0)\tilde{f}(0)$ (yr ⁻¹)	0.49 ± 0.02	0.47 ± 0.01	0.38 ± 0.01	0.50 ± 0.01
CPC Section	G	H	Y	
$\tilde{\alpha}_0$	0.70 ± 0.03	0.69 ± 0.03	0.62 ± 0.05	
$\tilde{\alpha}_\infty$	0.89 ± 0.04	0.89 ± 0.04	0.84 ± 0.03	
α	1.10 ± 0.08	1.10 ± 0.06	1.10 ± 0.03	
$\tilde{f}_0$	1.16 ± 0.17	1.14 ± 0.15	1.15 ± 0.13	
f_0	1.22 ± 0.48	1.31 ± 0.37	1.35 ± 0.43	
$\tilde{\tau}$ (yr)	6.16 ± 2.82	6.15 ± 2.73	7.03 ± 1.51	
τ (yr)	5.19 ± 1.21	5.25 ± 0.93	5.45 ± 0.41	
$\tilde{A}_0$ (yr ⁻¹)	0.35 ± 0.08	0.34 ± 0.07	0.40 ± 0.04	
A_0 (yr ⁻¹)	0.25 ± 0.05	0.22 ± 0.05	0.27 ± 0.02	
$\sigma_{\ln(\eta)}$	0.61	0.55	0.60	
$\tilde{A}(0)\tilde{f}(0)$ (yr ⁻¹)	0.55 ± 0.01	0.60 ± 0.01	0.49 ± 0.01	

Table 6.3: Technology-classification dependence of parameters associated with residual preferential attachment ($\tilde{\alpha}_0$, $\tilde{\alpha}_\infty$, $\tilde{f}_0$) and obsolescence-related ageing ($\tilde{\tau}$, $\tilde{A}_0$) in the citation rate given by Eq. (6.1), in conjunction with Eqs. (6.25a) & (6.25b). For comparison, parameters in the uncontrolled-for-fitness average rate $\bar{\lambda}_i \equiv (k_i^\alpha + f_0) A_0 \exp\{-t/\tau\}$ fitted to the data are also included. Both the saturation value $\tilde{\alpha}_\infty$ (the weighted average of the latest third among the extracted values for $\tilde{\alpha}$) and the value $\tilde{\alpha}_0$ for $t = 0$ (corresponding to the time of grant) are given. The time $\tilde{\tau}$ given here is the weighted average of obsolescence times extracted for the fixed- k bins. For reference, the empirically extracted standard deviation $\sigma_{\ln(\eta)}$ of the log of fitness η is also given. The values for $\tilde{A}(0)\tilde{f}(0)$ are obtained from the time evolution of uncited patents (Eq. (6.23)). Comparing these with $\tilde{A}_0\tilde{f}_0$ provides a satisfactory consistency check for demonstrating the overall suitability of the current approach.

6.5 Conclusions and Outlook

While node fitness in growing networks has been investigated endogenously in real-world networks, the work detailed in this chapter goes an important step further by determining fitness for individual patents in terms of network-independent quality measures. Obtaining fitness deterministically and exogenously enables the conclusive separation of its effects from those due to preferential attachment and obsolescence-induced ageing. These advances pave the way for broader studies of knowledge diffusion and could also serve to inform the design of meaningful impact measures for technological innovation.

One striking feature exhibited by the fitness-controlled citation rate is the significantly reduced exponent characterizing preferential attachment. Previous theoretical studies have suggested [124,282] that purely fitness-driven growth can cause phenomenologically observed preferential-attachment dynamics. As our estimate of fitness η pertains to the attributes of patents known at the time of grant, it could be expected that the mechanism for attracting citations at that time will be largely reflective of this fitness. In fact, if we believe that citations are indeed a noisy proxy for knowledge flow, then fitness may reflect the intrinsic properties of a patent that indicate the extent to which the knowledge contained in that patent is likely to diffuse, at least initially.

With this being said, $\tilde{\alpha}$ is still significantly different from zero even at the early times in a patent's citation history. It is of course likely that our quantitative fitness measure captures the true heterogeneity in fitness only imperfectly. It is impossible to say what portion of the empirical preferential attachment observed after controlling for observable heterogeneity in fitness represents unmeasured variation in fitness (related to the remaining 'ignorance gap' about patent quality [276]), and what portion represents true preferential attachment (the purely citation-number-driven kind as originally postulated, e.g., in Refs. [1,116]). It is interesting, however, that $\tilde{\alpha}$ rises over time until it plateaus after about 5 years. This pattern is consistent with an interpretation that the observed positive relationship between citation probability and previous citations reflects agents' understanding that accumulated citations convey information about the significance of an invention, with the reliability of that information increasing as time elapses.

With the observation that citation accumulation is in part determined by the fitness values calculated in this chapter, obsolescence times must also increase as a consequence. This is because all obsolescence measurements are conducted by analysing the decay in

the citation rate over time for groups of patents with the same number of citations at each point in time. A set of patents with 10 citations each after 3 months would be considered a highly successful group ($\bar{\eta} > 1$), while a set with the same number of citations each after 10 years would be considered positively mediocre ($\bar{\eta} \leq 1$). Therefore, if the fitness values calculated here are to reflect this difference it is expected that on controlling for these fitnesses, the rate function will be decreased for the former group due to their high fitness, and vice-versa for the latter group. This decreases the slope of the ageing function and hence increases $\tilde{\tau}$, as seen in Fig. 6.6.

Given that there is a kind of fitness ‘gradient’ that decreases as time goes on in these obsolescence measurements (for the same k -bin), the ability to accurately measure fitness for patents with different underlying levels of quality starts to become important. It is possible that due to their large numbers of citations at early times, and because the fitness metric is calibrated at grant, that the patents in the example above that obtain 10 citations very early in their lifetime have more accurately estimated fitness values than those patents that obtain 10 citations much later. Due to the positive skewness of the fitness distribution, this latter group is more likely to have an average fitness that is biased upwards. This leads to the slope of the ageing function being larger (and thus $\tilde{\tau}$ being smaller) than it perhaps should be if fitness were estimated with the same accuracy across all patents. This would be a particular problem for the small- k -bins which at long times contain unsuccessful patents of the type for which there was very little resolution in citation numbers at the time of calibration (leading to less accurate fitness estimation). At the same time, the calibration-at-grant will likely overestimate the importance of particular patent attributes. This would mean the fitness-controlled rate would be too small for patents that are successful early, and this would lead to the slope of the ageing function being smaller (and thus $\tilde{\tau}$ being larger), due to this being a bigger problem at small times when there is less data in these large- k bins for this effect to be mitigated by patents for which fitness is underestimated. These two effects together (underestimation of $\tilde{\tau}$ for small k and overestimation for large k) may produce the upward slope visible in the measurements of $\tilde{\tau}$ displayed in Figures 6.7, 6.10, & 6.11, and may be mitigated in the future by explicitly correcting for these biases given the known discretisation problem and the skewness of the fitness distribution.

An unexpected result in these obsolescence measurements, however, is the near-elimination of the deviation from exponential obsolescence that was observed in Section 4.5.2. It was postulated therein that these deviations may have been due to quality

differences that were not reflected accurately until some citations had been accrued or, alternatively, may be reflective of mechanisms of knowledge diffusion. The fact that these deviations are mostly unobserved after controlling for fitness appears to lend some evidence to the former explanation, but if fitness also plays a large role in the short-time diffusion process then this could also explain the results. More work is required to understand the role of fitness on the citation histories of patents at different points of their lifetime, as evidenced by both the disappearance of short-time deviations and by observations of increasing $\tilde{\alpha}$ leading to saturation after a short period of time (as seen in Fig. 6.7).

The work described in this chapter represents a novel way of integrating well-established knowledge on patent quality from the economics literature with well-established mathematical work on the effect of node heterogeneity on the growth of networks. Additionally, this chapter, in a more succinct form, has been published in Physical Review E as a prestigious ‘Rapid Communication’ [314]. There are many extensions to this research that may better inform us about the nature of patent quality and its role in knowledge diffusion, including but not limited to different empirical techniques, model augmentations, and the addition of new variables. These topics are discussed in Section 7.5.

6.6 Additional Figures for Chapter 6

Figure 6.10: Measuring preferential-attachment-type growth and ageing in the fitness-controlled citation rate of patents granted by the USPTO. Rows represent technology categories (excluding category A, for which the parameter estimates can be seen in Fig. 6.7) while columns represent the parameter measurements.

Rows: Parameter estimates for the cohorts of patents assigned, from top to bottom, categories B, C, and F.

Columns: **Left:** Obsolescence time $\tilde{\tau}$ extracted for each bin of patents with fixed number of citations k from fitting the fitness-controlled citation rate to the form of Eq. (6.25a) for times $t \geq t_0 = 2$ years. **Centre:** Preferential-attachment exponent $\tilde{\alpha}$, extracted for fixed times t from fits of the fitness-controlled citation rate to the form of Eq. (6.25b). **Right:** Constant contribution $\tilde{f}_0$ to the preferential-attachment part of the fitness-controlled citation rate, extracted for fixed times t from fits to Eq. (6.25b). Circles are fit-parameter values, solid lines their weighted averages, and black dashed (red dotted) curves show 95% confidence intervals for fit-parameter values (weighted averages). In the case of $\tilde{\alpha}$, averages and confidence intervals are calculated for the last third of the measurements.

Figure 6.11: Measuring preferential-attachment-type growth and ageing in the fitness-controlled citation rate of patents granted by the USPTO. Rows represent technology categories (excluding Category A, for which the parameter estimates can be seen in Fig. 6.7) while columns represent the parameter measurements.

Rows: Parameter estimates for the cohorts of patents assigned, from top to bottom, categories G, H, and Y.

Columns: **Left:** Obsolescence time $\tilde{\tau}$ extracted for each bin of patents with fixed number of citations k from fitting the fitness-controlled citation rate to the form of Eq. (6.25a) for times $t \geq t_0 = 2$ years. **Centre:** Preferential-attachment exponent $\tilde{\alpha}$, extracted for fixed times t from fits of the fitness-controlled citation rate to the form of Eq. (6.25b). **Right:** Constant contribution $\tilde{f}_0$ to the preferential-attachment part of the fitness-controlled citation rate, extracted for fixed times t from fits to Eq. (6.25b). Circles are fit-parameter values, solid lines their weighted averages, and black dashed (red dotted) curves show 95% confidence intervals for fit-parameter values (weighted averages). In the case of $\tilde{\alpha}$, averages and confidence intervals are calculated for the last third of the measurements.

Chapter 7

Conclusions and Outlook

7.1 Conclusions

Invention and the institutions in which invention occurs are complex, like so many human systems. The work detailed in this thesis constitutes a small step toward understanding a small segment of the innovation ecosystem, and in addition I hope it provides a useful template for productive research at the intersection of physics and economics, two conventionally disparate academic fields. In the course of the research detailed in this thesis, it has become clear to me that analytical and empirical tools are not field-specific, and that the novel application of tools traditionally associated with a different field may deliver great insight into old problems, providing they are applied with both domain-specific context and prior work in mind.

Combining the domain knowledge detailed in Chapter 2 with the novel analytical results in Chapter 3, lead directly to new findings – the reliable separation of growth and ageing mechanisms in citation accrual (Chapter 4) – in a field that has been active for decades. This result was then replicated on a different knowledge-based network (Chapter 5) with different field-specific considerations that needed to be taken into account. Empirical separation of these mechanisms is something that is generally assumed but never tested [35–39]. This work provides a strong foundation for investigating this phenomena to gain greater understanding about the relationship between diffusion and obsolescence on a more granular level. This pathway provides strong evidence for the claim made above: established tools transferred into unconventional fields can yield significant new results, provided the analysis and interpretation is done with the field-specific context in mind.

The work in Chapter 4 was then built on in Chapter 6, wherein new patent-specific information was integrated into the existing models. It was found that by combining relatively simple metadata into a single fitness metric, a large portion (20-30%) of the preferential attachment strength could be explained by some notion of patent quality – a truly novel finding in the field of network science with wide implications for the interpretation of results of any informetric analysis that uses citation data. The results provide not only further insight into the nature of invention and knowledge diffusion but also insight into the growth of other, unrelated, networks. This outcome is worth emphasising – a considerable development in our understanding of network growth generally was only made possible through the understanding and application of domain-specific knowledge typically unrelated to networks. Further, this domain-specific knowledge was already well-established and waiting to be applied in a new way. I find this particularly encouraging; there must exist many other established research fields that are yet to be integrated into a more general framework and have the potential to significantly add to our understanding of other real-world systems altogether unrelated to the original field.

The findings contained in this work and the associated published articles have already found significant traction in a number of fields. Chapters 4 and 5 are updated and expanded versions of my papers published in *Physical Review E* [157] and *Journal of Informetrics* [245] respectively, both among the most highly regarded journals in their respective fields. The work described in Chapter 6 was recently published in *Physical Review E* as a Rapid Communication [314], a category “intended for the accelerated publication of important new results.”¹ These publications have attracted citations of their own from variety of sources, from the economic literature [315] to the network science literature [124, 138]. The methodologies developed and results described in this thesis have even had a not-insignificant impact on the study of human collective memory, in an exciting paper published in *Nature Human Behaviour* [40].

The impact of the work contained in this thesis speaks for itself – there is an interest in these topics across multiple academic communities, which I believe provides a mandate for further research on these topics. For the remainder of this chapter, I will outline some of the extensions and exciting research avenues for the future that have emerged from the work in this thesis and describe some alternate frameworks that shed light on

¹journals.aps.org/pre/authors/editorial-policies-practices#articletypes

the nature of invention from a different angle.

7.2 Biological Approaches to Technological Evolution

Network models, while very useful for the study of knowledge diffusion and obsolescence, do not have a lot to say about why technology *evolves*. There are clear economic and social incentives that drive technology forward, however, both the rate of progression of these developments and the form they take may be better explored within a different framework. Concepts from the field of complex adaptive systems are the most common tools for this task, and they can be used to characterise and understand the ecosystem within which inventions are developed [316]. These tools were not applied to patent data in this thesis, but they may provide a useful framework for thinking about differences in network growth dynamics between different technology types and provide readers with an alternative view of technological development as a random process.

One of the most prevalent models from this literature adapts a framework from the field of evolutionary biology (and is itself analogous to earlier spin glass models [317–320]): the NK landscape [321–324]. This model treats inventors as agents exploring a combinatorial phase space in search of an invention with optimal ‘fitness’² [325,326]. This phase space consists of all possible combinations of N components, where each combination has a certain level of fitness. The NK landscape itself has ‘tunable ruggedness’ [320,322] as defined by the level of interdependence of the components, encapsulated by the parameter K. Inventors can then search this landscape for local or global maxima, which are often assumed to be patentable inventions [327].

This framework is suited to the understanding of how inventions are recombined, rather than the diffusion of those ideas in the first place. The idea that new ideas are often just combinations of older ones is not itself a new concept [328–330], so the concept of an NK landscape as applied to technological evolution is particularly appealing for the study of knowledge recombination on the scale of individual inventions.

Ecology also offers some parallels with technological development, and places more

²This fitness is not defined in precisely the same way as the concept of node fitness in Chapter 6, however, the concepts are related.

emphasis on the industrial and political environments in which inventions exist and interact [331]. There is a particular focus on the development of technological niches [332,333], and this line of enquiry generally assumes that evolution and natural selection are occurring due to external pressures with little concern for the micro-level ‘genetic’ approach that the combinatorial NK approach posits.

These two approaches look at technological development at two different, and complementary, scales; and while there is much interesting information to be garnered by these approaches, it is generally difficult to use them to study the specifics of dynamics of technological diffusion and obsolescence. A notable exception involves the synthesis of technological search on an NK-landscape with the social network in which the process of invention takes place [209]. In this work, the authors examine differences in knowledge diffusion through the different dimensions in which a social connection between inventors may be inferred: co-invention, firm membership, and geographic co-location. It is discovered that the complexity and interdependence (the parameter K from above) of ideas has a substantial effect on the transmission of knowledge in each of these dimensions and that the differences in transmission rates are most significant for inventions of middling complexity.

Diffusion extent and speed through both the social networks and technological networks in the temporal dimension needs to be examined in more depth, however. As described above, the social and industrial networks in which inventors and firms are embedded have a role to play in the facilitation of knowledge diffusion [74, 168, 209]. Despite this, very little work has been done to empirically examine the temporal dimension of knowledge diffusion through these networks, and how their structure affects this diffusion. In other words, how does network structure affect both the speed at which an invention catches on, or the rate at which it becomes obsolete? Answering these questions may provide understanding that allows the adoption of systems, programmes, or policies designed to maximise both the speed and extent of knowledge diffusion within and between industries to the benefit of all.

7.3 Changing Citation Behaviours & Network Structure

A conspicuous omission from this thesis is the comparison of the citation dynamics of cohorts of patents granted in different years. While this kind of temporal analysis was

not a goal of this research, it would be straightforward to extend this analysis in this way – providing special care is taken to control for, or examine, the effect of significant changes in patent law or patent office procedures and policies. The fitness analysis may be particularly useful for this purpose, as it provides an opportunity to look at the predictability of patent success, and changes in the determinants of this success. A comparison of cohorts would also allow for the analysis of changes in citation behaviours and how these are related to changes in the systems in which they are made. For example, do the citation networks constructed in different periods reflect changes in innovation policy? If so, at which scale are these differences most substantial: mesoscopic or macroscopic? We saw, for example, in comparing the patent and scientific citation networks that while the mesoscopic network structures are very different [95], the network growth and obsolescence behaviours on the macroscopic scale are very similar. A similar comparison could be made between different periods of the same network, and may lead to insights about the relationships between the two scales.

7.4 Comparisons to be Made (or Not)

7.4.1 Scientific Citations

Much of the work in this thesis was made possible by the standardised, systematic, and public recording of patent information. Scientific research and its associated records, although often funded by taxpayers, is not as easy to access (to the detriment of almost all parties) and often not systematised in a way that is useful for science-of-science [26] researchers. A fitness analysis as detailed in Chapter 6, for example, would be very difficult to conduct on scientific articles without a significant narrowing of topical scope, and thus a potential reduction in generality. Given this limitation, however, there is still much to be done that can utilise the information that is available.

Most pertinent to the research detailed in this thesis is further work on the important differences in network structure between the scientific and patent citation networks. The structural differences outlined in Section 7.3 illustrate the dangers of drawing conclusions about network growth mechanisms through analysis of a single scale, while also providing insight into the importance of the various network-growth mechanisms. After all, the two networks both represent the structure of knowledge flows, and so perhaps the specific local mechanisms, when aggregated, lose all significance and give way to the

macroscopic behaviours common to both networks. The changing importance of different knowledge diffusion mechanisms on different scales should be examined in more depth to understand better the human aspects of these mechanisms that rise above the systems in which they are acting.

Another interesting avenue is the investigation of changes in scientific citation behaviour through time, and the effect of new tools for literature search with special attention on preferential attachment and obsolescence. In particular, do search engines make science more meritocratic, or less? On the one hand, it is easier to find lesser-known papers via, for example, keyword search. On the other hand, the ranking algorithms explicitly weight highly-cited papers, leading directly to a real preferential attachment mechanism and perhaps boosting its effect. One may hypothesise, for example, that these search engines have proven a boon for articles with a moderate number of citations, but have done little to improve the prospects of less cited works. This kind of question may be addressed with the methodologies described in this thesis.

7.4.2 European Patent Office Data

The data used in this thesis comes primarily from the United States Patent and Trademarks Office (USPTO), but there are other jurisdictions whose patents and associated metadata may be analysed similarly to those of the USPTO. The most accessible and well-studied patents granted in these other jurisdictions are those granted by the European Patent Office (EPO), and as such, it is important to justify the restriction of the scope of the analyses in this thesis to USPTO-granted patents.

The EPO grants *European Patents*, which are valid in the subset of EPO-serviced countries selected by the applicant³. There is no duty of candor, and thus little incentive for applicants to include their knowledge of relevant prior art in their patent application unless it helps to establish the patentability of the invention in question [334]. This means that before looking at differences in examination procedures, the argument that citations are indicative of knowledge diffusion is weakened and results stemming from EPO citation data require significantly different interpretation than those derived from USPTO data [149].

³EPO patents are, therefore, not a single patent but a bundle of patents.

The procedures for prior-art search at the EPO also differ significantly from the USPTO. Most pertinent for the kind of analysis in this thesis, the EPO aims specifically to cite all relevant patents in as few citations as possible [52]. This, combined with the lack of a duty of candor, results in 60-90% fewer citations on EPO-granted patents when compared to USPTO-granted patents [52, 149]. Therefore, even if measuring preferential attachment strength in citation accrual is the only goal, EPO citation data offers significantly less spread in patent citation counts with which to make these measurements.

In short, differences in citation practice between the EPO and USPTO make it difficult (and in some contexts, nonsensical), to compare results of the same analyses between these jurisdictions. For this reason, the decision was made to restrict the focus of the analyses of patent citations to those made by and to USPTO-granted patents. While this is a limitation of this work, this choice is justified by the increased citation-count-resolution afforded by the USPTO data, the inclusion of applicant-provided prior art, and the larger body of existing academic research focusing on this jurisdiction that can be used for reference and comparison. This latter point perhaps itself embodying a kind of preferential attachment effect in favour of the USPTO data.

7.5 Extending Patent Fitness

The research detailed in Chapter 6 provides a blueprint for investigating the effect of patent quality on the growth of the citation network. This is a wholly novel approach and is therefore limited in scope so as to focus on the essential aspects that demonstrate its utility. It was a conscious decision to ensure that, to draw an analogy, the foundations of this house were stable before building additional stories. Now that this approach appears to be stable and useful, it is possible to suggest augmentations and modifications that may yield exciting new results.

Oblique Fitness Dimensions

In principle, we could use the results from our factor analysis in Section 6.3.1 to compare the different latent factors that are related to technological impact for different types

of inventions, but our methodology is not optimised for this purpose. In order to do this more rigorously, an oblique factor rotation could be utilised after the factors are extracted. In the current methodology detailed in that chapter, the factors are rotated while ensuring they remain orthogonal – this is merely a restriction on the form of the transformation matrix. This restriction ensures that the factors obtained after the scoring procedure are as uncorrelated as possible given the data. Orthogonality of factors is a suitable restriction for our purposes, as it allows the reasonable linear combination of factors that make up the fitness metric η to be mostly devoid of double-counted patent characteristics; therefore, orthogonality is ultimately retained for convenience and does not necessarily reflect the real-world dimensions of patent quality. Orthogonality may even introduce unwanted artefacts into the factor loadings as a result of this, as there may exist unobserved common influences on the factors themselves. This would not only mean an oblique set of factors may be the most interpretable structure to explain the observed variance in the data, but that an orthogonal set of factors would be more difficult to interpret due to the adjustments that must be made to the factor loadings to ensure this forced orthogonality of oblique dimensions [298].

Ideally, if integrating the patent quality dimensions into a growing-network framework was not the goal, oblique rotations and ex-post variables would add a lot of information and allow the resultant factors to be more easily interpreted. The simultaneity problems that ruled out using ex-post variables in the network growth model would no longer be relevant and forward citation counts could actually be included in the factor analysis itself.

Ex-Post Variables and Dynamic Fitness

The inclusion of ex-post variables also opens the door to more complex models of patent fitness, including the dynamic view that doesn't enforce the static intrinsic fitness obtained in Chapter 6. Information about the real utility of an invention may not be revealed until long after it is granted, and ex-post variables may provide the opportunity to capture the dynamic nature of the state of this knowledge. The dynamics can be built into the model either at the factor level, where the compositions of the factors change with time, or at the fitness level, where the factors remain the same, but their weights in the fitness metric can change with time. The appropriate approach depends on what the goal of the investigation is: the former may be better for understanding

the changeable nature of the dimensions of quality (which may be best captured by an oblique rotation), while the latter may be better suited for incorporation into the growing-network model investigated in this thesis. As discussed in Section 6.3.3, the point in time at which we define $t = 0$ is almost arbitrary, and therefore it may be possible to continuously update the composition of the fitness metric over time, providing simultaneity is dealt with adequately.

Inclusion of Additional Variables

Many potential patent-quality indicators were not included in the fitness analysis, of both the ex-ante and ex-post varieties. These are briefly covered below to indicate potential new or seldom-used sources of information that could be used in future work on this topic, in conjunction with both the existing set of ex-ante quality indicators and the common ex-post indicators briefly discussed in Section 6.1.1.

Bibliometric/event-based indicators. Some bibliometric or event-based indicators have been omitted from this thesis for practical reasons, including the characteristics of the inventors [335] and firms [335–337] associated with each patent. The main reason for this is to avoid loss of data due to undefined input variables – many patents are not assigned to a firm, for example, and many inventors do not have an invention history from which to extract information. For a smaller, select sample of patents for which variables derived from these attributes can be defined for the vast majority, however, there is much potential in the use of this data. This is also true for stock market-based [67, 78], auction-based [268], or licensing-based [286] indicators for which data is only available for a select group of patents that are not usually a representative sample.

Patent examination data are generally available for all patents, but this requires a large amount of pre-processing conducted with a deep understanding of the examination procedure. The information required to construct indicators from examination data is generally easily accessible, and includes both categorical variables such as the patenting route (such as direct application to the respective offices or via the Patent Cooperation Treaty process) or whether the application is a continuation or a divisional patent, and ordinal variables derived from the interactions between the examiners and the inventors/attorneys applying for the patent. This information, while not generally suited to factor analysis (although this may be remedied using a multiple correspon-

dence analysis [338]), has been found to correlate with more established patent quality indicators [264].

Although not possible for the vast majority of patents considered in this thesis, backward citations originating with applicants can be distinguished from those added by the examiner for all patents granted after 2001. It would be useful to incorporate consideration of the differences between these two types of citations [44, 87] in future work.

Network-based indicators. It has been noted at various points in this thesis that the notion that citations form a network was merely a helpful mental framework and did not influence the analyses conducted. This fact is still true for Chapter 6, even if node-fitness as a way to envision heterogeneity of inventions makes the network framework even more conceptually useful than in previous chapters. Taking into account network position and the local network environment of patents in the citation network may prove useful as indicators, as these are inherently node-level characteristics and thus ideal for integration into the fitness metrics.

A notable omission from this class of indicators is homophily – the tendency of nodes to link to other nodes with similar characteristics. Building this tendency into the model would effectively eliminate the need to disaggregate patents by their technology category, as the differences observed between these categories could be captured by some function that takes into account the probabilities of new citations forming between patents with certain intrinsic properties. This is a challenging task; however, it has been approached in other contexts (both theoretical and empirical) in many interesting ways [123, 339–341]. Additionally, community membership is closely related this concept [342], and thus homophily will likely explain much of the observed community structure in patent networks [159, 343].

Individual-node-level network-based characteristics that are much less daunting to calculate includes local clustering and centrality measures. Local clustering in citation networks can be calculated using a simple triangle-based approach [141] to check for citation links between two patents that are both cited by the same patent.⁴ Clustering may be associated with the existence of patent thickets [33, 344] and indicate a particularly

⁴Often this is achieved in directed networks by counting cycles [142], but these do not generally exist in citation networks.

crowded technological space. Centrality is a whole class of metrics that each capture some notion of the importance of each node to the structure of the networks. PageRank centrality [345], for example, is the result of a random walk on a network (mathematically this can be reduced to an eigen-decomposition), a computationally efficient way of ranking nodes based on the probability of passing through each of them during this random walk. The standard algorithm needs to be adapted somewhat for directed acyclic networks and has already been investigated in patent networks [210, 246]. Other common centrality measures that could be applied to fitness in patent citation networks are betweenness centrality [346], closeness centrality [347], and eigenvector centrality [348], each of which may reflect some new aspect of fitness related to network position, and in some cases may only be possible to calculate for cited patents (c.f. the Pedigree metric defined in Section 6.2).

Networks derived from patent metadata may also provide information about patent quality, and in particular, those derived the projections of bipartite networks [111, 349] such as the co-invention network of inventors, the co-assignment network of firms, and the co-classification network of technology categories, in addition to the individual characteristics of these entities. The network location of these metadata in their respective networks may have a large part to play in informing both the nature of knowledge flow between people [74, 169] and the effect of unique combinations of expertise or technologies [209]. An interesting piece of work in this vein is the recent study of Morgan et al. [350], the results of which highlight the institutional inequalities inherent in the diffusion of scientific knowledge. Additionally, citation flow between different firms and technology categories can potentially provide information on network structures reflective of knowledge flow [351].

Natural-Language-based indicators. Finally, while there has been some recent activity on the classification of patents through an examination of the textual information contained in them [177, 352], little appears to have been done to use this information to gain information about patent quality. Interesting results may be garnered via measures of, for example, the vagueness of the invention descriptions via a kind of sentiment analysis [353], or analysis of particular elements related to the language structures within the written descriptions (e.g. [354–356]). Other approaches may include ease of reading or other attributes that take advantage of the fact that both attorneys and examiners are human, and their citing behaviour may be somewhat dependent on the writing-style characterising the way particular inventions are described. Topic models [2] in partic-

ular lend themselves nicely to bipartite network construction (and projections thereof) in a similar fashion to the co-classification network of technology categories mentioned above [357]. These methodologies may be able to produce more natural and robust representations of technological relatedness, while also providing an alternative approach to technological distance and homophily through semantic relatedness [357] – this could prove particularly useful if one is attempting to disentangle relatedness from fitness and preferential attachment effects [123].

7.6 Parting Words

There are many ways to extend and refine the methodologies and results described in this thesis; it is a starting point for novel ways of thinking about both the empirical study of growing networks and the meaning of citations in this context. For example, if we think that obsolescence and knowledge diffusion are distinct and unrelated processes, then these assumptions should be made explicit, or better yet, tested. Further, if we believe that preferential attachment is simply a stand-in generative mechanism for a menagerie of unobserved processes and behaviours, then this should be stated, so readers know how to interpret results. It is my hope that these fundamental assumptions, among others, about the nature of citation accrual and the context in which it takes place have been confronted head-on and given the attention they deserve.

With luck, I look forward to being part of some of the efforts that will augment the work in this thesis.

Bibliography

- [1] Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
- [2] Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet allocation. *Journal of Machine Learning Research* **3**, 993–1022 (2003).
- [3] Vapnik, V. *The nature of statistical learning theory* (Springer Science & Business Media, 2013).
- [4] Barabasi, A.-L. & Oltvai, Z. N. Network biology: Understanding the cell’s functional organization. *Nature Reviews Genetics* **5**, 101 (2004).
- [5] Castellano, C., Fortunato, S. & Loreto, V. Statistical physics of social dynamics. *Reviews of Modern Physics* **81**, 591 (2009).
- [6] Lazer, D. *et al.* Computational social science. *Science* **323**, 721–723 (2009).
- [7] Rocha, L. E., Liljeros, F. & Holme, P. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLOS Computational Biology* **7**, e1001109 (2011).
- [8] Bettencourt, L. M. The origins of scaling in cities. *Science* **340**, 1438–1441 (2013).
- [9] Wang, Z. *et al.* Statistical physics of vaccination. *Physics Reports* **664**, 1–113 (2016).
- [10] Morehouse, N. F., Strotz, R. H. & Horwitz, S. An electro-analog method for investigating problems in economic dynamics: Inventory oscillations. *Econometrica: Journal of the Econometric Society* 313–328 (1950).
- [11] Anderson, P. W. *The economy as an evolving complex system* (CRC Press, 2018).

- [12] Caballero, R. J. & Jaffe, A. B. How high are the giants' shoulders: An empirical assessment of knowledge spillovers and creative destruction in a model of economic growth. In *NBER Macroeconomics Annual*, vol. 8, 15–86 (MIT Press, 1993).
- [13] Narin, F. Bibliometrics. In *Annual review of information science and technology* (Citeseer, 1977).
- [14] Ellis, P., Hepburn, G. & Oppenheim, C. Studies on patent citation networks. *Journal of Documentation* **34**, 12–20 (1978).
- [15] Jaffe, A. B. & de Rassenfosse, G. Patent citation data in social science research: Overview and best practices. *Journal of the Association for Information Science and Technology* **68**, 1360–1374 (2017).
- [16] Gallegati, M., Keen, S., Lux, T. & Ormerod, P. Worrying trends in econophysics. *Physica A: Statistical Mechanics and its Applications* **370**, 1–6 (2006).
- [17] Jungherr, A. & Theocharis, Y. The empiricist's challenge: Asking meaningful questions in political science in the age of big data. *Journal of Information Technology & Politics* **14**, 97–109 (2017).
- [18] Willinger, W., Alderson, D. & Doyle, J. C. Mathematics and the internet: A source of enormous confusion and great potential. *Notices of the American Mathematical Society* **56**, 586–599 (2009).
- [19] Caldarelli, G., Capocci, A., De Los Rios, P. & Munoz, M. A. Scale-free networks from varying vertex intrinsic fitness. *Physical Review Letters* **89**, 258702 (2002).
- [20] Vázquez, A. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E* **67**, 056104 (2003).
- [21] Fortunato, S., Flammini, A. & Menczer, F. Scale-free network growth by ranking. *Physical Review Letters* **96**, 218701 (2006).
- [22] D'Souza, R. M., Borgs, C., Chayes, J. T., Berger, N. & Kleinberg, R. D. Emergence of tempered preferential attachment from optimization. *Proceedings of the National Academy of Sciences* **104**, 6112–6117 (2007).
- [23] Lima-Mendez, G. & van Helden, J. The powerful law of the power law and other myths in network biology. *Molecular BioSystems* **5**, 1482–1493 (2009).

- [24] Stumpf, M. P. & Porter, M. A. Critical truths about power laws. *Science* **335**, 665–666 (2012).
- [25] Hidalgo, C. A. Disconnected, fragmented, or united? A trans-disciplinary review of network science. *Applied Network Science* **1**, 6 (2016).
- [26] Fortunato, S. *et al.* Science of science. *Science* **359** (2018).
- [27] Lane, J. Let's make science metrics more scientific. *Nature* **464**, 488 (2010).
- [28] Jaffe, A. B., Trajtenberg, M. & Henderson, R. Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics* **108**, 577–598 (1993).
- [29] Börner, K., Penumarthy, S., Meiss, M. & Ke, W. Mapping the diffusion of scholarly knowledge among major us research institutions. *Scientometrics* **68**, 415–426 (2006).
- [30] Albert, M. B., Avery, D., Narin, F. & McAllister, P. Direct validation of citation counts as indicators of industrially important patents. *Research Policy* **20**, 251–259 (1991).
- [31] Waltman, L. A review of the literature on citation impact indicators. *Journal of informetrics* **10**, 365–391 (2016).
- [32] Redner, S. Citation statistics from 110 years of physical review. *Physics Today* **58**, 49 (2005).
- [33] Strandburg, K. J., Csárdi, G., Tobochnik, J. & Érdi, P. Law and the science of networks: An overview and an application to the patent explosion. *Berkeley Technology Law Journal* **21**, 1293 (2006).
- [34] Yin, Y. & Wang, D. The time dimension of science: Connecting the past to the future. *Journal of Informetrics* **11**, 608–621 (2017).
- [35] Dorogovtsev, S. N. & Mendes, J. F. F. Evolution of networks with aging of sites. *Physical Review E* **62**, 1842 (2000).
- [36] Hajra, K. B. & Sen, P. Modelling aging characteristics in citation networks. *Physica A: Statistical Mechanics and its Applications* **368**, 575–582 (2006).
- [37] Valverde, S., Solé, R. V., Bedau, M. A. & Packard, N. Topology and evolution of technology innovation networks. *Physical Review E* **76**, 056118 (2007).

- [38] Golosovsky, M. & Solomon, S. Stochastic dynamical model of a growing citation network based on a self-exciting point process. *Physical Review Letters* **109**, 098701 (2012).
- [39] Wu, Y., Fu, T. Z. & Chiu, D. M. Generalized preferential attachment considering aging. *Journal of Informetrics* **8**, 650–658 (2014).
- [40] Candia, C., Jara-Figueroa, C., Rodriguez-Sickert, C., Barabási, A.-L. & Hidalgo, C. A. The universal decay of collective memory and attention. *Nature Human Behaviour* **3**, 82 (2019).
- [41] Holme, P. & Saramäki, J. Temporal networks. *Physics reports* **519**, 97–125 (2012).
- [42] Li, G.-C. *et al.* Disambiguation and co-authorship networks of the US patent inventor database (1975–2010). *Research Policy* **43**, 941–955 (2014).
- [43] Jaffe, A. B., Trajtenberg, M. & Fogarty, M. S. Knowledge spillovers and patent citations: Evidence from a survey of inventors. *American Economic Review* **90**, 215–218 (2000).
- [44] Cotropia, C. A., Lemley, M. A. & Sampat, B. Do applicant patent citations matter? *Research Policy* **42**, 844–854 (2013).
- [45] Graham, D. & Bachmann, T. T. *Ideation: The birth and death of ideas* (John Wiley & Sons, 2004).
- [46] Cooper, R. G. & Edgett, S. Ideation for product innovation: What are the best methods. *PDMA visions magazine* **1**, 12–17 (2008).
- [47] Rogers, E. M. *Diffusion of innovations* (Simon and Schuster, 2010).
- [48] Krugman, P. R. *Geography and trade* (MIT Press, 1993).
- [49] Jaffe, A. B. & Trajtenberg, M. *Patents, citations, and innovations: A window on the knowledge economy* (MIT Press, 2002).
- [50] Frumkin, M. The origin of patents. *Journal of the Patent Office Society* **27**, 143 (1945).
- [51] Prager, F. D. Brunelleschi's patent. *Journal of the Patent Office Society* **28**, 109 (1946).

- [52] Michel, J. & Bettels, B. Patent citation analysis. A closer look at the basic input data from patent search reports. *Scientometrics* **51**, 185–201 (2001).
- [53] Marco, A. C., Carley, M., Jackson, S. & Myers, A. F. The USPTO historical patent data files: Two centuries of innovation. *USPTO Economic Working Paper* (2015).
- [54] Marco, A. C., Sarnoff, J. D. & deGrazia, C. Patent claims and patent scope. *USPTO Economic Working Paper* (2016).
- [55] U.S. Const. art. I, §8, cl. 8.
- [56] USC 66 Stat. 792.
- [57] Fordis, J. B. & Sung, L. M. How to avoid patent rejection. *Nature Biotechnology* **13**, 42 (1995).
- [58] Schmookler, J. *Invention and economic growth* (Harvard University Press, 1966).
- [59] Schmookler, J. The level of inventive activity. *The Review of Economics and Statistics* **6**, 183–190 (1954).
- [60] Schmookler, J. Economic sources of inventive activity. *The Journal of Economic History* **22**, 1–20 (1962).
- [61] Pakes, A. & Griliches, Z. Patents and R&D at the firm level: A first report. *Economics Letters* **5**, 377–381 (1980).
- [62] Griliches, Z. Market value, R&D, and patents. *Economics Letters* **7**, 183–187 (1981).
- [63] Pakes, A. Patents, R&D, and the stock market rate of return: A summary of some empirical results. In *R&D, Patents, and Productivity*, 253–260 (University of Chicago Press, 1984).
- [64] Hausman, J., Hall, B. H. & Griliches, Z. Econometric models for count data with an application to the patents-R&D relationship. *Econometrica: Journal of the Econometric Society* 909–938 (1984).
- [65] Pakes, A. & Griliches, Z. Patents and R&D at the firm level: A first look. In *R&D, patents, and productivity*, 55–72 (University of Chicago Press, 1984).

- [66] Griliches, Z. Patent statistics as economic indicators: A survey. *Journal of Economic Literature* **28**, 1661–1707 (1990).
- [67] Hall, B. H., Jaffe, A. & Trajtenberg, M. Market value and patent citations. *RAND Journal of Economics* **36**, 16–38 (2005).
- [68] Mansfield, E. Patents and innovation: An empirical study. *Management Science* **32**, 173–181 (1986).
- [69] Blair, R. D. & Cotter, T. F. An economic analysis of damages rules in intellectual property law. *William & Mary Law Review* **39** (1998).
- [70] Ass’n for Molecular Pathology v. Myriad 133 US 2107 (2013).
- [71] Jaffe, A. B. & Lerner, J. *Innovation and its discontents: How our broken patent system is endangering innovation and progress, and what to do about it* (Princeton University Press, 2004).
- [72] Hall, B. H. Exploring the patent explosion. *The Journal of Technology Transfer* **30**, 35–48 (2004).
- [73] Kesan, J. P. & Gallo, A. A. The political economy of the patent system. *North Carolina Law Review* **87**, 1341 (2008).
- [74] Breschi, S. & Lissoni, F. Knowledge networks from patent data. In *Handbook of quantitative science and technology research*, 613–643 (Springer, 2004).
- [75] Jaffe, A. B. & Trajtenberg, M. International knowledge flows: Evidence from patent citations. *Economics of Innovation and New Technology* **8**, 105–136 (1999).
- [76] Maurseth, P. B. & Verspagen, B. Knowledge spillovers in Europe: A patent citations analysis. *The Scandinavian Journal of Economics* **104**, 531–545 (2002).
- [77] Jaffe, A. B. & Trajtenberg, M. Flows of knowledge from universities and federal laboratories: Modeling the flow of patent citations over time and across institutional and geographic boundaries. *Proceedings of the National Academy of Sciences* **93**, 12671–12677 (1996).
- [78] Kogan, L., Papanikolaou, D., Seru, A. & Stoffman, N. Technological innovation, resource allocation, and growth. *The Quarterly Journal of Economics* **132**, 665–712 (2017).

- [79] Lanjouw, J. O. & Schankerman, M. Characteristics of patent litigation: A window on competition. *RAND Journal of Economics* 129–151 (2001).
- [80] Harhoff, D., Narin, F., Scherer, F. M. & Vopel, K. Citation frequency and the value of patented inventions. *The Review of Economics and Statistics* **81**, 511–515 (1999).
- [81] Maurseth, P. B. Lovely but dangerous: The impact of patent citations on patent renewal. *Economics of Innovation and New Technology* **14**, 351–374 (2005).
- [82] Reitzig, M., Henkel, J. & Heath, C. On sharks, trolls, and their patent prey: unrealistic damage awards and firms strategies of being infringed. *Research Policy* **36**, 134–154 (2007).
- [83] Schumpeter, J. A. *Capitalism, Socialism and Democracy* (Harper & Brothers, 1942).
- [84] Harhoff, D., Scherer, F. M. & Vopel, K. Citations, family size, opposition and the value of patent rights. *Research Policy* **32**, 1343–1363 (2003).
- [85] Harhoff, D., Scherer, F. M. & Vopel, K. Exploring the tail of patented invention value distributions. In *Economics, law and intellectual property*, 279–309 (Springer, 2003).
- [86] Trajtenberg, M. A penny for your quotes: Patent citations and the value of innovations. *The RAND Journal of Economics* **21**, 172–187 (1990).
- [87] Moser, P., Ohmstedt, J. & Rhode, P. W. Patent citations – an analysis of quality differences and citing practices in hybrid corn. *Management Science* **64**, 1926–1940 (2017).
- [88] Trajtenberg, M., Henderson, R. & Jaffe, A. University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and New Technology* **5**, 19–50 (1997).
- [89] Carpenter, M. P., Narin, F. & Woolf, P. Citation rates to technologically important patents. *World Patent Information* **3**, 160–163 (1981).
- [90] Cockburn, I. M., Kortum, S. & Stern, S. Are all patent examiners equal? The impact of examiner characteristics. Working Paper 8980, National Bureau of Economic Research (2002).

- [91] Lemley, M. A. & Sampat, B. Examiner characteristics and patent office outcomes. *Review of Economics and Statistics* **94**, 817–827 (2012).
- [92] Alcácer, J., Gittelman, M. & Sampat, B. Applicant and examiner citations in US patents: An overview and analysis. *Research Policy* **38**, 415–427 (2009).
- [93] Roach, M. & Cohen, W. M. Lens or prism? Patent citations as a measure of knowledge flows from public research. *Management Science* **59**, 504–525 (2013).
- [94] Golosovsky, M. & Solomon, S. The transition towards immortality: Non-linear autocatalytic growth of citations to scientific papers. *Journal of Statistical Physics* **151**, 340–354 (2013).
- [95] Clough, J. R., Gollings, J., Loach, T. V. & Evans, T. S. Transitive reduction of citation networks. *Journal of Complex Networks* **3**, 189–203 (2015).
- [96] Watts, D. J. & Strogatz, S. H. Collective dynamics of small-world networks. *Nature* **393**, 440 (1998).
- [97] Albert, R., Jeong, H. & Barabási, A.-L. Internet: Diameter of the world-wide web. *Nature* **401**, 130 (1999).
- [98] Barabási, A.-L., Albert, R. & Jeong, H. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications* **272**, 173–187 (1999).
- [99] Newman, M. E. & Watts, D. J. Scaling and percolation in the small-world network model. *Physical Review E* **60**, 7332 (1999).
- [100] Kahn, A. B. Topological sorting of large networks. *Communications of the ACM* **5**, 558–562 (1962).
- [101] Garfield, E. & Sher, I. H. Diagonal display - a new technique for graphic representation of complex topological networks. Technical Report, Institute for Scientific Information Inc. (1967).
- [102] Garfield, E. Citation indexes in sociological and historical research. *American Documentation* **14**, 289–291 (1963).
- [103] Garfield, E. “Science Citation Index” - A new dimension in indexing. *Science* **144**, 649–654 (1964).

- [104] Garfield, E. *et al.* Can citation indexing be automated? In *Statistical association methods for mechanized documentation, symposium proceedings*, vol. 269, 189–192 (National Bureau of Standards, Miscellaneous Publication 269, Washington, DC, 1965).
- [105] Price, D. J. d. S. Networks of scientific papers. *Science* **149**, 510–515 (1965).
- [106] Garfield, E. *et al.* Citation indexing for studying science. *Nature* **227**, 669–671 (1970).
- [107] Dunlop, L. & Oppenheim, C. The effect of recycling on a patent citation network. *World Patent Information* **2**, 100–102 (1980).
- [108] Narin, F., Carpenter, M. P. & Woolf, P. Technological performance assessments based on patents and patent citations. *IEEE Transactions on Engineering Management* **31**, 172–183 (1984).
- [109] Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F. & White, D. R. Economic networks: What do we know and what do we need to know? *Advances in Complex Systems* **12**, 407–422 (2009).
- [110] Schweitzer, F. *et al.* Economic networks: The new challenges. *Science* **325**, 422–425 (2009).
- [111] Vasques Filho, D. & O’Neale, D. R. Degree distributions of bipartite networks and their projections. *Physical Review E* **98**, 022307 (2018).
- [112] Bang-Jensen, J. & Gutin, G. Z. *Digraphs: Theory, algorithms and applications* (Springer Science & Business Media, 2008).
- [113] Krapivsky, P. L., Redner, S. & Leyvraz, F. Connectivity of growing random networks. *Physical Review Letters* **85**, 4629 (2000).
- [114] Dorogovtsev, S. N., Mendes, J. F. F. & Samukhin, A. N. Structure of growing networks with preferential linking. *Physical Review Letters* **85**, 4633–4636 (2000).
- [115] Erdős, P. & Rényi, A. On random graphs, I. *Publicationes Mathematicae (Debrecen)* **6**, 290–297 (1959).
- [116] Price, D. d. S. A general theory of bibliometric and other cumulative advantage processes. *Journal of the Association for Information Science and Technology* **27**, 292–306 (1976).

- [117] Simon, H. A. On a class of skew distribution functions. *Biometrika* **42**, 425–440 (1955).
- [118] Merton, R. K. The Matthew effect in science: The reward and communication systems of science are considered. *Science* **159**, 56–63 (1968).
- [119] Dorogovtsev, S. N. & Mendes, J. F. Evolution of networks. *Advances in Physics* **51**, 1079–1187 (2002).
- [120] Eisenberg, E. & Levanon, E. Y. Preferential attachment in the protein network evolution. *Physical Review Letters* **91**, 138701 (2003).
- [121] Jeong, H., Nédá, Z. & Barabási, A.-L. Measuring preferential attachment in evolving networks. *Europhysics Letters* **61**, 567 (2003).
- [122] Krapivsky, P. L. & Redner, S. Network growth by copying. *Physical Review E* **71**, 036118 (2005).
- [123] Papadopoulos, F., Kitsak, M., Serrano, M. Á., Boguná, M. & Krioukov, D. Popularity versus similarity in growing networks. *Nature* **489**, 537 (2012).
- [124] Golosovsky, M. Mechanisms of complex network growth: Synthesis of the preferential attachment and fitness models. *Physical Review E* **97**, 062310 (2018).
- [125] Kong, J. S., Sarshar, N. & Roychowdhury, V. P. Experience versus talent shapes the structure of the web. *Proceedings of the National Academy of Sciences* (2008).
- [126] Zhu, H., Wang, X. & Zhu, J.-Y. Effect of aging on network structure. *Physical Review E* **68**, 056121 (2003).
- [127] Hajra, K. B. & Sen, P. Phase transitions in an aging network. *Physical Review E* **70**, 056103 (2004).
- [128] Berger, N., Borgs, C., Chayes, J. T., D’Souza, R. M. & Kleinberg, R. D. Competition-induced preferential attachment. In *International colloquium on automata, languages, and programming*, 208–221 (Springer, 2004).
- [129] Leskovec, J., Kleinberg, J. & Faloutsos, C. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining*, 177–187 (ACM, 2005).

- [130] Ding, Y. *et al.* Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology* **65**, 1820–1833 (2014).
- [131] Liao, H., Mariani, M. S., Medo, M., Zhang, Y.-C. & Zhou, M.-Y. Ranking in evolving complex networks. *Physics Reports* **689**, 1–54 (2017).
- [132] Aho, A. V., Garey, M. R. & Ullman, J. D. The transitive reduction of a directed graph. *SIAM Journal on Computing* **1**, 131–137 (1972).
- [133] Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S. & Tomkins, A. S. The web as a graph: Measurements, models, and methods. In *International computing and combinatorics conference*, 1–17 (Springer, 1999).
- [134] Vázquez, A. Disordered networks generated by recursive searches. *Europhysics Letters* **54**, 430 (2001).
- [135] Krapivsky, P. L. & Redner, S. Organization of growing random networks. *Physical Review E* **63**, 066123 (2001).
- [136] Evans, T. & Saramäki, J. Scale-free networks from self-organization. *Physical Review E* **72**, 026138 (2005).
- [137] Ren, F.-X., Shen, H.-W. & Cheng, X.-Q. Modeling the clustering in citation networks. *Physica A: Statistical Mechanics and its Applications* **391**, 3533–3539 (2012).
- [138] Golosovsky, M. & Solomon, S. Growing complex network of citations of scientific papers: Modeling and measurements. *Physical Review E* **95**, 012324 (2017).
- [139] Holme, P. & Kim, B. J. Growing scale-free networks with tunable clustering. *Physical Review E* **65**, 026107 (2002).
- [140] Wu, Z.-X. & Holme, P. Modeling scientific-citation patterns and other triangle-rich acyclic networks. *Physical Review E* **80**, 037101 (2009).
- [141] Newman, M. E. Clustering and preferential attachment in growing networks. *Physical Review E* **64**, 025102 (2001).
- [142] Fagiolo, G. Clustering in complex directed networks. *Physical Review E* **76**, 026107 (2007).

- [143] Hu, A. G. & Jaffe, A. B. Patent citations and international knowledge flow: The cases of Korea and Taiwan. *International Journal of Industrial Organization* **21**, 849–880 (2003).
- [144] Peri, G. Determinants of knowledge flows and their effect on innovation. *Review of Economics and Statistics* **87**, 308–322 (2005).
- [145] Popp, D. Lessons from patents: Using patents to measure technological change in environmental models. *Ecological Economics* **54**, 209–226 (2005).
- [146] Popp, D. They don't invent them like they used to: An examination of energy patent citations over time. *Economics of Innovation and New Technology* **15**, 753–776 (2006).
- [147] Popp, D. International innovation and diffusion of air pollution control technologies: The effects of NOX and SO2 regulation in the US, Japan, and Germany. *Journal of Environmental Economics and Management* **51**, 46–71 (2006).
- [148] Bacchiocchi, E. & Montobbio, F. Knowledge diffusion from university and public research. A comparison between US, Japan and Europe using patent citations. *The Journal of Technology Transfer* **34**, 169–181 (2009).
- [149] Bacchiocchi, E. & Montobbio, F. International knowledge diffusion and home-bias effect: Do USPTO and EPO patent citations tell the same story? *Scandinavian Journal of Economics* **112**, 441–470 (2010).
- [150] Verdolini, E. & Galeotti, M. At home and abroad: An empirical analysis of innovation and diffusion in energy technologies. *Journal of Environmental Economics and Management* **61**, 119–134 (2011).
- [151] Adams, J. D. & Clemmons, J. R. How rapidly does science leak out? A study of the diffusion of fundamental ideas. *Journal of Human Capital* **7**, 191–229 (2013).
- [152] Marco, A. C. The dynamics of patent citations. *Economics Letters* **94**, 290–296 (2007).
- [153] Bass, F. M. A new product growth for model consumer durables. *Management Science* **15**, 215–227 (1969).
- [154] Lee, M., Kim, K. & Cho, Y. A study on the relationship between technology diffusion and new product diffusion. *Technological Forecasting and Social Change* **77**, 796–802 (2010).

- [155] Cheng, A.-C. Exploring the relationship between technology diffusion and new material diffusion: The example of advanced ceramic powders. *Technovation* **32**, 163–167 (2012).
- [156] Argote, L. & Ingram, P. Knowledge transfer: A basis for competitive advantage in firms. *Organizational Behavior and Human Decision Processes* **82**, 150–169 (2000).
- [157] Higham, K. W., Governale, M., Jaffe, A. B. & Zülicke, U. Fame and obsolescence: Disentangling growth and aging dynamics of patent citations. *Physical Review E* **95**, 042309 (2017).
- [158] Li, X., Chen, H., Huang, Z. & Roco, M. C. Patent citation network in nanotechnology (1976–2004). *Journal of Nanoparticle Research* **9**, 337–352 (2007).
- [159] Érdi, P. *et al.* Prediction of emerging technologies based on analysis of the US patent citation network. *Scientometrics* **95**, 225–242 (2013).
- [160] Singh, J. Collaborative networks as determinants of knowledge diffusion patterns. *Management Science* **51**, 756–770 (2005).
- [161] Pugliese, E. *et al.* Unfolding the innovation system for the development of countries: Co-evolution of science, technology and production. *arXiv preprint arXiv:1707.05146* (2017).
- [162] Fritsch, M., Titze, M. & Piontek, M. Knowledge interactions in regional innovation networks: Comparing data sources. Working Paper No. 2018-003, Jena Economic Research Papers (2018).
- [163] Mehta, A., Rysman, M. & Simcoe, T. Identifying the age profile of patent citations: New estimates of knowledge diffusion. *Journal of Applied Econometrics* **25**, 1179–1204 (2010).
- [164] Popp, D., Juhl, T. & Johnson, D. K. Time in purgatory: Examining the grant lag for US patent applications. *Topics in Economic Analysis & Policy* **4**, 1–45 (2004).
- [165] Csárdi, G., Strandburg, K. J., Zálányi, L., Tobochnik, J. & Érdi, P. Modeling innovation by a kinetic description of the patent citation system. *Physica A: Statistical Mechanics and its Applications* **374**, 783–793 (2007).
- [166] Ghiglino, C. & Tabasso, N. The dynamics of innovations and citations. *Economics Letters* **131**, 94–97 (2015).

- [167] Balconi, M., Breschi, S. & Lissoni, F. Networks of inventors and the role of academia: An exploration of Italian patent data. *Research Policy* **33**, 127–145 (2004).
- [168] Breschi, S. & Lissoni, F. Mobility of skilled workers and co-invention networks: An anatomy of localized knowledge flows. *Journal of Economic Geography* **9**, 439–468 (2009).
- [169] Doonan, W., Higham, K. W., Governale, M. & Zülicke, U. Community structure in co-inventor networks affects time to first citation for patents. *Applied Network Science* **4**, 17 (2019). URL <https://doi.org/10.1007/s41109-019-0126-3>.
- [170] Wang, M., Yu, G. & Yu, D. Measuring the preferential attachment mechanism in citation networks. *Physica A: Statistical Mechanics and its Applications* **387**, 4692–4698 (2008).
- [171] Wang, M., Yu, G. & Yu, D. Effect of the age of papers on the preferential attachment in citation networks. *Physica A: Statistical Mechanics and its Applications* **388**, 4273–4276 (2009).
- [172] Geng, X. & Wang, Y. Degree correlations in citation networks model with aging. *Europhysics Letters* **88**, 38002 (2009).
- [173] Yang, R. & Song, A. Scaling properties of an evolving network model with aging mechanism. *International Journal of Modern Physics C* **20**, 781–787 (2009).
- [174] Hajra, K. B. & Sen, P. Aging in citation networks. *Physica A: Statistical Mechanics and its Applications* **346**, 44–48 (2005).
- [175] Medo, M., Cimini, G. & Gualdi, S. Temporal effects in the growth of networks. *Physics Review Letters* **107**, 238701 (2011).
- [176] Sørensen, J. B. & Stuart, T. E. Aging, obsolescence, and organizational innovation. *Administrative Science Quarterly* **45**, 81–112 (2000).
- [177] Bergeaud, A., Potiron, Y. & Raimbault, J. Classifying patents based on their semantic content. *PLOS ONE* **12**, e0176310 (2017).
- [178] Yan, B. & Luo, J. Measuring technological distance for patent mapping. *Journal of the Association for Information Science and Technology* **68**, 423–437 (2017).

- [179] Alstott, J., Triulzi, G., Yan, B. & Luo, J. Mapping technology space by normalizing patent networks. *Scientometrics* **110**, 443–479 (2017).
- [180] Gao, Y., Zhu, Z., Kali, R. & Riccaboni, M. Community evolution in patent networks: Technological change and network dynamics. *Applied Network Science* **3**, 26 (2018).
- [181] Murray, F. & Stern, S. Do formal intellectual property rights hinder the free flow of scientific knowledge?: An empirical test of the anti-commons hypothesis. *Journal of Economic Behavior & Organization* **63**, 648–687 (2007).
- [182] Aoki, R. & Spiegel, Y. Pre-grant patent publication and cumulative innovation. *International Journal of Industrial Organization* **27**, 333–345 (2009).
- [183] Henderson, R., Jaffe, A. B. & Trajtenberg, M. Universities as a source of commercial technology: A detailed analysis of university patenting, 1965–1988. *Review of Economics and Statistics* **80**, 119–127 (1998).
- [184] Mowery, D. C., Nelson, R. R., Sampat, B. N. & Ziedonis, A. A. The growth of patenting and licensing by US universities: An assessment of the effects of the Bayh–Dole Act of 1980. *Research Policy* **30**, 99–119 (2001).
- [185] Hall, B. H., Jaffe, A. B. & Trajtenberg, M. The NBER patent citation data file: Lessons, insights and methodological tools. Working Paper 8498, National Bureau of Economic Research (2001).
- [186] Petersen, A. M., Pan, R. K., Pammolli, F. & Fortunato, S. Methods to account for citation inflation in research evaluation. *Research Policy* **48**, 1855–1865 (2019).
- [187] Wang, D., Song, C. & Barabási, A.-L. Quantifying long-term scientific impact. *Science* **342**, 127–132 (2013).
- [188] Milojević, S. Power law distributions in information science: Making the case for logarithmic binning. *Journal of the American Society for Information Science and Technology* **61**, 2417–2425 (2010).
- [189] Dyson, F. A meeting with Enrico Fermi. *Nature* **427**, 297 (2004).
- [190] Mayer, J., Khairy, K. & Howard, J. Drawing an elephant with four complex parameters. *American Journal of Physics* **78**, 648–649 (2010).

- [191] Bianconi, G. & Barabási, A.-L. Competition and multiscaling in evolving networks. *Europhysics Letters*. **54**, 436–442 (2001).
- [192] Simkin, M. V. & Roychowdhury, V. P. A mathematical theory of citing. *Journal of the American Society for Information Science and Technology* **58**, 1661–1673 (2007).
- [193] Radicchi, F., Fortunato, S. & Castellano, C. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences* **105**, 17268–17272 (2008).
- [194] Stringer, M. J., Sales-Pardo, M. & Amaral, L. A. N. Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal. *Journal of the American Society for Information Science and Technology* **61**, 1377–1385 (2010).
- [195] Eom, Y.-H. & Fortunato, S. Characterizing and modeling citation dynamics. *PLOS ONE* **6**, e24926 (2011).
- [196] Waltman, L., van Eck, N. J. & van Raan, A. F. Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology* **63**, 72–77 (2012).
- [197] Sheridan, P. & Onodera, T. A preferential attachment paradox: How preferential attachment combines with growth to produce networks with log-normal in-degree distributions. *Scientific Reports* **8**, 2811 (2018).
- [198] Burton, R. E. & Kebler, R. The half-life of some scientific and technical literatures. *American Documentation* **11**, 18–22 (1960).
- [199] Stinson, E. R. & Lancaster, F. W. Synchronous versus diachronous methods in the measurement of obsolescence by citation studies. *Journal of Information Science* **13**, 65–74 (1987).
- [200] Nakamoto, H. Synchronous and diachronous citation distributions. *Informetrics* *87/88* 157–163 (1988).
- [201] Glänzel, W. Towards a model for diachronous and synchronous citation analyses. *Scientometrics* **60**, 511–522 (2004).
- [202] MacRae Jr, D. Growth and decay curves in scientific citations. *American Sociological Review* **34**, 631–635 (1969).

- [203] Matricciani, E. The probability distribution of the age of references in engineering papers. *IEEE Transactions on Professional Communication* **34**, 7–12 (1991).
- [204] Egghe, L. *et al.* Citation age data and the obsolescence function: Fits and explanations. *Information Processing & Management* **28**, 201–217 (1992).
- [205] Mingers, J. & Burrell, Q. L. Modeling citation behavior in management science journals. *Information Processing & Management* **42**, 1451–1464 (2006).
- [206] Parolo, P. D. B. *et al.* Attention decay in science. *Journal of Informetrics* **9**, 734–745 (2015).
- [207] Martin, T., Ball, B., Karrer, B. & Newman, M. Coauthorship and citation patterns in the Physical Review. *Physical Review E* **88**, 012814 (2013).
- [208] Börner, K., Maru, J. T. & Goldstone, R. L. The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences* **101**, 5266–5273 (2004).
- [209] Sorenson, O., Rivkin, J. W. & Fleming, L. Complexity, networks and knowledge flow. *Research Policy* **35**, 994–1017 (2006).
- [210] Mariani, M. S., Medo, M. & Zhang, Y.-C. Identification of milestone papers through time-balanced network centrality. *Journal of Informetrics* **10**, 1207–1223 (2016).
- [211] Nigel Gilbert, G. Referencing as persuasion. *Social Studies of Science* **7**, 113–122 (1977).
- [212] Leydesdorff, L. Theories of citation? *Scientometrics* **43**, 5–25 (1998).
- [213] Brock, W. A. & Durlauf, S. N. A formal model of theory choice in science. *Economic Theory* **14**, 113–130 (1999).
- [214] Klammer, A. & Van Dalen, H. P. Attention and the art of scientific publishing. *Journal of Economic Methodology* **9**, 289–315 (2002).
- [215] Bramoullé, Y. & Saint-Paul, G. Research cycles. *Journal of Economic Theory* **145**, 1890–1920 (2010).
- [216] Van Dalen, H. P. & Klammer, A. Is science a case of wasteful competition? *Kyklos* **58**, 395–414 (2005).

- [217] Van Raan, A. F. Sleeping beauties in science. *Scientometrics* **59**, 467–472 (2004).
- [218] Van Dalen, H. P. *et al.* Signals in science - On the importance of signaling in gaining attention in science. *Scientometrics* **64**, 209–233 (2005).
- [219] Costas, R., van Leeuwen, T. N. & van Raan, A. F. Is scientific literature subject to a sell-by-date? A general methodology to analyze the durability of scientific documents. *Journal of the American Society for Information Science and Technology* **61**, 329–339 (2010).
- [220] Braun, T., Glänzel, W. & Schubert, A. On sleeping beauties, princes and other tales of citation distributions. *Research Evaluation* **19**, 195–202 (2010).
- [221] Ke, Q., Ferrara, E., Radicchi, F. & Flammini, A. Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences* 201424329 (2015).
- [222] Ye, F. Y. & Bornmann, L. Smart girls versus sleeping beauties in the sciences: The identification of instant and delayed recognition by using the citation angle. *Journal of the Association for Information Science and Technology* **69**, 359–367 (2018).
- [223] Flaig, G. & Stadler, M. Success breeds success. The dynamics of the innovation process. *Empirical Economics* **19**, 55–68 (1994).
- [224] Yule, G. U. *et al.* II. A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, F.R.S. *Philosophical Transactions of the Royal Society of London B* **213**, 21–87 (1925).
- [225] Redner, S. How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B - Condensed Matter and Complex Systems* **4**, 131–134 (1998).
- [226] White, D. R. & Houseman, M. The navigability of strong ties: Small worlds, tie strength, and network topology. *Complexity* **8**, 72–81 (2002).
- [227] Newman, M. E. The structure and function of complex networks. *SIAM Review* **45**, 167–256 (2003).
- [228] Berger, N., Borgs, C., Chayes, J. T. & Saberi, A. On the spread of viruses on the internet. In *Proceedings of the sixteenth annual ACM-SIAM symposium on discrete algorithms*, 301–310 (Society for Industrial and Applied Mathematics, 2005).

- [229] Bassetti, B., Zarei, M., Lagomarsino, M. C. & Bianconi, G. Statistical mechanics of the “Chinese restaurant process”: Lack of self-averaging, anomalous finite-size effects, and condensation. *Physical Review E* **80**, 066118 (2009).
- [230] Blei, D. M., Griffiths, T. L. & Jordan, M. I. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)* **57**, 7 (2010).
- [231] Mitzenmacher, M. A brief history of generative models for power law and lognormal distributions. *Internet mathematics* **1**, 226–251 (2004).
- [232] Price, D. J. d. S. *Little science, big science... and beyond* (Columbia University Press New York, 1986).
- [233] Watts, D. J. The “new” science of networks. *Annual Review of Sociology* **30**, 243–270 (2004).
- [234] Newman, M. E. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences* **98**, 404–409 (2001).
- [235] Gmür, M. Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics* **57**, 27–57 (2003).
- [236] Ciotti, V., Bonaventura, M., Nicosia, V., Panzarasa, P. & Latora, V. Homophily and missing links in citation networks. *EPJ Data Science* **5**, 7 (2016).
- [237] Newman, M. E. Scientific collaboration networks. i. network construction and fundamental results. *Physical review E* **64**, 016131 (2001).
- [238] American Physical Society. APS data sets for research (2017). URL <http://journals.aps.org/datasets>.
- [239] Rijcke, S. d., Wouters, P. F., Rushforth, A. D., Franssen, T. P. & Hammarfelt, B. Evaluation practices and effects of indicator usea literature review. *Research Evaluation* **25**, 161–169 (2016).
- [240] Meystre, P. Review changes. *Physical Review Letters* **111**, 180001 (2013).
- [241] Molenkamp, L. W. Scope and standards of PRB. *Physical Review B* **87**, 170001 (2013).

- [242] Meho, L. I. & Yang, K. Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar. *Journal of the American Society for Information Science and Technology* **58**, 2105–2125 (2007).
- [243] Pan, R. K., Petersen, A. M., Pammolli, F. & Fortunato, S. The memory of science: Inflation, myopia, and the knowledge network. *Journal of Informetrics* **12**, 656–678 (2018).
- [244] van Raan, A. F. Dormitory of physical and engineering sciences: Sleeping beauties may be sleeping innovations. *PLOS ONE* **10**, e0139786 (2015).
- [245] Higham, K., Governale, M., Jaffe, A. & Zülicke, U. Unraveling the dynamics of growth, aging and inflation for citations to scientific articles from specific research fields. *Journal of Informetrics* **11**, 1190 – 1200 (2017).
- [246] Mariani, M. S., Medo, M. & Lafond, F. Early identification of important patents: Design and validation of citation network metrics. *Technological Forecasting and Social Change* (2018).
- [247] Leydesdorff, L., Bornmann, L., Comins, J. A. & Milojević, S. Citations: Indicators of quality? The impact fallacy. *Frontiers in Research Metrics and Analytics* **1**, 1 (2016).
- [248] MacRoberts, M. H. & MacRoberts, B. R. Problems of citation analysis. *Scientometrics* **36**, 435–444 (1996).
- [249] Wallin, J. A. Bibliometric methods: Pitfalls and possibilities. *Basic & Clinical Pharmacology & Toxicology* **97**, 261–275 (2005).
- [250] Barabási, A.-L. Network science: Luck or reason. *Nature* **489**, 507–508 (2012).
- [251] Wasserman, M., Zeng, X. H. T. & Amaral, L. A. N. Cross-evaluation of metrics to estimate the significance of creative works. *Proceedings of the National Academy of Sciences* **112**, 1281–1286 (2015).
- [252] Sinatra, R., Wang, D., Deville, P., Song, C. & Barabási, A.-L. Quantifying the evolution of individual scientific impact. *Science* **354** (2016).
- [253] Zeng, A. *et al.* The science of science: From the perspective of complex systems. *Physics Reports* **714-715**, 1–73 (2017).

- [254] Hendy, S. *Get off the grass: Kickstarting New Zealand's innovation economy* (Auckland University Press, 2013).
- [255] Scherer, F. M. & Harhoff, D. Technology policy for a world of skew-distributed outcomes. *Research Policy* **29**, 559–566 (2000).
- [256] Silverberg, G. & Verspagen, B. The size distribution of innovations revisited: An application of extreme value statistics to citation and value measures of patent significance. *Journal of Econometrics* **139**, 318–339 (2007).
- [257] Sanders, B. S., Rossman, J. & Harris, L. J. The economic impact of patents. *Patent, Trademark & Copyright Journal of Research & Education* **2**, 340 (1958).
- [258] Guellec, D. & de la Potterie, B. v. P. Applications, grants and the value of patent. *Economics Letters* **69**, 109–114 (2000).
- [259] Lanjouw, J. O. & Schankerman, M. Patent quality and research productivity: Measuring innovation with multiple indicators. *The Economic Journal* **114**, 441–465 (2004).
- [260] Pakes, A. Patents as options: Some estimates of the value of holding European patent stocks. *Econometrica* **54**, 755–784 (1986).
- [261] Schankerman, M. & Pakes, A. Estimates of the value of patent rights in European countries during the post-1950 period. *The Economic Journal* **96**, 1052–1076 (1986).
- [262] Lanjouw, J. O., Pakes, A. & Putnam, J. How to count patents and value intellectual property: The uses of patent renewal and application data. *The Journal of Industrial Economics* **46**, 405–432 (1998).
- [263] Reitzig, M. Improving patent valuations for management purposes - validating new indicators by analyzing application rationales. *Research Policy* **33**, 939–957 (2004).
- [264] Van Zeebroeck, N. & Van Pottelsberghe de la Potterie, B. Filing strategies and patent value. *Economics of Innovation and New Technology* **20**, 539–561 (2011).
- [265] Harhoff, D. & Wagner, S. The duration of patent examination at the European Patent Office. *Management Science* **55**, 1969–1984 (2009).

- [266] Austin, D. H. An event-study approach to measuring innovative output: The case of biotechnology. *The American Economic Review* **83**, 253–258 (1993).
- [267] Nicholas, T. Does innovation cause stock market runups? Evidence from the great crash. *The American Economic Review* **98**, 1370–1396 (2008).
- [268] Fischer, T. & Leidinger, J. Testing patent value indicators on directly observed patent valuean empirical analysis of Ocean Tomo patent auctions. *Research Policy* **43**, 519–529 (2014).
- [269] Harhoff, D. & Reitzig, M. Determinants of opposition against EPO patent grants - the case of biotechnology and pharmaceuticals. *International Journal of Industrial Organization* **22**, 443–480 (2004).
- [270] Lanjouw, J. O. & Schankerman, M. Stylized facts of patent litigation: Value, scope and ownership. Working Paper 6297, National Bureau of Economic Research (1997).
- [271] Torrance, A. W. & West, J. D. All patents great and small: A big data network approach to valuation. *Virginia Journal of Law and Technology* **20**, 466 (2016).
- [272] Van Zeebroeck, N. The puzzle of patent value indicators. *Economics of Innovation and New Technology* **20**, 33–62 (2011).
- [273] Thoma, G. Composite value index of patent indicators: Factor analysis combining bibliographic and survey datasets. *World Patent Information* **38**, 19–26 (2014).
- [274] Wang, B. & Hsieh, C.-H. Measuring the value of patents with fuzzy multiple criteria decision making: Insight into the practices of the industrial technology research institute. *Technological Forecasting and Social Change* **92**, 263–275 (2015).
- [275] Lee, C., Kwon, O., Kim, M. & Kwon, D. Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting and Social Change* **127**, 291–303 (2018).
- [276] Gambardella, A., Harhoff, D. & Verspagen, B. The value of European patents. *European Management Review* **5**, 69–84 (2008).
- [277] Petruzzelli, A. M., Rotolo, D. & Albino, V. Determinants of patent citations in biotechnology: An analysis of patent influence across the industrial and organizational boundaries. *Technological Forecasting and Social Change* **91**, 208–221 (2015).

- [278] de Rassenfosse, G., Jaffe, A. B. & Webster, E. Low-quality patents in the eye of the beholder: Evidence from multiple examiners. Working Paper 22244, National Bureau of Economic Research (2016).
- [279] van Zeebroeck, N. & van Pottelsberghe de la Potterie, B. The vulnerability of patent value determinants. *Economics of Innovation and New Technology* **20**, 283–308 (2011).
- [280] Squicciarini, M., Dernis, H. & Criscuolo, C. Measuring patent quality: Indicators of technological and economic value. Working Paper 2013/03, OECD Science, Technology and Industry Working Papers (2013).
- [281] Ferretti, L., Cortelezzi, M., Yang, B., Marmorini, G. & Bianconi, G. Features and heterogeneities in growing network models. *Physical Review E* **85**, 066110 (2012).
- [282] Pham, T., Sheridan, P. & Shimodaira, H. Joint estimation of preferential attachment and node fitness in growing complex networks. *Scientific Reports* **6**, 440 (2016).
- [283] Newman, M. E. J. & Leicht, E. A. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences* **104**, 9564–9569 (2007).
- [284] Ronda-Pupo, G. A. & Pham, T. The evolutions of the rich get richer and the fit get richer phenomena in scholarly networks: The case of the strategic management journal. *Scientometrics* **116**, 363–383 (2018).
- [285] Ergn, G. & Rodgers, G. Growing random networks with fitness. *Physica A* **303**, 261–272 (2002).
- [286] Abrams, D. S., Akcigit, U. & Popadak, J. Patent value and citations: Creative destruction or strategic disruption? Working Paper 19647, National Bureau of Economic Research (2013).
- [287] Scotchmer, S. Standing on the shoulders of giants: Cumulative research and the patent law. *Journal of Economic Perspectives* **5**, 29–41 (1991).
- [288] Czarnitzki, D., Hall, B. H. & Oriani, R. The market valuation of knowledge assets in US and European firms. In *The management of intellectual property*, 111–131 (Edward Elgar Publishing, 2006).

- [289] Sapsalis, E. & van Pottelsberghe de la Potterie, B. The institutional sources of knowledge and the value of academic patents. *Economics of Innovation and New Technology* **16**, 139–157 (2007).
- [290] Jefferson, O. A. *et al.* Mapping the global influence of published research on industry and innovation. *Nature Biotechnology* **36**, 31–39 (2018).
- [291] Benson, C. L. & Magee, C. L. Quantitative determination of technological improvement from patent data. *PLOS ONE* **10**, e0121635 (2015).
- [292] Mukherjee, S., Romero, D. M., Jones, B. & Uzzi, B. The nearly universal link between the age of past knowledge and tomorrow’s breakthroughs in science and technology: The hotspot. *Science Advances* **3**, e1601315 (2017).
- [293] de Rassenfosse, G. & Jaffe, A. B. Are patent fees effective at weeding out low-quality patents? *Journal of Economics and Management Strategy* **27**, 134–148 (2018).
- [294] Lerner, J. The importance of patent scope: An empirical analysis. *RAND Journal of Economics* **25**, 319–333 (1994).
- [295] Cooperative Patent Classification. <https://www.uspto.gov/web/patents/classification/cpc/html/cpc.html> (accessed 12 October 2018).
- [296] Régibeau, P. & Rockett, K. Innovation cycles and learning at the patent office: Does the early patent get the delay? *The Journal of Industrial Economics* **58**, 222–246 (2010).
- [297] Breitzman, A. & Thomas, P. Inventor team size as a predictor of the future citation impact of patents. *Scientometrics* **103**, 631–647 (2015).
- [298] Mulaik, S. A. *Foundations of factor analysis* (CRC Press, 2009).
- [299] Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. & Strahan, E. J. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods* **4**, 272 (1999).
- [300] Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24**, 417 (1933).
- [301] Brown, T. A. *Confirmatory factor analysis for applied research* (Guilford Publications, 2014).

- [302] Kaplan, D. *Structural equation modeling: Foundations and extensions* (Sage Publications, 2008), Second edn.
- [303] Kaiser, H. F. An index of factorial simplicity. *Psychometrika* **39**, 31–36 (1974).
- [304] Ghadge, S., Killingback, T., Sundaram, B. & Tran, D. A. A statistical construction of power-law networks. *International Journal of Parallel, Emergent and Distributed Systems* **25**, 223–235 (2010).
- [305] Gualandi, S. & Toscani, G. Human behavior and lognormal distribution. A kinetic description (2018). preprint arXiv:1809.01365.
- [306] Roff, M. Some properties of the communality in multiple factor theory. *Psychometrika* **1**, 1–6 (1936).
- [307] Guttman, L. Multiple rectilinear prediction and the resolution into components. *Psychometrika* **5**, 75–99 (1940).
- [308] Horn, J. L. A rationale and test for the number of factors in factor analysis. *Psychometrika* **30**, 179–185 (1965).
- [309] Kaiser, H. F. The application of electronic computers to factor analysis. *Educational and Psychological Measurement* **20**, 141–151 (1960).
- [310] Kaiser, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**, 187–200 (1958).
- [311] Thomson, G. The factorial analysis of human ability. *British Journal of Educational Psychology* **9**, 188–195 (1939).
- [312] Thomson, G. *The Factorial Analysis of Human Ability* (University of London Press, London, UK, 1950), Fifth edn.
- [313] Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464 (1978).
- [314] Higham, K. W., Governale, M., Jaffe, A. B. & Zülicke, U. Ex-ante measure of patent quality reveals intrinsic fitness for citation-network growth. *Phys. Rev. E* **99**, 060301 (2019).
- [315] Triulzi, G., Alstott, J. & Magee, C. L. Estimating technology performance improvement rates by mining patent data (2018).

- [316] Basalla, G. *The evolution of technology* (Cambridge University Press, 1988).
- [317] Edwards, S. F. & Anderson, P. W. Theory of spin glasses. *Journal of Physics F: Metal Physics* **5**, 965 (1975).
- [318] Kirkpatrick, S. & Sherrington, D. Solvable model of a spin-glass. *Physical Review Letters* **35**, 1792–1796 (1975).
- [319] Amitrano, C., Peliti, L. & Saber, M. Population dynamics in a spin-glass model of chemical evolution. *Journal of Molecular Evolution* **29**, 513–525 (1989).
- [320] Weinberger, E. D. Local properties of Kauffmans NK model: A tunably rugged energy landscape. *Physical Review A* **44**, 6399 (1991).
- [321] Kauffman, S. & Levin, S. Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology* **128**, 11–45 (1987).
- [322] Kauffman, S. A. & Weinberger, E. D. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology* **141**, 211–245 (1989).
- [323] Weinberger, E. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biological Cybernetics* **63**, 325–336 (1990).
- [324] Kauffman, S. A. *The origins of order: Self-organization and selection in evolution* (Oxford University Press, USA, 1993).
- [325] Kauffman, S., Lobo, J. & Macready, W. G. Optimal search on a technology landscape. *Journal of Economic Behavior & Organization* **43**, 141–166 (2000).
- [326] Fleming, L. Recombinant uncertainty in technological search. *Management Science* **47**, 117–132 (2001).
- [327] Fleming, L. & Sorenson, O. Technology as a complex adaptive system: Evidence from patent data. *Research Policy* **30**, 1019–1039 (2001).
- [328] Poincaré, H. & Newman, J. Mathematical creation. *Scientific Work and Creativity: Advice from the Masters* **1**, 177–183 (1908).
- [329] Usher, A. P. *A history of mechanical inventions: Revised edition* (Courier Corporation, 2013).
- [330] Schumpeter, J. A. *Theory of economic development* (Routledge, 2017).

- [331] Barnett, W. P. The organizational ecology of a technological system. *Administrative Science Quarterly* **35**, 31–60 (1990).
- [332] Podolny, J. M. & Stuart, T. E. A role-based ecology of technological change. *American Journal of Sociology* **100**, 1224–1260 (1995).
- [333] Schot, J. & Geels, F. W. Niches in evolutionary theories of technical change. *Journal of Evolutionary Economics* **17**, 605–622 (2007).
- [334] Akers, N. J. The referencing of prior art documents in european patents and applications. *World Patent Information* **22**, 309–315 (2000).
- [335] Gambardella, A., Harhoff, D. & Verspagen, B. The value of patents. In *EPIP Conference, Munich* (2006).
- [336] Hall, B. H. & Ziedonis, R. H. The patent paradox revisited: An empirical study of patenting in the US semiconductor industry, 1979-1995. *RAND Journal of Economics* **32**, 101–128 (2001).
- [337] Cremers, K. Determinants of patent litigation in Germany. Discussion Paper 04-72, Zentrum für Europäische Wirtschaftsforschung (2004).
- [338] Greenacre, M. & Blasius, J. *Multiple correspondence analysis and related methods* (Chapman and Hall/CRC, 2006).
- [339] Kim, J.-W., Hunt, B. & Ott, E. Evolving networks with multispecies nodes and spread in the number of initial links. *Physical Review E* **66**, 046115 (2002).
- [340] Menczer, F. Evolution of document networks. *Proceedings of the National Academy of Sciences* **101**, 5261–5265 (2004).
- [341] Shalizi, C. R. & Thomas, A. C. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research* **40**, 211–239 (2011).
- [342] McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* **27**, 415–444 (2001).
- [343] Caminati, M. & Stabile, A. The pattern of knowledge flows between technology fields. *Metroeconomica* **61**, 364–397 (2010).
- [344] Clarkson, G. Objective identification of patent thickets: A network analytic approach. *Harvard Business School Doctoral Thesis* (2004).

- [345] Page, L., Brin, S., Motwani, R. & Winograd, T. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab (1999).
- [346] Freeman, L. C. A set of measures of centrality based on betweenness. *Sociometry* 35–41 (1977).
- [347] Okamoto, K., Chen, W. & Li, X.-Y. Ranking of closeness centrality for large-scale social networks. In *International Workshop on Frontiers in Algorithmics*, 186–195 (Springer, 2008).
- [348] Bonacich, P. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology* **2**, 113–120 (1972).
- [349] Ramasco, J. J., Dorogovtsev, S. N. & Pastor-Satorras, R. Self-organization of collaboration networks. *Physical Review E* **70**, 036106 (2004).
- [350] Morgan, A. C., Economou, D. J., Way, S. F. & Clauset, A. Prestige drives epistemic inequality in the diffusion of scientific ideas. *EPJ Data Science* **7**, 40 (2018).
- [351] Gress, B. Properties of the USPTO patent citation network: 1963–2002. *World Patent Information* **32**, 3–21 (2010).
- [352] Magerman, T., Van Looy, B. & Song, X. Exploring the feasibility and accuracy of latent semantic analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics* **82**, 289–306 (2009).
- [353] Liu, B. & Zhang, L. A survey of opinion mining and sentiment analysis. In *Mining text data*, 415–463 (Springer, 2012).
- [354] Kataria, S., Mitra, P. & Bhatia, S. Utilizing context in generative Bayesian models for linked corpus. In *Association for the advancement of artificial intelligence*, vol. 10, 1 (2010).
- [355] Gerken, J. M. & Moehrle, M. G. A new instrument for technology monitoring: Novelty in patents measured by semantic patent analysis. *Scientometrics* **91**, 645–670 (2012).
- [356] Letchford, A., Moat, H. S. & Preis, T. The advantage of short paper titles. *Royal Society Open Science* **2**, 150266 (2015).

- [357] Curran, B., Higham, K., Ortiz, E. & Vasques Filho, D. Look who's talking: Two-mode networks as representations of a topic model of New Zealand parliamentary speeches. *PLOS ONE* **13**, e0199072 (2018).