Technology Recommendations from the US Patent Database

Avon Huxor

EXECUTIVE SUMMARY

The aim of this project is to develop tools that can offer recommendations to firms for technology areas that they might consider for new R&D. The recommendations should be based on the firms existing patent portfolio, and will exploit the massive patent database of the US Patent and Trademark Office (USPTO) to reveal the most likely productive for the firm to move into, based on the behaviour of other inventors/firms.

The literature argues that higher value patents come from recombinative innovation, in which different technologies are combined in new ways. It is also expected that this is true only to a certain degree of radically (the conceptual distance between the technologies being mixed), beyond which the innovators themselves, or the market cannot absorb the ideas fully and the innovation has lower value. These two forces together create an expected inverted U-shape curve. This project thus aimed to: firstly identify the optimal technology distance to get the greatest expected patent value; and secondly, use this distance to recommend specific technologies to a firm.

The project originally used the NBER dataset for 1997-2006. These data have been used for many studies, and included data cleaning of patent assignees.

The results are applied to an example, showing how they can be used to suggest future technological directions for a computer company.

Table of contents

List of figures **v**

List of tables **vii**

1 Introduction **1**
1.1 Project Aims . . . 1
1.2 The Nature of Innovation . . . 1
1.2.1 Recombination . . . 2
1.2.2 Encouraging Radical Innovation . . . 3
1.3 Recommendation, not Prediction . . . 5
1.4 The Patent System . . . 6
1.4.1 Patent Classification . . . 7
1.4.2 Patent Citations . . . 9
1.5 The Inverted U-shape relation . . . 9

2 The Dataset(s) **15**
2.1 Introduction . . . 15
2.1.1 NBER . . . 15
2.1.2 NBER Data Description . . . 17
2.2 The Yan and Luo Dataset . . . 23
2.2.1 Cleaning . . . 23
2.2.2 Subset by epoch . . . 23
2.3 PatentsView . . . 24

3 Analysis and Results **27**
3.1 Coding and Analysis Environment . . . 27
3.2 Yan and Luo IPC data . . . 27
3.2.1 Effect on patent value of combinations . . . 28
3.3 Postscriptum: PatentsView data . . . 32

3.3.1 Results for PatentsView data 32
3.3.2 Co-occurence measures 39
3.3.3 An Example 48

4 Conclusions and Discussion **51**
4.1 Evaluation and Discussion 51
4.1.1 Evaluation 51
4.1.2 Discussion 53
4.2 Limitations of Study 54
4.2.1 Provenance of the IPC classifications 54
4.2.2 Issues of the patent system overall 54
4.3 Future Work 56
4.3.1 Heat diffusion 56
4.3.2 Patent text analysis 57
4.3.3 Scientific networks 58

References **61**
4.4 Python Codes 68

List of figures

1.1 Example of structure of IPC classification for H01S 3/00, from Fall et al. (2003). 8
1.2 A schematic diagram of the expected relation between technological distance (diversity, radicalness) and impact of an innovation. From vom Stein et al. (2015). 10
1.3 Relations between various patent metrics and technological improvement. From Benson and Magee (2015). 13

2.1 Number of patents having the IPC sections A–H, in the NBER dataset. . . . 17
2.2 Relationship between rank of assignee firms, and the number of patents they hold. They linear fit is applied to rank>100. Below this value the data clearly depart from Zipf's Law. 18
2.3 Plot of the number of forward citations (logged) against the rank of IPC4 code. 19
2.4 Plot of log of forward citations as a function of originality, for patents in years 1976–2000. 20
2.5 Plot of log(number of forward citations) against log(number of patents) for the ten originality bins. 21
2.6 Slope of power-law fits for originality values. 22
2.7 Plot of number of claims as a function of originality, for patents in years 1976–2000. 23
2.8 Plot of number of forward citations found for the years after patent granting date. 24

3.1 Number of forward for number of sections 28
3.2 Slopes for various numbers of sections used in a patent. 29
3.3 Histogram of for various numbers of IPC3 used in a patent. 30
3.4 Slopes for various numbers of IPC3 allocated to a patent. 31

3.6 Slopes for patents combining 1 to 6 sections used in a patent using PatentsView CPC classification. 33
3.5 Plot of forward citations for number of sections using PatentsView CPC classification 33
3.8 Slopes for patents combining 1 to 9 CPC3s used in a patent using PatentsView CPC classification. 34
3.7 Plot of forward citations for number of CPC3 using PatentsView CPC classification 34
3.9 Plot of forward citations for number of CPC4 codes using PatentsView . . . 35
3.10 Slopes for patents combining 1 to 10 CPC4s 36
3.11 Plot of best-fit slopes as a function of CPC4 combinations. 37
3.12 Slopes for patents combining 1 to 5 CPC4s used in each patent using PatentsView. Section G, and only for firms. 38
3.13 Plot of best-fit slopes as a function of CPC4 combinations. Section G, and only for firms. 38
3.14 Raw patent counts for pairs of CPC sections, including the single section values (e.g. A-A). 40
3.15 Normalised patent counts for pairs of CPC sections. 41
3.16 Forward citation ratio: (number of patents with more than 20 forward citations)/(number below 20) for pairs of CPC sections. 42
3.17 Patent value (forward citation ratio) as a function of technological distance (co-occurance) for CPC technologies in each Section. The origin CPC section is shown in the upper right. The dotted blue horizontal line shows the forward citation rate for patents that only contain the single CPC3 code, and acts as a baseline. 44
3.18 Normalised patent counts for pairs of CPC3 codes in Section G. Note that this plot is not log-scaled in colour, due to the zeros in the data. 46
3.19 Forward citation ratio: (number of patents with more than 20 forward citations)/(number below 20) for pairs of CPC3 codes in Section G. 47
3.20 Similar to Fig. 3.17 but for CPC3 technologies within Section G. 49

4.1 from Alston and Mueller (2015). 53
4.2 Figure 1 from Olsson (2000). 54
4.3 Figure 1 from **?** 56

List of tables

1.1 Distribution of patents for diversified and specialised firms 4

2.1 IPC sections 16
2.2 Top five ranked full IPC codes in the NBER data. 20

3.1 IPC3 codes from the Yan/Luo data, and Google patents, for most cited patents with four different sections. 29
3.2 IPC3 codes from the Yan/Luo data, and Google patents, for most cited patents with four different IPC3 codes. 31
3.3 Descriptions of technologies covered by CPC3 codes in section G 45
3.4 Highly-cited patents with both G06 and G07 classifications. 50

Chapter 1

Introduction

1.1 Project Aims

The primary aim of the project described in this dissertation is to explore the ideas needed to develop computer-based tools that can recommend an area of technology to a firm that is planning to invest in research and development (R&D). The recommendation is based on both the firm's own patent portfolio and the patenting behaviour of other firms, as revealed by a massive public patent database. We are explicitly interested in those firms that want to broaden their R&D activities, and not just deepen their existing technology base. The motivation for doing so is, as we shall see in the literature review below, the growing evidence that a widening the technology base of innovations increases their likely value. Moreover, we assume that firms will require more advice (recommendations) in areas that are new to them, than within their existing R&D expertise.

1.2 The Nature of Innovation

To develop a tool that can aid invention, it is important to understand something of the nature of the innovative process. Innovation seems to cover a wide range of activities. Olsson (2005) suggests that there are three kinds of innovation: incremental innovation, radical innovation, and discovery. Incremental innovations are small non-revolutionary changes. Radical innovations are created by combining existing, disparate, ideas. Discoveries are those innovations that appear so radical that they are initially viewed as an anomaly to the dominant technological paradigm. Keijl et al. (2016) describes using an intermediate level,

between incremental and radical innovation, namely *adjacent innovation*. It is likely that there is actually a continuum of radicalness, rather than distinct regions.

But why the interest in novel or radical innovation: inventions that cut across technology boundaries? Radical inventions are considered important because a consensus of studies indicate that such innovations are the drivers of technological, industrial and social change (Schoenmakers and Duysters, 2010). Shane (2001) found that three attributes influence the likelihood that a a new firm with be created to exploit an innovation: technology-importance, radicalness, and patent scope. We see that the radicalness of an innovation is an important factor in a how an innovation is valued. One does not normally set up up a new firm for a low value, incremental, innovation.

How are we best to try and understand the notion of *radical innovation* and its relationship to incremental innovation?

1.2.1 Recombination

One key idea in innovation is that is can be considered as a search over a technological landscape (Kauffman et al., 2000). In this view, a firm will traverse a conceptual landscape of technological alternatives. There is a *distance* inherent in this model, typically seen as related to the number of changes to an existing design needed to achieve an innovative new design. Kauffman et al. (2000) notes that most firms will do a “local search" near to its current technology portfolio. But he shows that if a firm is already in a poor – or even an average – place in the landscape, then a local search will be sub-optimal; they must look at more distant locations in the landscape. But searching far away comes with costs, thus the aim is to find the *optimal distance* for the firm to focus its efforts.

Another, related, idea underlying much work in innovation studies posits that the the principal source of invention is *recombinant search*: the combination of new and existing technologies (Fleming, 2001; Weitzman, 1998), an idea that goes back at least to the key work of Joseph Schumpeter (Schumpeter, 1939). Recent studies, (Strumsky and Lobo, 2015; Youn et al., 2015) support this idea, and find – from an an analysis of the US patent data – that the bulk of invention comprises the combining of different, but existing, technologies.

An example of innovating by effectively combining two disparate technology areas has been documented by Bruck et al. (2016). They show, using patent citations, how the laser printer evolved as at the convergence of two extant technologies: “sequential printing mechanism" and “static image production".

This result is further underpinned by other analyses of the patent statistics, which found that inventors who acted as interfaces or brokers between different groups of inventors tended to apply for patents that were wider in their technologies (were more “original" in the

terminology used in the literature) Sternitzke et al. (2008). It is suggested that they achieve this originality by be able to access a diverse set of knowledge and skills in their own personal and social networks.

Only rarely does a new technology class appear; the kind of invention that many might think of as truly innovative. Indeed, Youn et al. (2015) find that the rate of appearance of these “new" technologies appears to be slowing down, and that the recombination of the (now many) technology components available, is becoming even more dominant.

Luan et al. (2014) show that having diverse technologies in a patent also increase its impact. However, although significant inventions are more diversified, still keeping a focus on some core technology domains maybe better for creating significant inventions.

But it seems that going too far in radicalness one actually creates problems rather than valuable innovations. Many of the measures of patent “novelty" will also let through the “wacky" patents (Czarnitzki et al., 2011), those wild ideas that have no impact or value. And if the technology mix is small, if the degree of innovativeness is low, the solution rather obvious and the value is also low. There appears to be a sweet spot at a certain “technological distance" (vom Stein et al., 2015), an idea that appears often in the literature (Adamopoulos and Tuzhilin, 2014; Ardito et al., 2016; Katila and Ahuja, 2002; Mastrogiorgio and Gilsing, 2016).

It should be noted that there are dissenting voices. For example, Kasmire et al. (2012) believe that radical innovations are not a great leap, but a form of incremental change that causes non-linear behaviour in the system. This is because we have been confusing a radical effect from a patent, with a radical development in a design.

1.2.2 Encouraging Radical Innovation

One major theme in innovation studies is how to help organisations move into new niches that are outside their “comfort zone". Firms typically use local search in R&D, in which existing approaches and technologies used by them are furthered (Stuart and Podolny, 1996). This caution can be detrimental. Breschi et al. (2003) show (Table 1.1) that “Diversified Innovators" (having patents in more than one of their categories) have a much greater share of patents compared to the “Specialized Innovators" (staying in one category) .

Ahuja and Lampert (2001) list three “pathologies" that inhibit the invention of radical inventions:

- Favouring the familiar: technologies already known to the firm.
- Favouring the mature: technologies that have been in the industry for a while and are well understood.

Table 1.1 Distribution of patents for diversified and specialised firms, from Breschi et al. (2003).

	Diversified Innovators	Specialised Innovators	Total
Share of firms (%)	30.2	69.8	100
Share of patents (%)	89.5	10.5	100

- Favouring propinquity: solutions that lie close to existing ones.

They conclude, from their study of the chemical industry, that firms should experiment with novelty to overcome these pathologies and generate innovative inventions – as commercial benefits are associated with breakthrough inventions.

But how can a firm avoid these pathologies? Evidence suggest that the firms/organisations which have a large input from the sciences avoid these traps. One explanation (Fleming and Sorenson, 2004) for this argues that science changes the way in which inventors search a problem space. Although most invention is a local search, just looking at nearby solutions and changes; science encourages a more “global" search and hence more novel inventions.

Other data support this notion: inventors with science degrees are more likely to generate patents that cross technological boundaries than inventors with an engineering degree; and a doctoral degree is linked with increased combinations of technologies for all groups of inventors (Gruber et al., 2013). Another study found that citations from US patents to US “research papers" grew 300% from 17,000 in the years 1987/88 to 50,000 for 1993/94. Yet in this same period the US patent system grew by only 30% (Narin et al., 1997). This suggests that there is a benefit from using science as an input.

Core + radical technologies

All this is not to suggest that firms should throw out their existing set of technologies and innovate wildly. The available empirical data, using historical patents, show that innovations need one foot in the known – the conventional – as well as another in the radical.

For example, although valued patents often include what Schoenmakers et al. (2010) called “radical" innovation, they also found that radical innovations are actually based on existing knowledge. That is, although radical innovations involve a recombination over different domains, they are usually grounded in a core technology, which may include a “frosting" from another technology area that gives the idea a more innovative tone. Comparable results have been reported elsewhere (e.g. Benson and Magee, 2015; Falk and Train, ming; Keijl et al., 2016; Kim, Daniel et al., 2016): patents with conventional cores and a novel/radical addition do better than average. Indeed, entirely novel inventions fair the worst in measures

of patent impact. And useful unconventionality is more likely to occur with experience, in teams and in large organizations (Della Malva and Riccaboni, 2014). This all suggests we need a managed radically to get truly valuable innovation. As Benson and Magee (2015) puts it, " conventionality does not collide over novelty, but conventionality illuminated by novelty can help an invention's influence".

These considerations find themselves in the behaviour of firms. Breschi et al. (2003) find that when firms move into new technology areas, they do so into ones that are related to the existing activities. It also has an impact on how we might design a tool to aid firms in innovation. It should build on the firms existing technology base, as expressed in its own patent portfolio, but advise of possible novel components to add.

1.3 Recommendation, not Prediction

Due to the cost of undertaking research and development (R&D) in many technology areas, it would be useful to provide innovators with tools that support the decision-making involved in initiating R&D projects, especially if these are in technological fields new to any organisation.

Increasingly, R&D has become a collaborative effort. It is more difficult for a single inventor to cross technical fields, but organisations that undertake patentable research can recruit new staff or re-allocate existing staff if required. Knowing which staff to re-allocate or recruit requires an estimate of which technology areas are "up-and-coming'. Patent technology recommendations can show organisations (which may include companies, universities and quasi-governmental agencies) which areas may be worth their attention in the near future. But we envisage this as a recommendation, not a prediction, as it will have inherent uncertainty, and will be only part of the information that goes into the decision-making process.

Focus on the firm not the inventor

Typically, the inventor of a device or process (we focus on so-called *utility patents*, e.g. devices, methods, compounds and software), will sign over their rights to an *assignee*. These owners can then either use the patent to (exclusively) make and sell a product or service; sell the patent to others, license it to others for a financial reward; or use it as part of a cross-licensing agreement. This assignee is normally a firm – often the employer of the inventor. The other kind of inventor is the lone-inventor, who is typcally the legal assignee.

In this dissertation, we mainly consider the situation from the position of a firm, mainly because, as noted above, a lone inventor is unlikely to able to corral the resources to innovate by adding new technologies as these require additional skills and equipment.

The literature provide further reasons to focus on the firm. For example, it has been found Falk and Train (ming) that having only an individual person as the assignee (representative of an individual inventor) negatively affects the number of forward citations, which are known to be associated with valuable patents (as we shall see below). Also, lone inventors stay close to technology areas that are the same as, or very similar to, the ones they have been successful in before, while firms exhibit a more diverse technology portfolio. Thus the firms are a more appropriate user-group for recommendations.

Shane (2001) has also shown that when an academic-led innovation leads to a start up company, the assignees are almost always the University in which the idea started, and this appears to be typical, so the decision not to use individual person assignees should not bias us against such innovations.

Finally, a focus on the firm simplifies matters. Cymer Inc (who manufacture UV light sources used by integrated circuit manufacturers) are the assignee for 1,542 patents in our (initial) dataset, although these have 219 first named inventors.

Novel/Useful Recommendations

It is notable that a concern with giving recommendations, and with seeking novelty/radicalness in innovation converge in the literature. The value of a recommendation system comes from surprise too. Simple measurements of accuracy will, however, preference algorithms that return the most obvious (Adamopoulos and Tuzhilin, 2014; Lu et al., 2012; Zhou et al., 2010). For example, a film recommender that is so accurate that it returns a list of prequels and sequels to a film you have seen is not so useful. Most viewers will already have considered these. Real value comes from the slightly surprising, a film similar to those they have seen, but one that the viewer might not have considered.

1.4 The Patent System

To understand the project, it is necessary to briefly explain the key features of the patent system, and the data within in. This project uses data from the US Patent and Trademark Office (USPTO). An excellent summary of the US patent system is given in Elliott (2007) and in the USPTO website[1].

The patent system aims to promote technical progress by giving inventors secure rights over their invention for a limited period for having made it public. It thus increases the extent of knowledge in the public domain, and the speed with which it becomes available. The

[1] http://www.uspto.gov

patent system is thought to be a crucial aspect of the economy. The advantages of patents and their legal aims are discussed in Spulber (2015), – although many disagree and consider the patent system bad for innovation e.g. Torrance (2009).

Patents have been used extensively as *indicators* of innovation and economic development, although it has been noted that they too have both advantages and disadvantages in this role (Han and Park, 2006). On one hand, the invention has to pass through an independent assessor (the patent examiner), which maintains quality; the applicant must spend time and money in writing and pursuing the application; and maintaining once granted patents can also be costly. On the other hand, not all inventions are patented. The propensity to patent varies across industrial sectors (Fontana et al., 2013), and the role of patents differs across countries. Nonetheless, Acs et al. (2002) conclude, from empirical evidence, that patents do still provide a fairly good measure of innovative activity. Due to the perceived importance of technological developments to the economy, there are a wealth of studies into using the patent data to try and evaluate patents, and predict technology developments. Indeed, the field has its own journal, *World Patent Information*.

The process of obtaining a patent is, in theory, simple. The inventor will apply for a patent on their idea, which is examined by specialists examiners, who have skills in that technology area. The examiners will grant the application if the idea fulfils the requirements for a valid patent. They will also – as part of the process – add to any citations of previous patents in the application that were included by its authors. Note that these citations can also include scientific journal papers, as well as other preceding patents. The examiners will also include classification codes to represent the “technical content" of the patent. These citations and classification are central to this project and are described below.

The database of the U.S. Patent and Trade Office contains all the data since 1790, retrievable as images of the patent documents; after 1976 the patents are also as full text (Leydesdorff, 2008). Hence many studies fully begin after 1976.

1.4.1 Patent Classification

It is through the classification codes given to a patent that we determine technological novelty. There are a number of patent classifications, the main systems of interest are the USPTO system, the International Patent Classification (IPC) system, and the Co-operative Patent Classification (CPC) system.

USPTO system

There is an original patent classification used by the US patent and trade mark office (USPTO). These are three digits (which can include leading zeros)[2] e.g. “184" represents “Lubrication". This system has been used in the vast majority of research that has exploited the USPTO database, being the *native* classification for most of the history of the US patent system.

Recent work had revealed possible weaknesses of the US patent classifications. For example, in one study reported by Barirani et al. (2013), for the time period in their investigation, the USPTO code for nanotechnology (code number 977) was used for about 4000 patents, while a lexical search of the patent texts returned over 50,000 patents.

Moreover, the USPTO system has not been used since 1st June 2015 for all utility patents, as they move to the CPC (see below). For this reason, we do not use it in this project.

IPC

The International Patent Classification[3] (IPC) divides technology into eight main sections. Each subdivision has a symbol consisting of Arabic numerals and letters of the Latin alphabet, the subdivisions representing almost 70,000 technology classes in total. The full IPC is a hierarchy, with a depth of up to eight levels (see Fig. 1.1).

One comparison Harris et al. (2010) of the IPC and USPC classifications and showed the latter to be better for many applications, as it is more technology focussed. But this work only studied chemistry patents, so the result may not extend beyond this domain. However, others (Alstott et al., 2017) found that in their work, the USPC and IPC4 (using the first 4 characters of the full IPC classification) were comparable. One can also combine both UPSC and IPC, which has been suggested gives a better ontology Benson and Magee (2016).

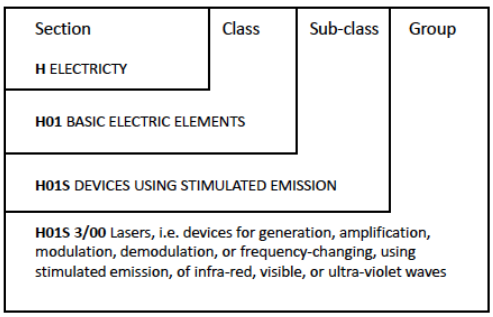

Fig. 1.1 Example of structure of IPC classification for H01S 3/00, from Fall et al. (2003).

We (initially) use the IPC patents.

[2]https://www.uspto.gov/web/patents/classification/selectnumwithtitle.htm
[3]http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf

CPC

There is a new system, the recently agreed Cooperative Patent Classification (CPC) system, which is a bilateral agreement between the USPTO and EPO, which will provide more than 250,000 detailed subgroupings. It is very similar to the IPC, on which it is based. It is increasingly become the system of choice across the globe, not only in the US and Europe.

We originally chose not to use the CPC classification, in case there were teething problems with its ongoing introduction. However, as detailed in later chapters, problems with our original datasets led us to eventually turn the the CPC.

1.4.2 Patent Citations

Another crucial component of the patent data is that of “citations". Patent can cite older patents (backward citation) or be cited itself by other patents (forward citations) in the future (Suh, 2015). If “patent B cites patent A, it implies that patent A represents a piece of previously existing knowledge upon which patent B builds, and over which B cannot have a claim Bronwyn H. Hall (2005)" – so citations have an important legal claim, and for this reason the applicant has a duty to disclose them. The patent examiners can also add new citations if they feel it is appropriate.

Some research as tried to ground the relation of citations with actual process of innovation by interviewing inventors (Jaffe et al., 2000). They found that there is indeed a flow of technical information detectable through the citations, but it is “noisy". About one half of the citations did not reveal of any knowledge flow, while one-quarter show a clear knowledge flow. The large number of the with an apparent lack of knowledge flow appears to be due to the addition of citations by the patent examiner, which although may be legally important, do not reflect the flow of ideas between innovators (Li et al., 2014b; Sorensen and Stuart, 2000).

1.5 The Inverted U-shape relation

In this section, I briefly discuss the key research theme that guide the implementation of this project. The main project aim is to find a means to identify radical combinations of technologies that a firm might apply to its existing technology base: these recommendations need to be novel/radical, and yet rational – giving the best chance of a profitable and impactful innovation outcome.

Our approach to this is to use a feature of the innovation systems that has appeared in the literature: the inverted U-shape relation between the technology “mix" of innovation and its value (Fig. 4.2). It captures the notion (discussed above) that moving away from the

simple, incremental, invention towards something more radical will increase the value of any innovation – but becoming too radical will reduce its value. This reduction in value is due to the absorptive capacity of the firm and/or the market. That is, a very radical innovation may be intrinsically excellent, but if too different from the existing technology mix, then either the development team in the firm will have difficulties working out the details of implementation to “product"; or the market into which the product is planned will not be able to accept it.

The significance of the inverted U-shape to the field of innovation studies can be judged by vast number of and range of studies that find evidence for it (e.g. Ahuja and Lampert, 2001; Cassi and Plunket, 2014; Gilsing et al., 2008; Hsu and Lim, 2014; Nooteboom et al., 2007; Petruzzelli et al., 2015; Wuyts et al., 2005). In some cases, these studies look more widely than the technology distance (mix) inherent in the innovations themselves, and are concerned with the technology mix found within a community of inventors (Cecere and Ozman, 2014), or between firms that collaborate (Fornahl et al., 2011) or undergo mergers or acquisitions (Cloodt et al., 2006).

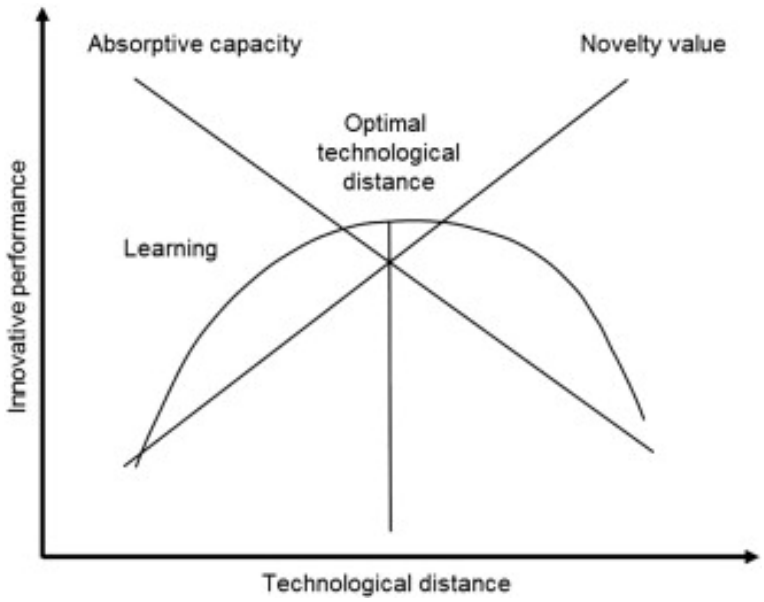

Fig. 1.2 A schematic diagram of the expected relation between technological distance (diversity, radicalness) and impact of an innovation. From vom Stein et al. (2015).

In many cases the detection of the inverted U-shape relationship is statistical, and may be too small to be of effective use. That is, the form of the relationship may not be as striking as shown in the schematic (Fig. 4.2), but very subtle, and when noise effects are taken into account, too small to be use.

Thus our first goals is to try and identify the relationship in our patent data

Deriving a measure for technological distance

To detect the inverted U-shape, we need a measure for the “technology distance", the radicalness of an innovation compared to its predecessors. Moreover, this measure must be derivable from the data we have to hand in our dataset.

Combination degree measures

There are a couple of measures of technologic distance that we use. One is directly related to the notion of recombinant innovation. The measure simply looks at the number of different classes that are in the innovation, as expressed in the patent classifications. This approach looks at the range of patent “technology area" classification codes used in a patent, or in the patents that it cites as prior art (Strumsky and Lobo, 2015) . That is, the number and type of differing classification codes given to any patent is used to estimate innovativeness.

Co-occurance measures

Verhoeven et al. (2016) note that just having how many different IPC codes, for example, are in cited patents does not reflect how difficult a given recombination is – the extent of the “conceptual distance" involved. That is, the distance between section D (paper/textiles) and G (physics) is not equivalent to that between G (physics) and H (electricity). So another approach looks at the specific shared patent classification codes, but treating the different combinations of classes as different (e.g. Breschi et al., 2003; Dolfsma and Leydesdorff, 2011). Breschi et al. (2003) have a measure of distance, based on the co-occurrence of patent classification codes, to get a matrix of co-occurrence of classes within each patent. Jaffe (1986) does similar, but uses the set of patent classes aggregated by firm, rather than in each patent. The notion has also been applied more widely: Ejermo (2005) use the co-occurance approach as seen across different Swedish geographical “regions".

Deriving a measure of patent value

The other key component in fitting the NBER data to the U-shaped curve is a measure for patent value.

Forward citation counts

Lanjouw and Schankerman (2004) show that many factors: forward citations, such as the number of backward citations, the number claims, and the number of countries for which the patents was applied for, all act as proxies for patent quality. And Harhoff et al. (1997) they also adds that whether annual maintenance fees are paid by the assignee is a useful indicator.

But the literature suggests that the most effective measure for patent value is the number of forward citations that a patent receives. Early evidence suggested that the importance of a

patent can be estimated from its citations received Carpenter et al. (1981). They found that a sample of technologically important patents were more than twice as likely to be cited than a random comparison sample. Bronwyn H. Hall (2005) also finds that citations to a patent reflect its value, with each citation increasing its market value by 3%.

In a case study of Computer Tomography, it was found that simple patent counts reflect input R&D (as one might expect), but citation weighted (i.e. weight simple counts by number of citations each patent received) counts of patent better reflect measures of "social value" Trajtenberg (1990). One problem we have is that the NBER data ends at 2006, limiting the period available to collect forward citations. It has been found, however, that the recent citation counts are the most important for patent value, and Kim, Daniel et al. (2016) for example, only look five years ahead of the patent publication date.

Harhoff et al. (1997) provide details to the general notion that the value of a patent can be estimated by the extent to which it is cited by future patents. Indeed, they argue that a single forward citation in the US implies an value of more than $1 million Harhoff et al. (1997). Benson and Magee (2015) discuss some key features of technology development. They have used technology specific measures of performance, across a range of technology types, to investigate how well patent data acts as a tracer of these. They thus have an independent measure of the validity of the patent data. Their results can be seen in Fig. 1.3, which show how various patent metrics correlate with the technology performance measures; panel (A) reveals little correlation between performance and simple patent counts in that domain. This is probably due to the widely varying propensity to patent across technologies Fontana et al. (2013). Panel (B) shows a weak correlation with the number ratio of patents that receive more than 20 forward citations. A good correlation was found (panel C) with the average number of forward citations, received per patent in a domain, within 3 years of publication.

Number of claims

When a patent is filed it must explicitly list the claims that it makes. The number of claims in any patent is also a useful measure of value (e.g Lanjouw and Schankerman, 2004). Nemet and Johnson (2012) indicates that the number of claims are a better measure of its "breadth", which may be the intermediate variable for value. It has been argues that the number of claims only indicates value up to a certain number between 40 and 70), after which it flattens or may even be detrimental[4].

I

[4]http://www.incrementaladvantage.com/articles-objective-analysis/patent-claims-are-determinative-of-patent-value

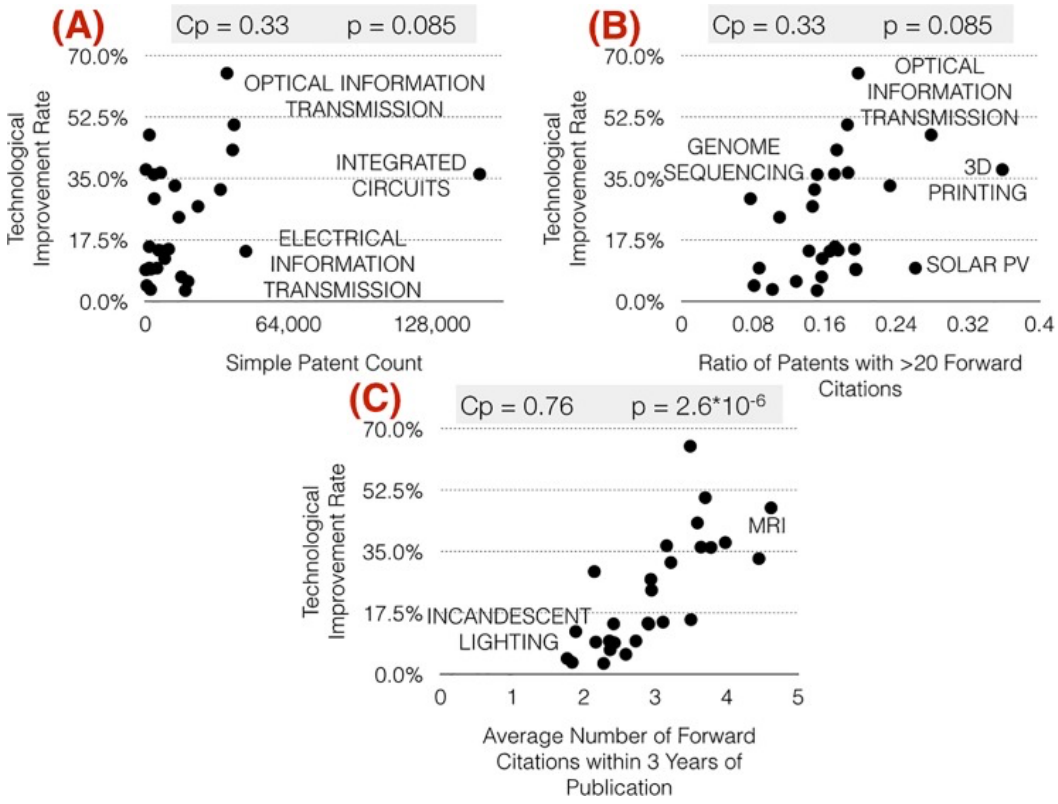

Fig. 1.3 Relations between various patent metrics and technological improvement. From Benson and Magee (2015).

Chapter 2

The Dataset(s)

2.1 Introduction

There are a number of patent databases available that one could use. For the most part, researchers in patent studies adopt either the US patent and trade office (USPTO) data, or that of the European Patent Office (EPO). The US patent data is the most extensive and studied. It stretches back over two centuries, and due to the importance of the US market, many major non-US firms seek patent protection with a US patent.

The complete database of the USPTO contains all the data since 1790, retrievable as images of the patent documents; after 1976 the patents are also as full text – hence most studies fully begin after this date (Leydesdorff, 2008). Although there is a current database (PatentsView,[1]), it was decided[2] to use the NBER database (Hall et al., 2001). It is a standard dataset for patent studies and is still being used up to the present day (e.g. Block et al., 2013; Li et al., 2014a).

2.1.1 NBER

Many of the studies post-1976 were undertaken by the National Bureau of Economic Research (NBER), who created a value-added database as part of their own work, but which they made public. It is documented, in part, by Hall et al. (2001). It is this NBER database that we initially used in this project. The primary driver for this decision is the need for good and clean representation of the assignees (mainly firms) of the patent. The NBER team used text matching to try and get all variants of names for assignee, as well as taking mergers and

[1] http://www.patentsview.org/

[2] As we shall see, this decision to use the well-studied NBER database proved, in hindsight, to be mistaken. After many problems with the NBER and other datasets, the PatentsView data was used in the final analysis.

acquisitions into account (although as noted below, this cleaning was imperfect). Also, early investigations of the PatentsView data indicated some issues, which may be expected for a new system. Specifically, the PatentView system did have its failings in the assignee (firm) naming[3].

Subsets of NBER data

The patent classification system we use, the IPC, has eight broad sections (A–H), described in Table 2.1.

Table 2.1 IPC sections

IPC Section	Description
A	Human Necessities
B	Performing Operations, Transporting
C	Chemistry, Metallurgy
D	Textiles, Paper
E	Fixed Constructions
F	Mechanical Engineering, Lighting, Heating, Weapons
G	Physics
D	Electricity

In addition to looking at the full sample of patents, however, we also chose to focus in on one IPC Section. It has been found that the various technology categories in the US patent data exhibited differed in their overall statistical properties (Gress, 2010). Hence, drawing inferences across technology sectors may not be valid. Specifically, Fontana et al. (2013) to show that propensity to patent differs across technology sectors, with chemical/pharmaceuticals and process engineering having the greatest propensity. One could also focus on technology areas that were (for that period covered by the NBER) new and disruptive, and hence worthy of study, such as mobile phone technologies (Dalum et al., 2005), and liquid crystal technologies (Schadt, 1992).

The propensity for radical patents appears to be dependent on the industrial sector being considered. Schoenmakers and Duysters (2010) found that the industry sector with the most radical patents was “Organic chemistry" followed by “Electric communication technique", while some sectors had no radical patents at all (e.g. Textiles, and Paper). This topic, the varying behaviour of different industries with regard to innovation and patenting is one that arises through this dissertation. We need to be wary of thinking of the “industrial-technological system" as a singular, with common practices.

[3]This issue was reduced by using the NBER assignee data, when we turned to the PatentsView data later.

The section that we focus on, when required, is the G section, as it includes “computing". This section also has the largest sample of patents of all sections (Fig. 2.1), which will help in making the results more robust, especially when the subset is further subdivided into smaller sub-samples for deeper analysis.

2.1.2 NBER Data Description

The NBER dataset was obtained from their Patent Project website[4], and comprised a number of individual data files (and descriptive documents). The data are divided into smaller files to make them more manageable. The full dataset covers the years 1976 – 2006. Below we briefly describe the NBER data and its overall descriptive statistics.

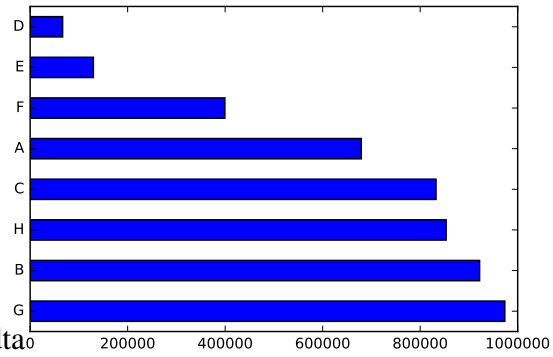

Fig. 2.1 Number of patents having the IPC sections A–H, in the NBER dataset.

The main patent data file is pat76_06_assg.dta with 3,279,509 records. There are 3,210,361 unique patents. Some patents have multiple assignees, each with it own record, giving rise to the larger number of records overall.

There are 223,958 unique assignee numbers. A further 505,557 patents have assignee = “NaN", which on inspection are lone inventors – not firms/organisations. That is, there are twice as many lone inventors/assignees as there are firms (which can include universities, government agencies etc., not just commercial firms).

The number of unique patents compared to assignees seems to imply many patents belong to each assignee. But the reality is quite different, the assignees are very unevenly distributed across the patents. There are four assignees (firms) with over 25,000 patents granted in the period of the full data (1976–2006). These are IBM (45,146 patents), Canon (31,975 patents), Hitachi (28,356 patents) and GEC (26,188 patents). And yet 122,771 assignees (firms) possess only one patent from this period.

A log-log plot of rank against number of patents per assignee (Fig. 2.2) shows that Zipf’s Law basically applies to most the data: having a slope of –1.09, with an R^2 of 0.97. In other words, the proportion of assignees with p patents is inversely proportional to p. For the top 100 assignees, the data shows a tailing-off. This Zipf’s relationship has been found widely in

[4]https://sites.google.com/site/patentdataproject/Home

both patenting and journal paper authorship (Mehta, 2005), and indicates how concentrated successful R&D outcomes are amongst innovators.

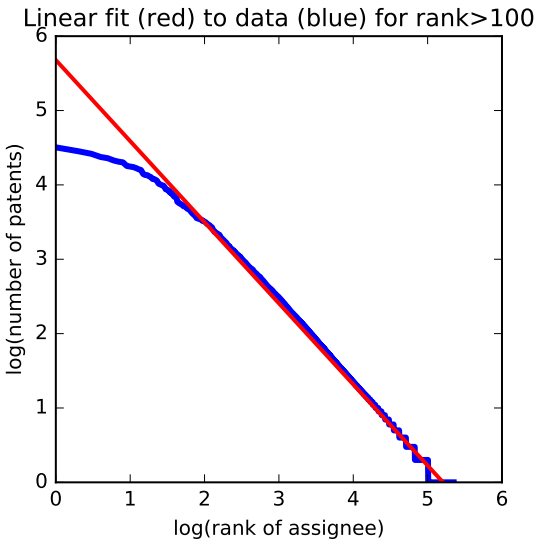

Fig. 2.2 Relationship between rank of assignee firms, and the number of patents they hold. They linear fit is applied to rank>100. Below this value the data clearly depart from Zipf's Law.

Assignees

The file 'assignee.dta' relates the ID number used by NBER for each assignee to the assignee's standard name. These assignee IDs are created by the NBER through extensive term matching to get a unique identifier for the various names that belong to an "assignee". For example, IBM has more than a 100 different spelling, misspellings and abbreviations that were used throughout the original USPTO data. The source of the misspellings is not always clear. In some cases looking to an image of the original patent document shows it to be a typographic error by those writing the patent itself. In other cases it may be due to OCR errors when scanning these documents. As we shall see later, this cleaning by the NBER was not totally successful.

IPC class

The data file of most interest to this dissertation is *pat76_06_ipc.dta* which contains the patent classification data for each patent. This is a subset of the NBER main data file (*pat76_06_assg*), with many columns removed (e.g. day and month of patent publication). There are, however, a few columns in pat76_06_assg that are of interest to us: "allcites" (the number of forward citations to the patent), and "nclaims" (the number of claims in a patent). Recall that Lanjouw

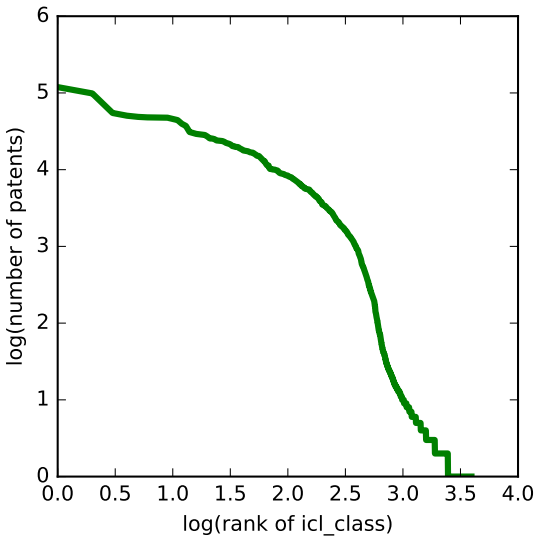

Fig. 2.3 Plot of the number of forward citations (logged) against the rank of IPC4 code.

and Schankerman (2004) suggest that the number of claims can be used as a proxy for patent quality, as well as the more widely used forward citation number.

Fig. 2.3 shows the number of patents in this data as a function of the rank of IPC4 (the first four characters of the IPC full classification, such as “G06F") patent codes used in the patents. The most prevalent IPC4 classes, with more than 50,000 uses in the dataset are: G06F (electric digital data processing: 150054 patents) ; H01L (semiconductor devices; electric solid state devices not otherwise provided for: 119112 patents); A61K (preparations for medical, dental, or toilet purposes: 98515 patents); G01N (investigating or analysing materials by determining their chemical or physical properties: 54784 patents),; and A61B (diagnosis; surgery; identification: 50484 patents).

We can see that these cover the two greatest areas of development in the past few decades: the semiconductor/computer industries and the biomedical industry. The descriptions of the IPC4 codes also illustrate just how broad they are in their coverage.

A list of the top-ranked full IPC codes (Table 2.2) is equally revealing. Of the first ranked five full IPC classes, only one “C12Q 168" (measuring or testing processes involving enzymes or microorganisms involving nucleic acids) does not involve the computer/semiconductor industry. And the latter gives an early pointer as to the growing importance of the genomics industry. Looking at the next five ranked codes are all still in the computing/semiconductor sector.

Originality mesure

The NBER dataset has one value-added feature of use to us. They include a derivation, for most patents, of the *Originality* measure, which is a good proxy for the “radicalness" Hall et al. (2001).

The Originality measure is defined as:

Table 2.2 Top five ranked full IPC codes in the NBER data.

Rank	IPC	Count	Description
1	G06F 1300	9551	Interconnection of, or transfer of information or other signals between, memories, I/O devices or CPUs
2	G06F 1730	8070	Digital computing or data processing equipment or methods, specially adapted for specific applications
3	G11C 700	7469	Arrangements for writing information into, or reading information out from, a digital store
4	C12Q 168	6779	Measuring or testing processes involving enzymes or microorganisms – those involving nucleic acids
5	G06K 900	6263	Methods or arrangements for reading or recognising printed or written characters or for recognising patterns

$$Originality_i = 1 - \sum_{j}^{n_i} s_{ij}^2 \tag{2.1}$$

where s_{ij} denotes the proportion of citations made by patent i that belong to patent class j, out of n_i patent classes (and excluding the focal patent's own class) Only 78% of the patent have an originality measure, as many patents do not cite "out" of their own patent class.

Although Hall et al. (2001) do not discuss the motivation for this approach to estimate the mix of technologies in a patent, we speculate that it is because of their decision to use the US patent classification system in their work. In the data, only one US patent class in attributed to each patent, making it impossible to immediately determine the mix of technologies used by the invention. The NBER data, by contrast does contain multiple IPC classes. The obvious solution would be to look at patents cited by our patent of interest and see how wide the technology mix is for these – assuming that these are in some way an input to the invention.

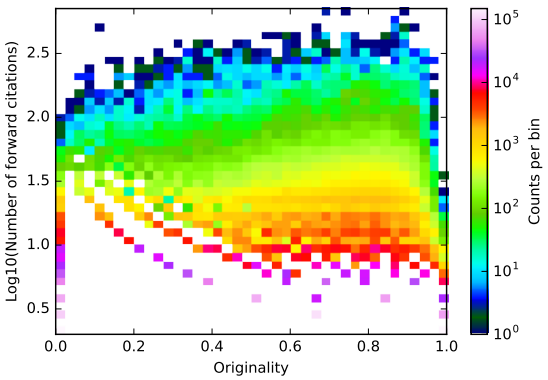

Fig. 2.4 Plot of log of forward citations as a function of originality, for patents in years 1976–2000.

The NBER originality measure does, however, allow us to get obtain an early indication of the impact of originality on patent value. Inspection of Figure 2.4 does seem to support the idea that higher originality gives greater value. However, for any value of originality,

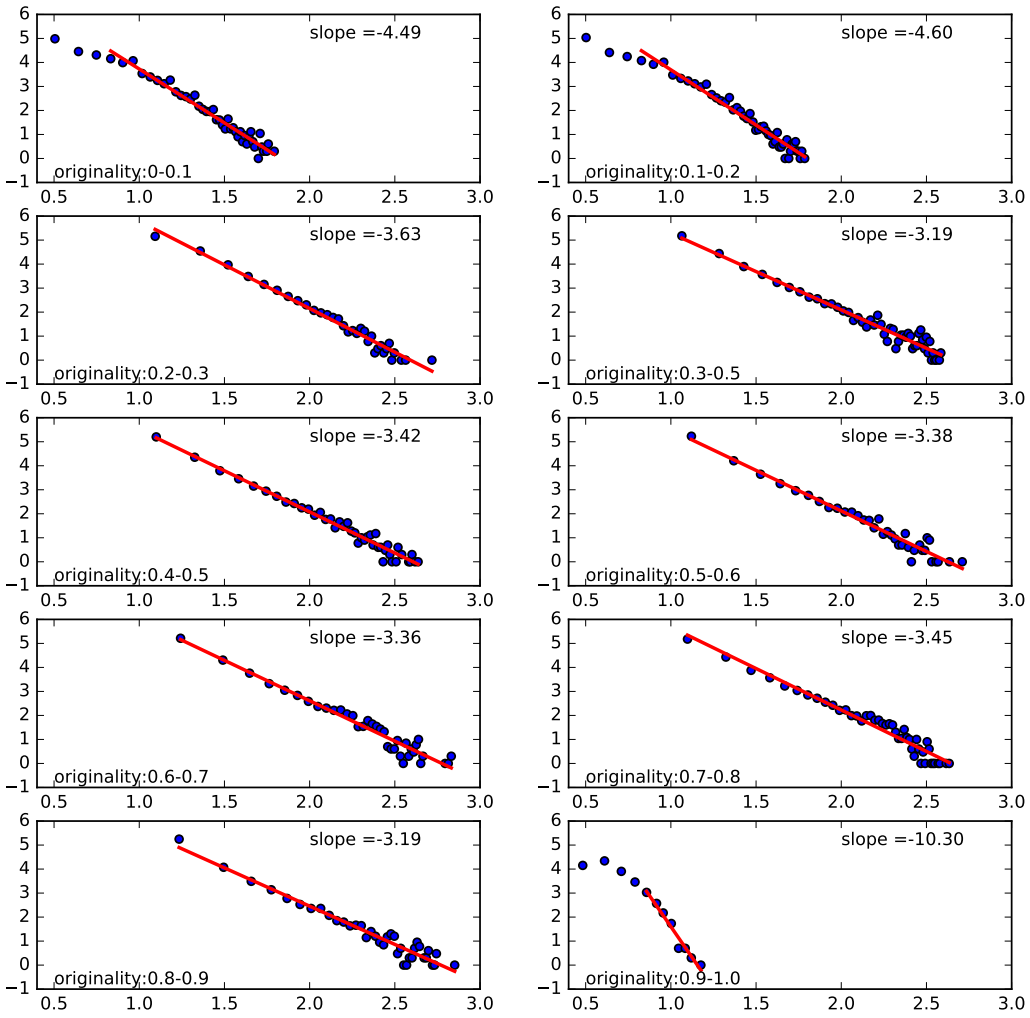

Fig. 2.5 Plot of log(number of forward citations) against log(number of patents) for the ten originality bins.

the distribution of patent forward citations is distributed in a power-law like manner. Most patents, at any originality, have very few forward citations, and very few patents have many forward citations. Hence, it would be inappropriate to apply traditional statistics.

But we can put numbers to this situation. We take bins of 0.1 width in originality and plot the log of the number of patents against the log of the number of forward citations. If more highly-cited patents exist in any bin, as a proportion of the total number of patents, the slope will be shallow. That is, we use the slope of the fit as a measure of the ratio of highly-cited patents in the bin and hence is “value".

When we plot the slopes across the range of originality (Fig. 2.6), however, we do not see a clean inverted U-shape overall . The values have flattish top for a wide range of originality, have somewhat lower values for very low originality, and drop off steeply for very high

values of originality (Fig. 2.6). This indicates that indeed the more *original* patents do have an apparent higher value, until the originality become very high, above about 0.9.

But there is a problem with the interpretation, due to the nature of the equation used to derive originality. At high values of originality, such as when the patents cited are equally distributed amongst all available patent classes, that measure never reaches a value of 1. So the final bin unreliable.

We do have the apparent increase in patent value with originality though. It may simply be the case that we are seeing the first hints that the patent system is self-regulating. That is, maybe very radical inventions never actually make it to the patent system. The cost and effort to secure a patent may deter firms from pursuing such ideas to a full patent.

The detections of the inverted U-shape in the literature are often given as statistical. It may be if the gain in patent value decreases with originality, the best fit is curvilinear, but we only see the approach to the peak, and not beyond it.

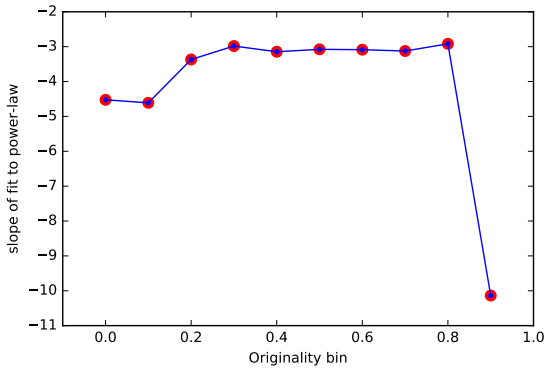

Fig. 2.6 Slope of power-law fits for originality values.

The lack of a clear inverted U-shape may, however, be due to the forward citation number not being a good measure of patent value. If we use the number of claims, which has also been suggested as a measure (se §1.5). However, a plot of originality against claim number (Fig. 2.7), shows less of a signal than for the forward citations. For this reasons, for the rest of this dissertation we use the forward citation count as our chosen measure of patent value.

Problems with the NBER data

Although the chosen dataset for the project, throughout the analysis it became increasingly clear that something was amiss. Occasional checks of particular patents of interest with Google Patents[5] showed that the majority had somewhat different IPC classifications than those given in the NBER data. Initially this was attributed to the fact that patent are reclassified when new standards are released, and also to some OCR errors. But eventually the disparity became stark and problematic, as we rely of these classifications being representative of *ground truth* in some way.

[5]https://patents.google.com

The literature (e.g. Hall et al., 2001) did not describe the origin of the IPC codes (they were after all interested in the US patent classification). Eventually it was found that the IPC codes in the NBER dataset are not native to the patents but generated from the USPTO classification codes via a concordance between the two classification systems [6].

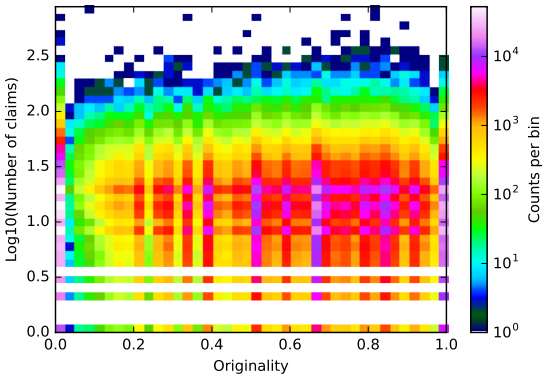

Fig. 2.7 Plot of number of claims as a function of originality, for patents in years 1976–2000.

2.2 The Yan and Luo Dataset

Following the discovery of the issues with the NBER IPC codes, a decision was made to use a new dataset. This is data that is not public, but kindly made available to us by Bowen Yan, and used by him in one recent study, Yan and Luo (2017). In this dataset the classifications only go down to IPC3 (with three characters, e.g. H01). It is not possible to study beyond this "resolution" of patent classification (which is, as we have seen rather broad).

2.2.1 Cleaning

After loading in Bowen's IPC data. First we take a subset to match the period (1997 – 2006) for the NBER dataset, as the Yan and Luo (2017) data covers 1997–2015. This allows us to later join this data with the NBER for forward citations, assignee names etc., as we believe these data to still be reliable.

It was found that a few of the IPC codes did not match the formats in the IPC standard. Regular expressions, based on the IPC standards, were used to reject invalid records. However, from an initial database of 3,694,831 records, only 7837 were rejected. This result initially gave us confidence in the quality of the data, with respect to the IPC classifications.

2.2.2 Subset by epoch

The NBER data goes from 1996–2006, the Yan and Luo (2017) data from a longer period. As we use the forward citations for impact measurements, and these are truncated (that is, a recent patent has not had the time accrue forward citations. Previous authors have used ∼5

[6] Bronywn Hall, priv. comm. 31st August 2017

years look-ahead. However when we plotted the number of forward citations as a function of the years after the patent was granted we find that it is reasonable to have a longer look-ahead period.

We see (Fig. 2.8) that the forward citations accrued increases, on average, up to 15 years after the publication of a patent. Thus, as we use the NBER forward citations data to 2006, we cut-off our sample of patents to 1990, so as to get to probe when the citations numbers are still increasing. This is a longer period than most other studies. For example, Keijl et al. (2016) allows a ten year period, highlighting that this is greater than the typical 5 year period. This 15 year period impacts greatly on the size of our final working sample, and its recency. Interestingly, evidence indicates that a major change in the US patent system seems to have occurred between the years 1990 and 2000 (Strandburg, 2009). Having a cut-off at 1990 for our main sample minimise the impact of this event, although of course it may still effect the number and nature of forward citations that occur in that period of change.

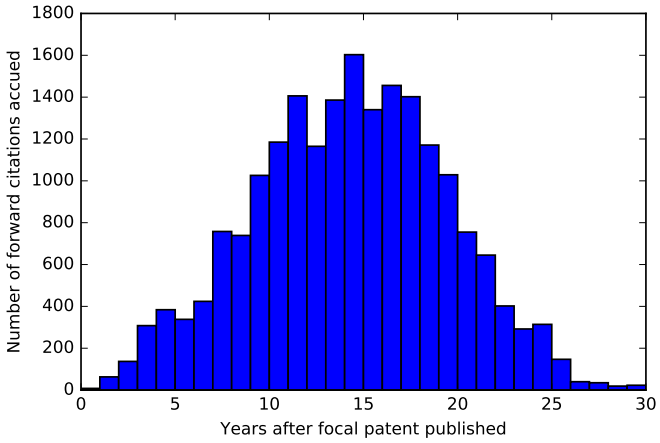

Fig. 2.8 Plot of number of forward citations found for the years after patent granting date.

Visual inspection of a random number of patents further showed that some IPC codes did not match those available in the various version of the classification standard. These are described in the next chapter. One consequence of this finding was the decision to turn to the *third* dataset for the project. This is the recent PatentsView data.

2.3 PatentsView

Although not able to provide assignee data well, we can still exploit some of the more recent and extensive PatentsView data[7].

PatentsView is actually is a patent data analysis and visualization platform intended to increase the value and utility of patent data. Most importantly for our purposes, it is built on a regularly updated database that links inventors, their organizations, locations, and overall

[7]http://www.uspto.gov/about-us/news-updates/new-uspto-tool-allows-exploration-40-years-patent-data (accessed on 30 July 2016)

patenting activity. These individual datasets can also be downloaded from the PatentsView website [8].

[8] http://www.patentsview.org/download/

Chapter 3

Analysis and Results

In this chapter, the experiments undertaken with the data are described along with results. The early experiments, using the NBER data are not included due to the IPC codes used were found to not be from the patents, but derive from the the US patent classes. These results are, however, to be found in the poster (which is available in SAFE). However, the NBER data remains as the base for all the sample, with only the patent classification codes being from from other sources.

With regard to the analysis we try, where possible, to follow the advice of Benner and Waldfogel (2008), who argues for aggregating over years and to course classification granularity to get a good sample size, and also using all technology classes used in a patent classification. Reattempt to keep the period analysed as large as possible, given other constraints, and also begin with the broadest classifications available in the classification systems.

3.1 Coding and Analysis Environment

The project work was undertaken in Python 3, using a Jupyter Notebook. Various libraries were also used, such as Numpy, Pandas, and Scikit-learn. The primary notebooks are submitted in SAFE, and selected codes are included in an appendix of this dissertation.

3.2 Yan and Luo IPC data

In this section we use the Yan and Luo (2017) data for the IPC codes, joined with the NBER data for forward citation counts.

Recall that one goal of the project was to determine if the inverted U-shape curve could be detected in our data, using an appropriate measure of patent value and of technological novelty. We look at two ways of determining technological distance, beginning with a simple count of the number of technological components (as given by the IPC codes) within each patent.

3.2.1 Effect on patent value of combinations

Effect on patent value of combining IPC “sections"

A plot of the number of forward citations as a function of the number of IPC sections in each patent (Fig. 3.1) show that the maximum number of sections found in any patent (of this sample) is four (out of a possible eight).

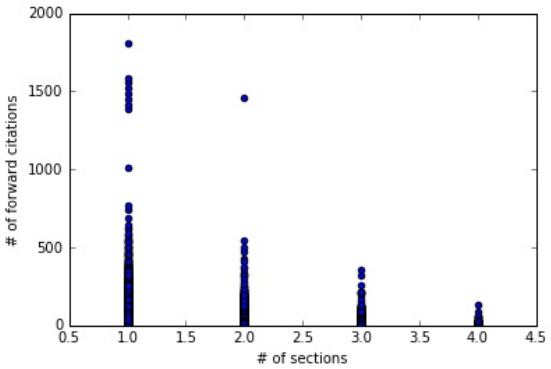

Fig. 3.1 Number of forward for number of sections

But Fig. 3.1 is hard to interpret due to the power law distribution of each bin, with the most cited patents only having one section. And there are many more patents with only one section allocated, than patents with four sections.

So, as before, we derive the slopes for each of the section number bins (Fig. 3.2). We find that the overall value of the patents samples in each bin increases with the number of IPC sections combined: the slopes become shallower. This result reflects that found in the originality/forward-citation plot in the NBER data (Fig. 2.4).

The few patents with four sections combined are naturally of interest, especially those that have garnered many forward citations. Table 3.1 gives the seven most cited (all with more than 50 forward citations) from the four section sample. The table gives the IPC3 codes, rather than just the single section code (which is of course the first letter of the IPC3), as these will better illustrate an issue that arises.

The issue is that Table 3.1 highlights problems with the IPC codes in the Yan and Luo data too (the 3 character codes are required to see that, for example, patent 4394930 has an A61 code in the current Google pants database, but A47 in Yan/Luo. Both would be A if we only looked at the section). For these seven patents, the match between the data and those IPC codes given in Google patents is not good. For patent 4233694 the Yan and Luo data

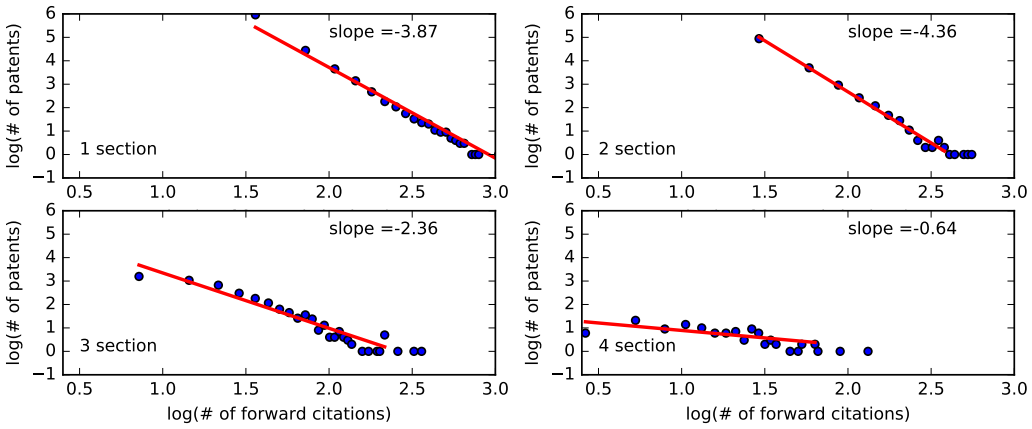

Fig. 3.2 Slopes for various numbers of sections used in a patent.

Table 3.1 IPC3 codes from the Yan/Luo data, and Google patents, for most cited patents with four different sections.

Patent number	IPC3 codes (Yan/Luo data)	IPC3 codes (Google patent)	# forward cites
4394930	'A47' 'B32' 'C08' 'H05'	A61	132
4693556	'A61' 'B32' 'F21' 'G02'	F21, A61, G02	88
4483334	'A61' 'B25' 'E04' 'F16'	A61, F16	64
4081277	'B03' 'D06' 'G03' 'H01'	H01, G03, H04, G03	63
4233694	'A47' 'E04' 'F24' 'H01'	A61	62
4668506	'A61' 'B29' 'C08' 'G03'	A61, G02	51
4532414	'A61' 'B67' 'F24' 'H05'	A61, H05	51

bears no relation to the current Google patent document. Of course, if there are errors (due to typos from OCR, for example), one might expect them to preferentially appear in the extrema of the data, departing from the normal population. And, as noted before, some changes have occurred in the patent classification system over time, but this table does raise doubts.

Effect on impact of combining "IPC3"

As with the NBER data, we can move to a more resolved description of the patents, using IPC3 codes (the three character codes). We find a similar pattern to the IPC section result, although there is now a maximum combination of five IPC3 classes (Fig. 3.3). This is to be expected as patents can have different IPC3 cods, such as G02 and Go3, but would still be in the same section (G).

There are, however, too few patents in the fifth bin to be able to derive a slope.

Fig.3.4 shows a constant shallowing of slope with increasing numbers of IPC3 codes for any given patent. That is, these data suggest that a firm should simply try to increase the number of IPC3 codes its innovations aim for, to get a greater patent value.

The data give only three patents with five IPC3 codes. Again, these are of interest as they may illustrate key features of highly recombinative, radical innovation. These three have patent numbers 4471538, 4325774 and 4792645. We still find an issue with the data. Although the Yan and Luo data give these as having five IPC3 codes, Google patents gives these as having 2, 1, and 2 IPC3 classes respectively.

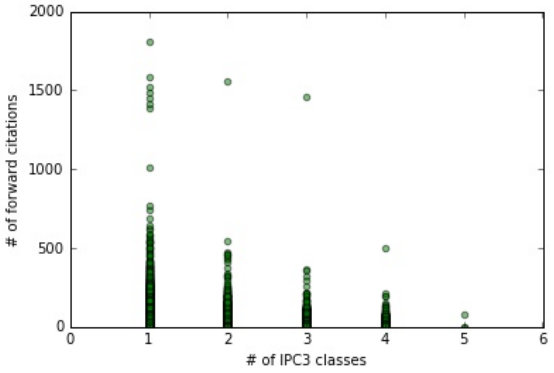

Fig. 3.3 Histogram of for various numbers of IPC3 used in a patent.

There is increasing evidence of significant errors in the IPC data of the Yan and Luo (2017) dataset, although it may still only impact on these extrema (but it is these extrema that interest us, as examples of radical innovation). Looking at those that our data as having four IPC3 codes (Table 3.2), the situation is not very reassuring, with no good correlation between the Yan and Luo data and Google patents. Firstly, the Google patent archive mostly gives fewer than four IPC3 codes to many patents, in others, there are more than four, making the measure of recombination unreliable. Secondly, and maybe more crucial, there is little overlap between between the actual IPC3 codes. The reason for this disparity is not obvious. The errors are too great to be attributed a few OCR typographic errors.

One patent does stand out here, and is worthy of comment. The patent 4863538 has Google patent classes of B05, C23, B29, B23, G05, B22, compared to [’B27’ ’B32’ ’B23’ ’B29’] in our data. That is, the Google patent record has six IPC3s, two more than given in the Yan/Luo data. Their data give this patent as having 199 forward citations; the modern record (up to the present, so there has been time for more citations to accrue) has been cited 682 times. This patent is clearly significant, and and radical in our terms. It is actually a key patent for the new technology of *additive manufacturing* (commonly known as “3-D printing").

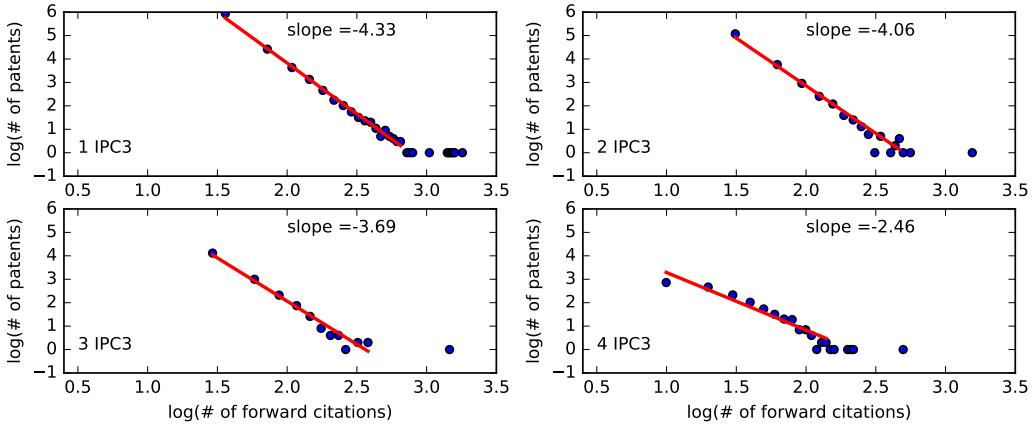

Fig. 3.4 Slopes for various numbers of IPC3 allocated to a patent.

Table 3.2 IPC3 codes from the Yan/Luo data, and Google patents, for most cited patents with four different IPC3 codes.

Patent number	IPC3 codes	IPC3 codes (Google patent)	# forward cites
3962503	'A01' 'A47' 'B32' 'B65'	A01	497
4002886	'B65' 'G06' 'G09' 'H04'	G07, G09, G06	214
4863538	'B27' 'B32' 'B23' 'B29'	B05, C23, B29, B23, G05, B22	199
4595950	'G11' 'G06' 'G07' 'H04'	G11, H04, G06, G07	191
4038519	'A61' 'F16' 'F24' 'H05'	F16, H05, B29, A61, F24	152
4612083	'B44' 'B29' 'C09' 'C23'	H01	140
4162282	'B29' 'B01' 'C09' 'G21'	B01	134
4394930	'A47' 'B32' 'C08' 'H05'	A61	132
4844775	'B44' 'B05' 'C23' 'H01'	H01	124
4752352	'B44' 'B32' 'C23' 'C03'	H01, B23, B21, B29, B22, B32, G05	121
4642160	'B44' 'B29' 'C23' 'C03'	H05	111
4502914	'B44' 'B29' 'C03' 'H01'	C23, H01	106
4601777	'B44' 'C03' 'C23' 'H01'	H01, B41	106
4740410	'B44' 'C03' 'C23' 'H01'	B23, B29, G03, H01	103
4688036	'B62' 'B60' 'G08' 'G06'	E05, G07	102

3.3 Postscriptum: PatentsView data

Following the discovery that the data of Yan and Luo (2017) also has problems, it was decided, during the final days of the project, and after the results presented above, to use the most recent data available, namely the PatentsView data[1]. In particular we only use the file that contains the patent classification data (cpc_current.tsv), using it to replace the patent classification codes of the NBER data,

PatentsView is moving to CPC, replacing both the US patent classification systems, and the IPC (although the CPC is based on the IPC) [2]. Thus, for this section, the results refer to CPC classification codes, rather than IPC, but they are almost identical with some specific changes. Most notably, the CPC classification also includes the recently created section “Y", designed for certain cross-sectional inventions.

To be consistent with the data used above, we use the PatentView data for the period 1976–1990. That is we select the patents in the PatentsView CPCV file for this epoch only, to allow comparison with the results above on the Yan and Luo data.

One reassuring aspect of this dataset is that a list of the unique CPC sections in the data showed there only to be ’A’,’B’,’C’,’D’,’E’,’F’,’G’,’H’,’Y’. The PatentsView data seems to be have been cleaned relatively well (this was unlike the Yan and Luo data, which required some additional cleaning).

3.3.1 Results for PatentsView data

Impact of combining CPC Sections

The same basic analysis described above is now applied to the PatentsView data.

A plot of the forward counts for each mix of CPC sections (Fig. 3.5), now has up to six sections combined. There were 31 patents with six sections, of which four had more than 50 forward citations: patent numbers 4663230, 4344999, 4960643 and 4234907. These were individually inspected in Google patents to see if this data matched.

Reassuringly, for these four patents, the PatentsView data was consistent with the Google patents data. So it is from this data that we consider the most reliable results will come. The results from the Yan and Luo data are given above, however, to allow comparison, and for other researchers to consider the impact that data error may have had on the Yan and Luo (2017) results.

[1] http://www.patentsview.org/download/

[2] http://www.cooperativepatentclassification.org//cpcSchemeAndDefinitions/table.html

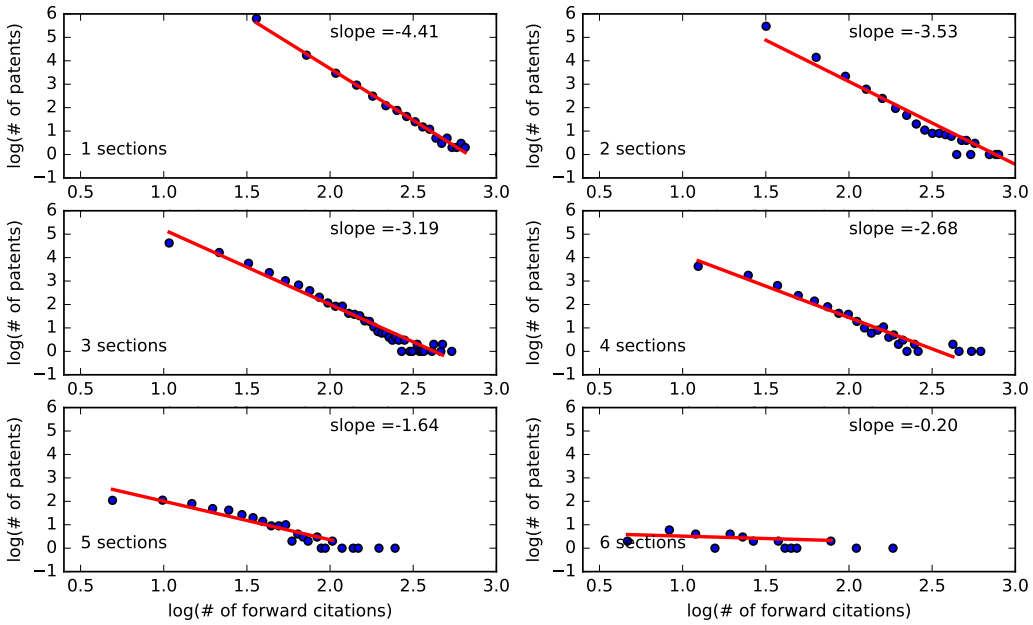

Fig. 3.6 Slopes for patents combining 1 to 6 sections used in a patent using PatentsView CPC classification.

These four highly-recombinative, highly-cited patents are for many technologies that we take for granted today, such as carbon fibre, and other composite synthetic materials.

Inspection of the slopes for the various number of CPC sections (Fig. 3.6) shows – as for the Yan and Luo (2017) data – a constant shallowing of slope with increasing numbers of sections. There is still no sign of the decrease in patent value (as measured by forward citations) with high recombination.

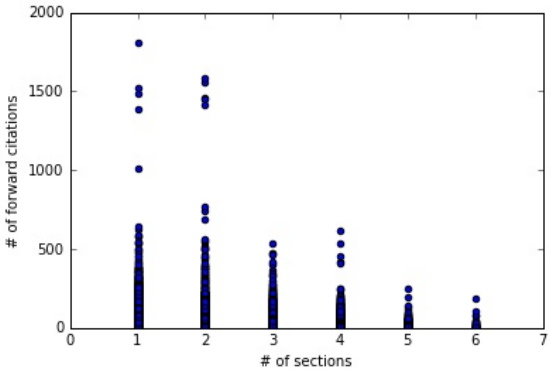

Fig. 3.5 Plot of forward citations for number of sections using PatentsView CPC classification

Impact of combining CPC3

As before, we can now move to more resolved patent classes. The results for CPC3 (equivalent to IPC3 above), are shown in Fig. 3.7, with some patents having as many as 13 unique CPC3 codes. Beyond 8-9 CPC3s in a patent, however, there are too few in the bins to derive a slope (Fig. 3.8).

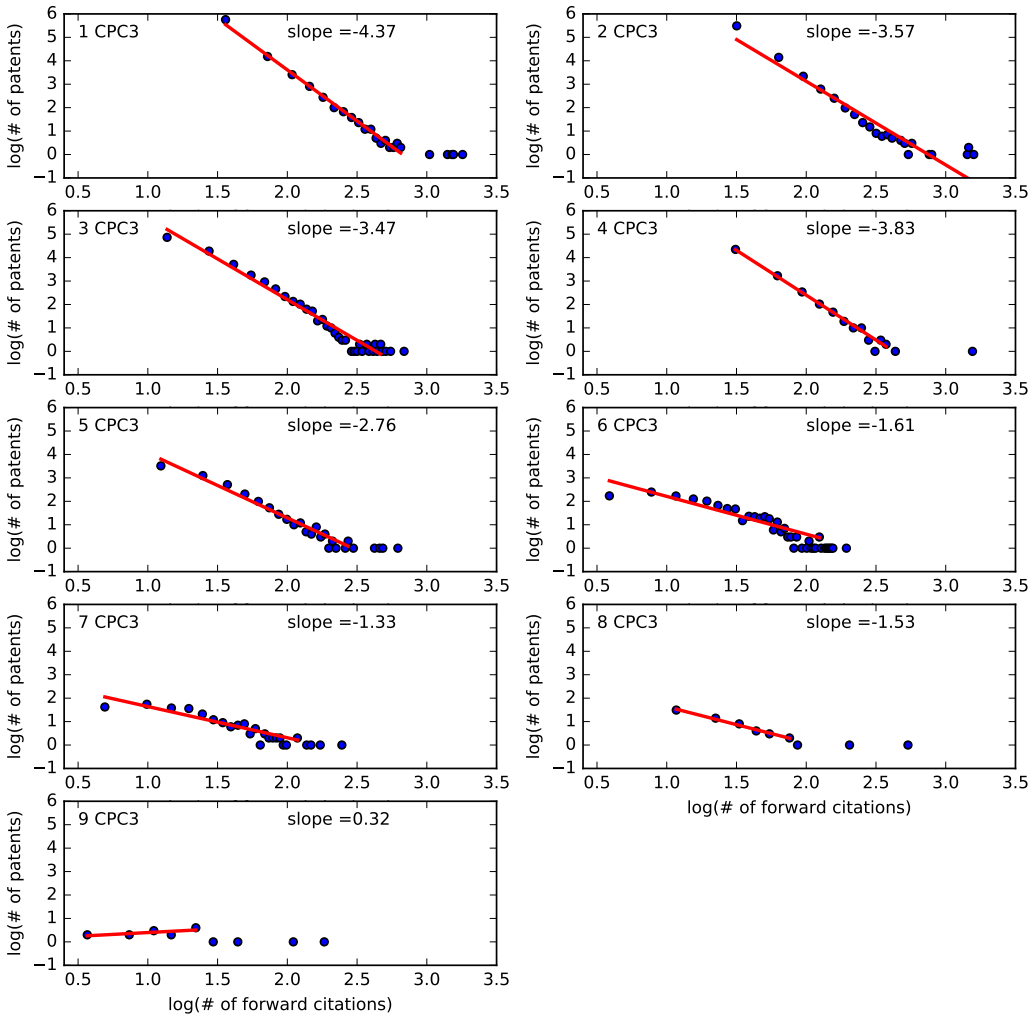

Fig. 3.8 Slopes for patents combining 1 to 9 CPC3s used in a patent using PatentsView CPC classification.

The slopes for CPC3 may tell a different story from the CPC sections. Overall there is an increasing patent value as the number of CPC3 codes combined increases. But there is an apparent lowering of patent value for four CPC3s, the slope being steeper than both two and three CPC3s.

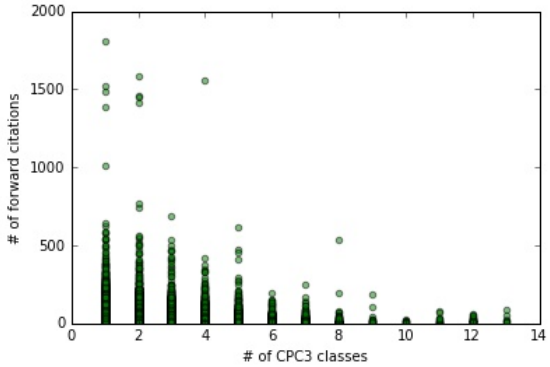

Fig. 3.7 Plot of forward citations for number of CPC3 using PatentsView CPC classification

Impact of combining CPC4

Unlike the Yan and Luo (2017) data, the PatentsView data has the CPC4s, so we can use the equivalent to the IPC4 codes (CPC4), those with four characters (e.g. G06C). The maximum combination of CPC4 is higher than CPC3, at 15 (Fig. 3.9)

The plot of slopes as a function of combinations (Fig. 3.12), returns to a constant shallowing of the slopes with increasing combination, at least up to 10 CPC3s, after which the samples are too small. This result suggest that the feature found in the CPC3 data above, with a steepening at four CPC3s, is possibly due to noisy data or an artefact of the classification system.

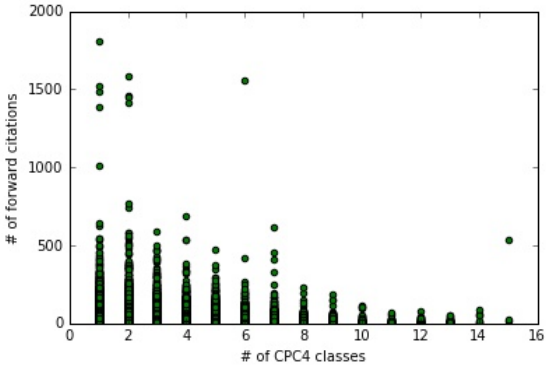

Fig. 3.9 Plot of forward citations for number of CPC4 codes using PatentsView

Investigating the most extreme cases in the CPC4 sample, three patents have 15 CPC4 codes: 4773182, 4209843 and 4415512. There first of these has accrued (in the PatentsView data used) 536 forward citations. yet it is for a decorative flower pot cover. Closer inspection of this patent showed that the citations are – for the most part – from the same company as the inventor, and then from the company that later acted as his trustees. It is an extreme form of self-citation.

The second patent of these has no current assignee, suggesting its value may not be high, although having 23 forward citations. The last of these three, although having 19 forward citations (close to our threshold of 20) also has no assignee as the inventor failed to pay the fee required to maintain the patent. the inventor certainly did not value it too highly. The failure to maintain a patent is actually seen as a good indicator of patent value in itself (Moore, 2005). This information is available for the PatentsView data, but only for post-2005 patents, not those in our chosen epoch (1976-1990). But these, admittedly unrepresentative, patent examples do point to the maintenance of a patent by an assignee as a significant means to clarify the value of a patent.

Although the extreme end of the sample, maybe these three patents illustrate that the behaviour at the highly radical end of innovation (mixing many CPC section/classes) is being driven by the actions of lone inventors. These may have personal motivations that may make them more likely to persevere with, and patent, very radical ideas.

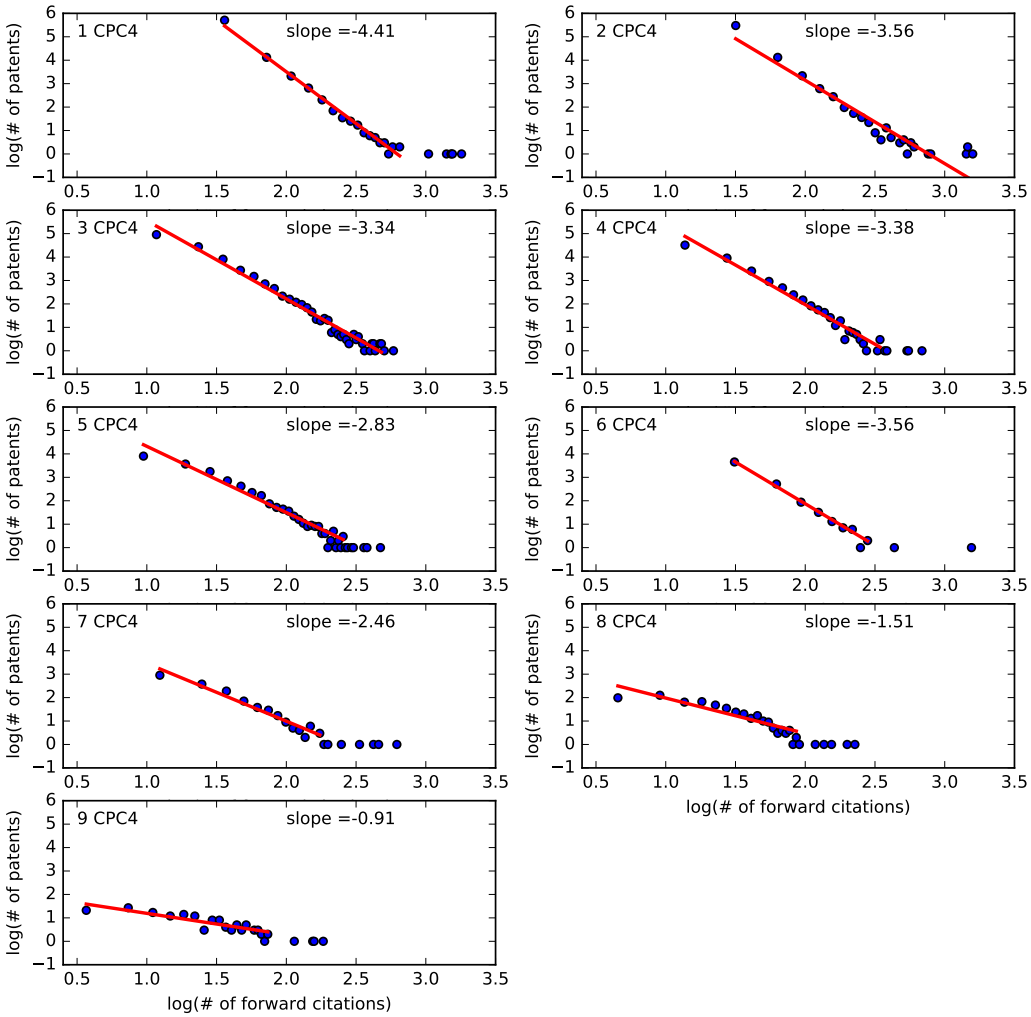

Fig. 3.10 Slopes for patents combining 1 to 10 CPC4s .

The lone inventors have been kept in the sample to-date as it was though likely that they would suffer less from the decision making found in firms, which might avoid spending time and money on either very small incremental innovation, or the very radical, risky, innovations. But maybe these lone-inventors scatter the data, hiding any underlying inverted U-shape relationship.

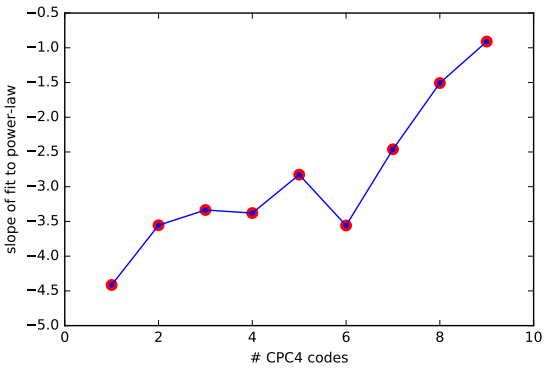

Fig. 3.11 Plot of best-fit slopes as a function of CPC4 combinations.

Hence, for the next experiment we thus select for patents that have a firm as an assignee, i.e. the patent has an assignee code (lone-inventors are indicated by a NaN in the database).

CPC4 for firms only

In addition to restricting the sample to those patent owned by “firm", we further limit the sample to one CPC section, in this case section G (Physics). This is because, as noted by Gress (2010), each industry varies in its propensity to patent, and how its activities have been historically classified by the patent system.

The plots of slopes (Figs. 3.12 and 3.13) maintains the pattern seem earlier, with slope shallowing, interpreted as an increasing ratio of highly-cited patents, as the inventors combine more and more CPC4 sections.

The overall results from the effect of simple combinations of CPC codes does, on the whole, support the idea that the most valuable patents are those that combine many diverse technology types. Of course, we should be cautious about recommending that the firm only need to add random technologies to their existing portfolio. It may well be the case that there is a selection effect in play in the data. As noted above, many firms will self-regulate and *not* patent ideas that look to extreme and complex to get through the patent process, to build, or to market.

Deeper study of the types of combinations will be required to identify any such effects.

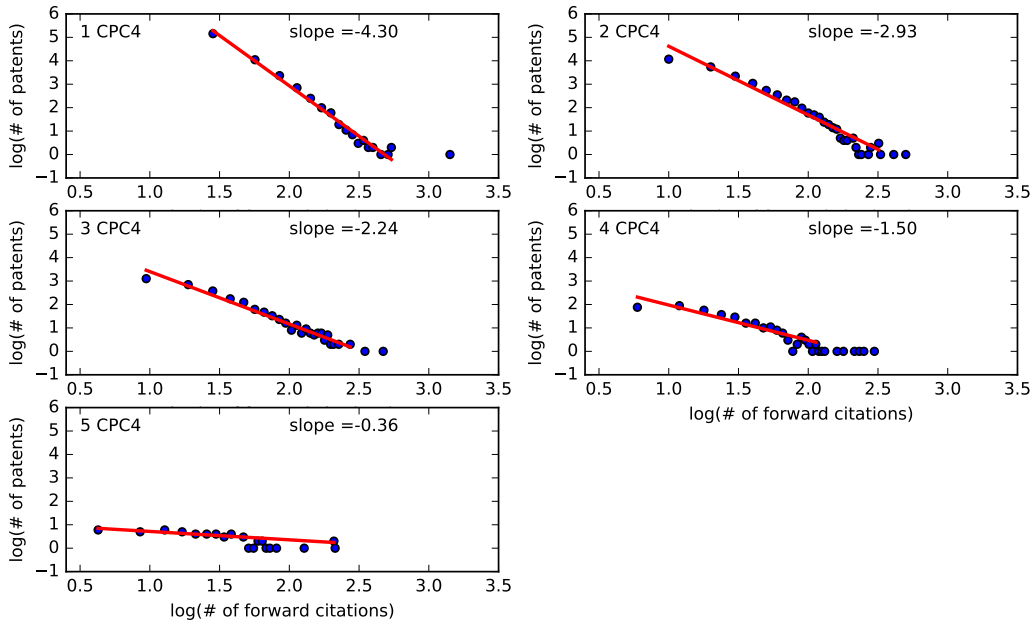

Fig. 3.12 Slopes for patents combining 1 to 5 CPC4s used in each patent using PatentsView. Section G, and only for firms.

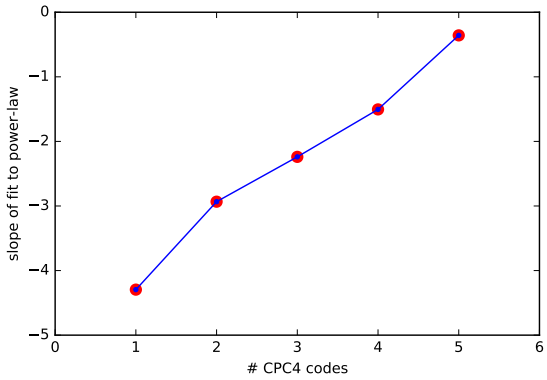

Fig. 3.13 Plot of best-fit slopes as a function of CPC4 combinations. Section G, and only for firms.

3.3.2 Co-occurence measures

The problem with simply counting the number of sections in a patent, as we have above, is that it fails to recognise that the sections differ greatly in their nature. For example, for a patent primarily in section G (Physics) is probably more likely to be related to the H section (Electricity) than to section D (Textiles/Paper). Yet the pairs G/H and G/D are treated equality above, as simply being the combination of two sections.

One approach to improve our “technological distance" measure is to use the likelihood that the section pairs are found together in the dataset and normalise the counts accordingly. This is a *normalised co-occurrence* measure as used by Yan and Luo (2017).

Co-occurence measures for CPC Sections

To achieve this we first derive the number of patents that share a pair of CPC sections. We exclude those those patents that have additional sections to these two sections of interest, so as to create a purer sample. If the pair are accompanied by any one from range of third sections, the diverse nature of these additional sections may impact on the results in unpredictable way.

The resulting values for all patents from 1976–1990 are shown in Fig. 3.14. In this figure the higher diagonal numbers, e.g. for A-A, are the numbers where the patent only has an A section as its classification.

These raw counts are best understood in the context of the relative numbers of patents in each section as a whole (the numbers shown in each cell of Fig. 3.14. As we have seen the G-section has many more patents than the D-section. Thus we normalise these raw numbers using the total number of patents having only the two sections in question. This normalisation gives Fig. 3.15.

This normalised co-occurrence matrix gives an estimate of the “technology distance" between patents using this pair of sections from the single section patents.

One way to imagine how this works is to consider an region of rural medieval England, prior to proper roads, railways etc. Let us assume that people move randomly, they are more likely to find themselves in a village near to their home than to are in more distant one. Taking a database of marriage records, which act as a proxy for people settling elsewhere, and which record the place of birth of the couple involved. The aggregated co-occurrences of the home villages in the marriage records database will reflect the “distances" between the villages – allowing a map to be drawn.

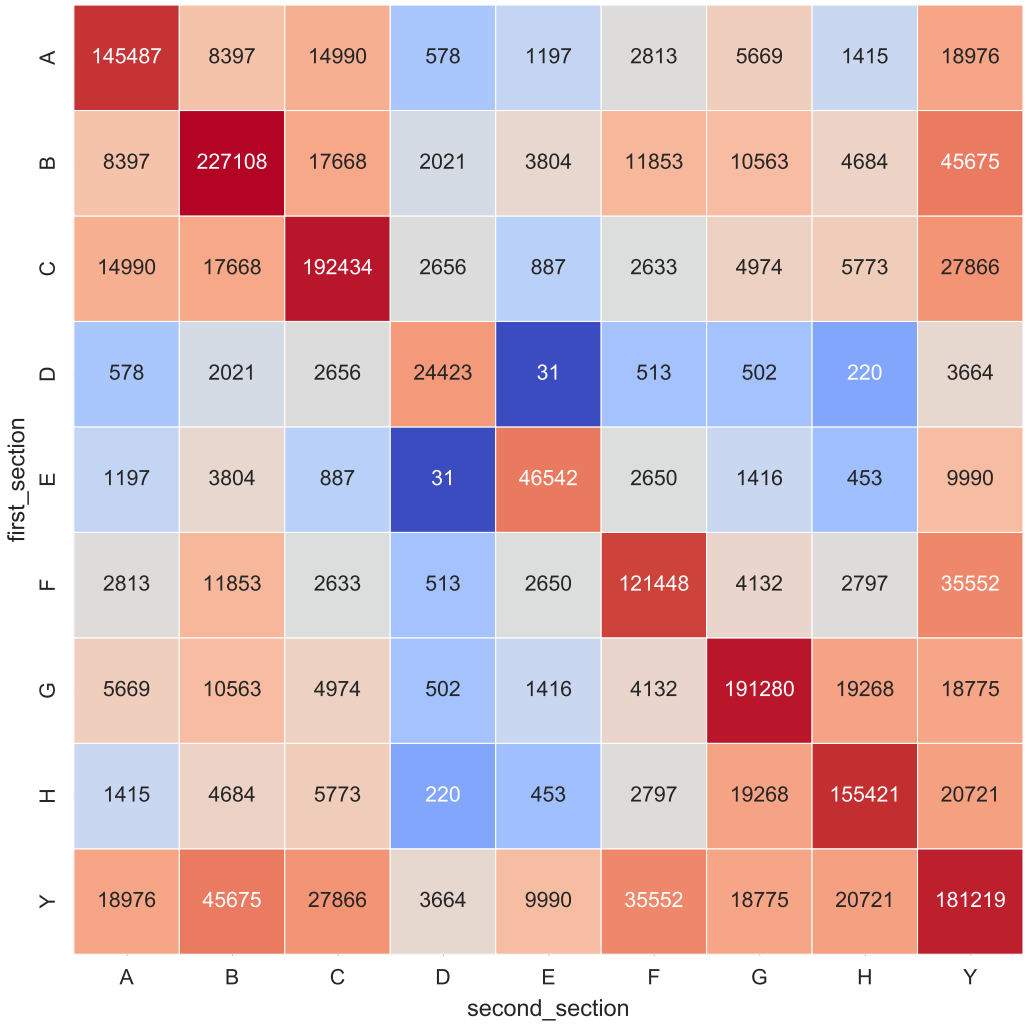

Fig. 3.14 Raw patent counts for pairs of CPC sections, including the single section values (e.g. A-A).

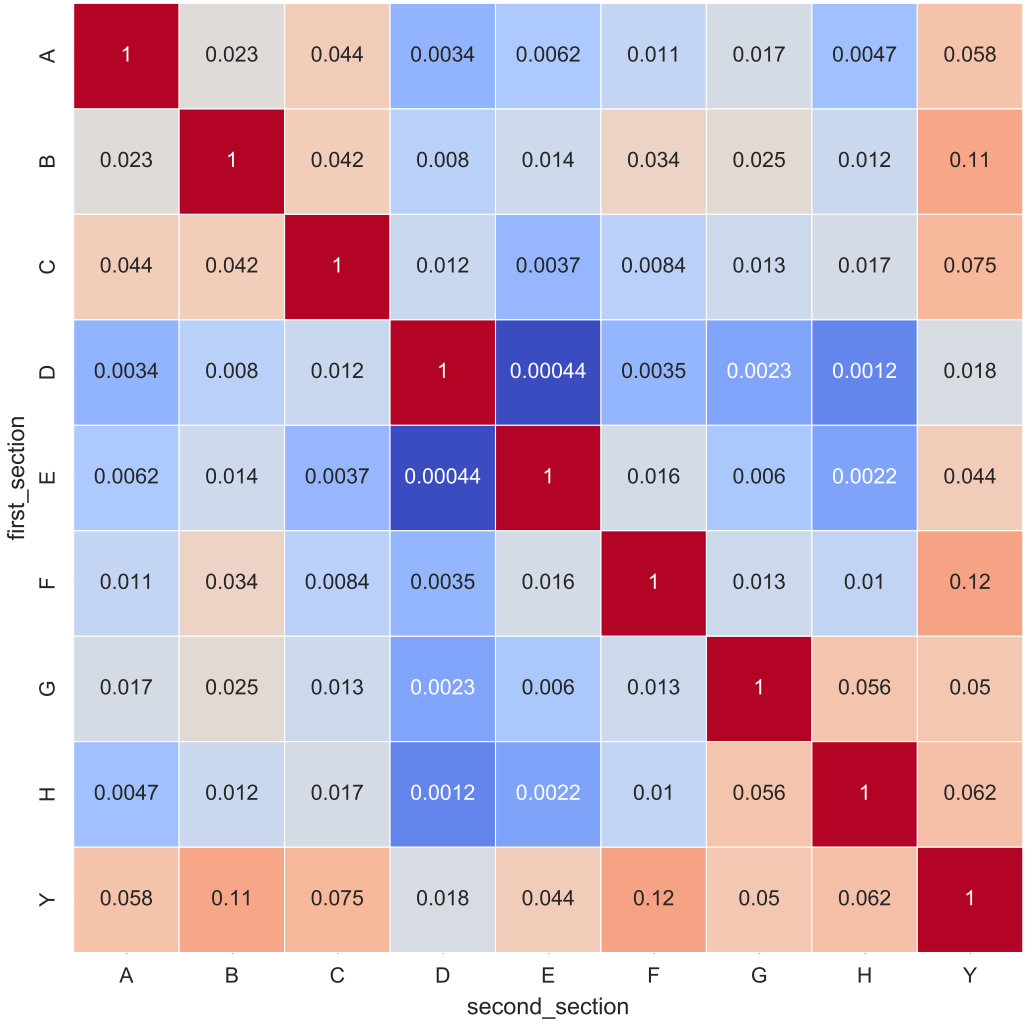

Fig. 3.15 Normalised patent counts for pairs of CPC sections.

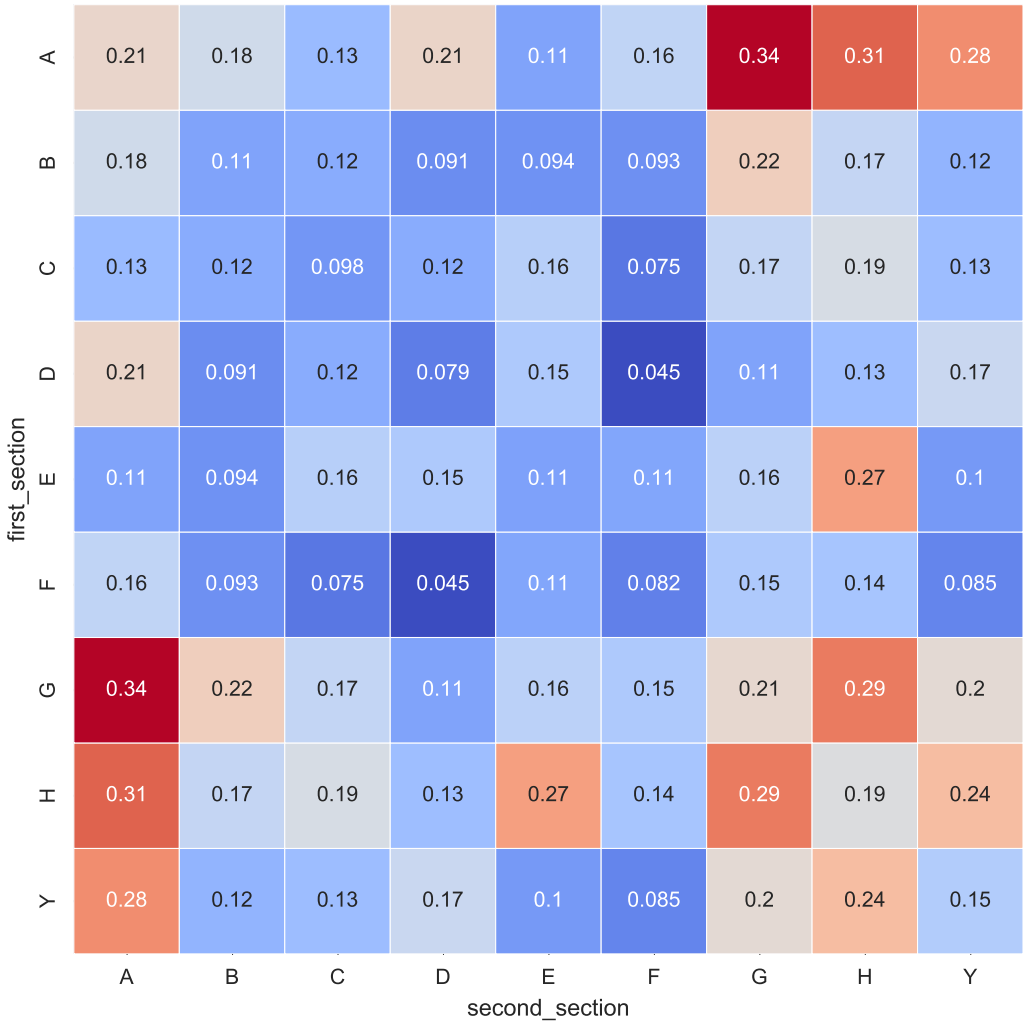

Fig. 3.16 Forward citation ratio: (number of patents with more than 20 forward citations)/(number below 20) for pairs of CPC sections.

So we have our new measure of technological distance. The measure of patent value is, as before, based on forward citations. We need to determine the impact (value) of the patents within each cell defines by the pairs of classification codes.

To do this, for each pair, we take the patents in the matrix cell, and determine the ratio of patents that have more than 20 forward citations to the number of patents with 20 or less forward citations. The higher this ratio, the greater on average are the values of the patents in that cell. This value of 20 is chosen as one found by Schoenmakers and Duysters (2010) to indicate a significant change in the nature of the citation rate.

The plot of forward citation ratios (our measure of patent value) in the matrix of CPC section pairs (Fig. 3.16) shows a number of features. The new CPC section “Y" is unlike

the others in having highly-cited papers across the range when combined with all other sections. This is not unexpected. Section Y is formally described as “General tagging of new technological developments; general tagging of cross- sectional technologies spanning over several sections of the IPC"[3]. It includes older as well as new patents that cross particular boundaries.

Equally notable is that section D (Textiles/Paper) has a very low impact when combined with any other section (Y excepted). The combination of D with E (Fixed constructions; mainly covering civil engineering and building) is the least valuable by our measures. This result concurs with that of Schoenmakers and Duysters (2010).

If we exclude Y, we see two apparent groups of sections whose combinations offer higher value, these are “A, B, C" and then “G, H". There is an overall *geography* of the technological system here, but it is very unresolved. Again, the result for G (Physics) and H (Electricity) is consistent with Schoenmakers and Duysters (2010).

One notable finding is that the forward citation rate for those patents that only have one CPC section in their classification (e.g. A/A) is never the highest. It is, on average, always better to have some combination with another section, although some combinations can be lower than the single CPC section.

For sections D, F, G, H and Y, the most forward citations are all with section A. For sections A and B it the most forward citations are with section G. Finally for C and E the most comes with section H.

This may be a result of the nature of how the sections described the technological domain. Section A is “Human necessities" (typically medical); it is as much about the application of a patent, than the technology inherent in the invention; although these may be correlated – certain technologies may be required to make a patent applicable to medical use.

We now come to the most important results in so far as our initial project aims are concerned.

We take each section, use the co-occurrence ratio as a measure of technological distance to the remaining sections. Ordering the other sections by distance, we can now plot the impact of the patents that have this “distance", with the citation ratio as our proxy for patent impact. It is this plot that we need to compare with our idealised schematic from Chapter 1 (Fig. 4.2), in our search for the inverted U-shaped relationship.

The plots of the patent value (from forward citation ratio with a threshold of 20), against co-occurrence distance (Fig. 3.17), although “noisy" does, gives a hint of the inverted U-shape relation. There is a a peak in patent value lying roughly the distances for all the sections that can be combined with out starting CPC section. But these peaks have very low

[3] http://www.cooperativepatentclassification.org//cpc/scheme/Y/scheme-Y.pdf

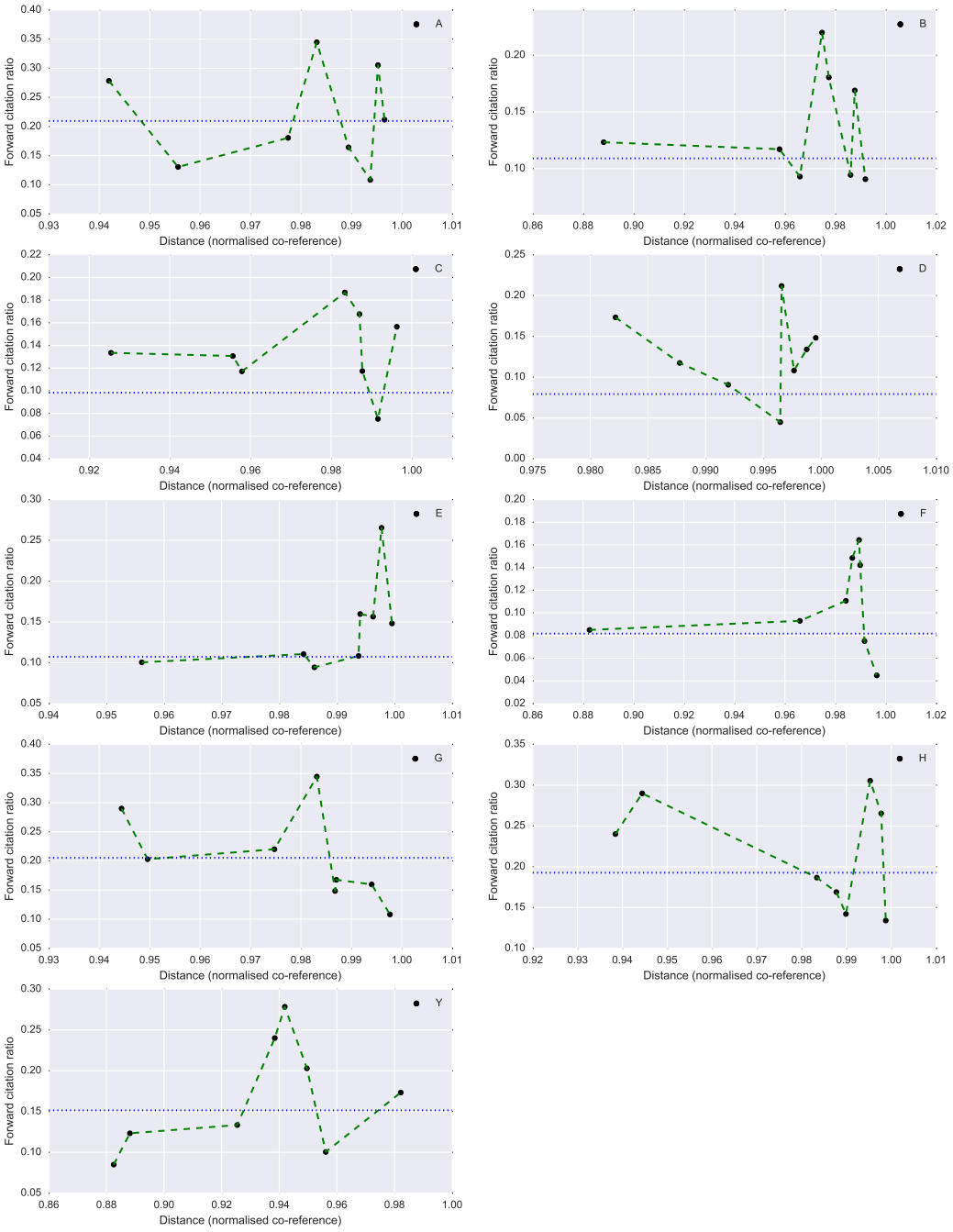

Fig. 3.17 Patent value (forward citation ratio) as a function of technological distance (co-occurance) for CPC technologies in each Section. The origin CPC section is shown in the upper right. The dotted blue horizontal line shows the forward citation rate for patents that only contain the single CPC3 code, and acts as a baseline.

Table 3.3 Descriptions of technologies covered by CPC3 codes in section G

CPC3 code	Description
G01	Measuring, testing
G02	Optics
G03	Photography, cinematography, holography, analogous techniques using waves other than optical
G04	Horology
G05	Controlling, regulating
G06	Computing, calculating, counting
G07	Checking-devices
G08	Signalling
G09	Educating, cryptography, display, advertising, seals
G10	Musical instruments, acoustics
G11	Information storage
G12	Instrument details
G21	Nucleonics

patent value often have data-points either side. We certainly do not see the clear topology of the idealised schematic, but the results can still be used to generate recommendations.

Co-occurence measures for CPC3 codes in Section G (Physics)

Due to the particular properties of various industrial sectors (as noted in Chapter 1), an obvious set is to focus in on our G section, using the more resolved CPC3 codes. The hope is that the firms working in the G section share working practices, and approaches to patenting, allowing any features in the plots to be “cleaner".

So, in this subsection we apply the same analysis as above to patents in that have at least one G-section CPC code, and use the CPC3 codes.

The forward citation ratio reveals certain features of the G section. The least combined CPC3 code is G21 (Nuclear Physics) suggesting that firms in this business do not look far outside their own field for innovations. Given the safety and security issues surrounding this industry, such caution is understandable.

The code G12 (Instrument details) is almost as isolated. This code is currently only used for instrument details ’not otherwise provided for’. It is a box for patents that do fit in well elsewhere in the CPC, and is only sparsely populated.

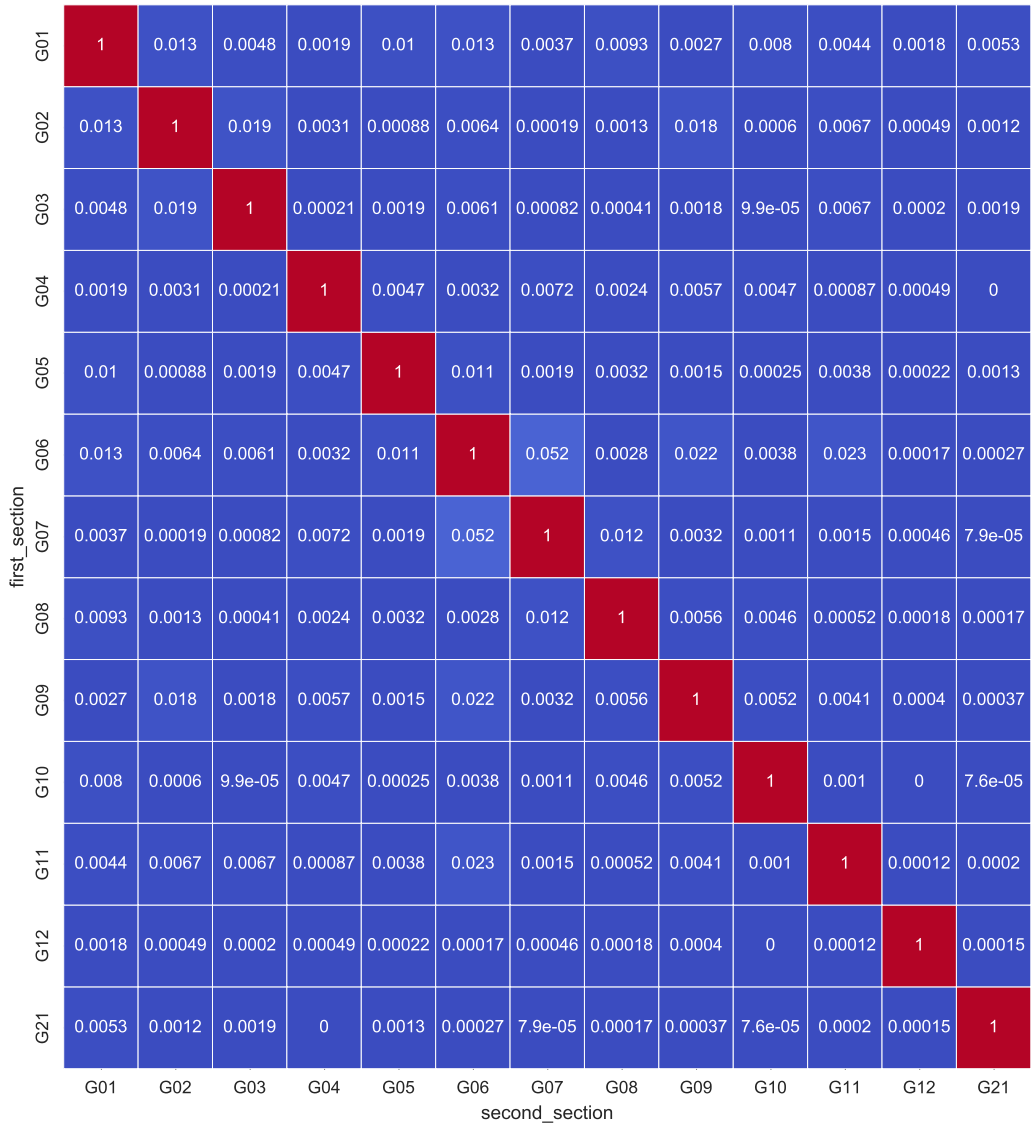

Fig. 3.18 Normalised patent counts for pairs of CPC3 codes in Section G. Note that this plot is not log-scaled in colour, due to the zeros in the data.

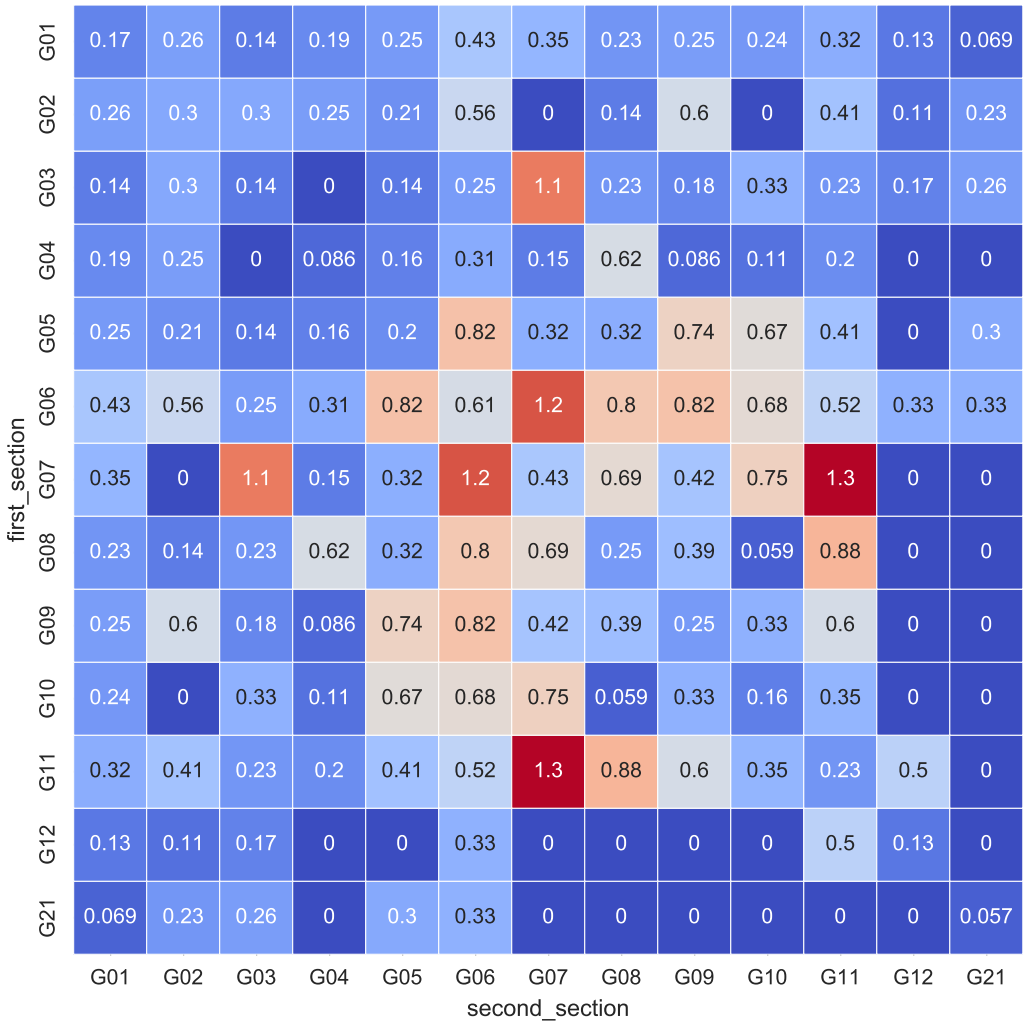

Fig. 3.19 Forward citation ratio: (number of patents with more than 20 forward citations)/(number below 20) for pairs of CPC3 codes in Section G.

3.3.3 An Example

Given the results above, we can start to see how they might be used to recommend technology areas for a firm.

Let us imagine an imaginary firm that has specialised in computing (G06), and is interested in recommendations of technology areas that it might move into.

If we take the CPC3 code “G06" as an example – the three-character code for “computing", we can see how the patent values (measured by forward citation) vary as a function of the “technology proximity" (the inverse of distance – measured by co-occurrence). This plot (Fig. 3.20) is, as in the example above, derived from the values in Fig 3.18 and Fig. 3.19 above.

We see the most proximate CPC3 (Distance 0.05) has the greatest value. This CPC3 is “G07", the code for “Checking Devices". So our initial recommendation to our firm would be to look at this type of technology. But this is a very broad description. Specific suggestions and possible directions of travel can be found by looking at specific high-value (from forward citations) patents that combine G06 and G07.

If we look at a sample of highly cited patents that have both G06 and GO7 classifications, we find the following (Table 3.4). It is notable that these patents cover two types of applications: those that use computing devices to track objects (such as people or animals); and those that involve a computing device with intermediating in markets (be it goods or in services, especially gambling).

Recall that these “patents of interest" to our imaginary computing firm are (due to the nature of our dataset) from patents between 1976 and 1990, using forward citations up to the year 2006 (due to the need to have meaningful time for forward citations to accrue).

In one sense these recommendations would have been eminently sensible, given the time period that the data relates to. The period of the patents used predates – and foresees – many aspects of the modern economy, with online shopping, trading and gambling. All these combine computing devices with the particular need to track/check some (either material or virtual) objects.

The challenge with the technique used above is that the need to have a reasonable period of forward citations on which to judge the value of a patent means that it will always be a few years out-of-date. It may be worthwhile using the PatentsView data alone (which goes up to 2016), and only going back some 5 years (to the year 2000 say). The shorter period of forward citations may have a serious impact on the patent value measurements, but if not, then it would be useful tool to identify technologies to explore now.

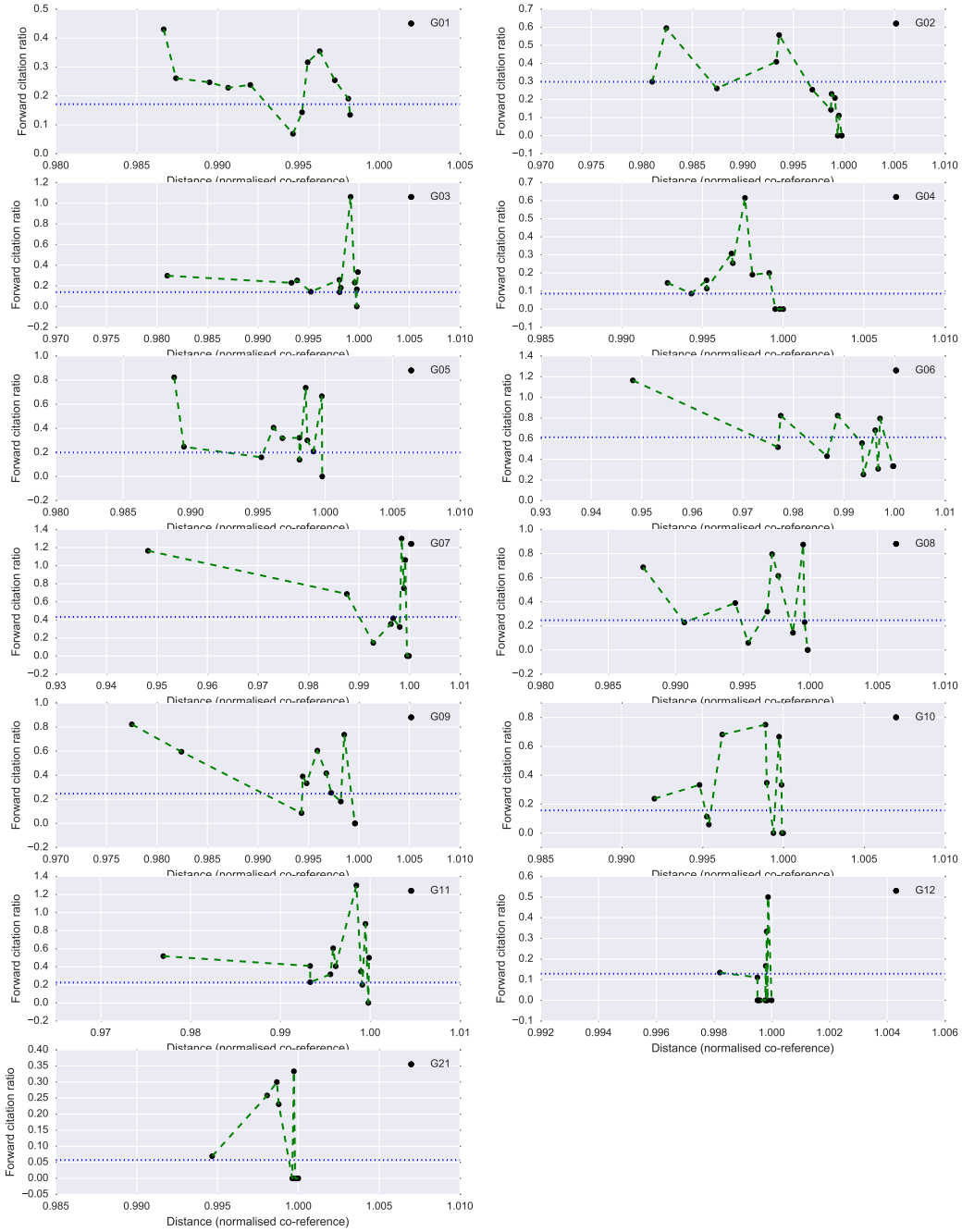

Fig. 3.20 Similar to Fig. 3.17 but for CPC3 technologies within Section G.

Table 3.4 Highly-cited patents with both G06 and G07 classifications.

Patent number	Patent Title	# Forward Citations
4799156	Interactive market management system	471
4002886	Electronic price display unit	214
4652998	Video gaming system with pool prize structures	207
4952928	Adaptable electronic monitoring and identification system	168
4815741	Automated marketing and gaming systems	162
4567359	Automatic information, goods and services dispensing system	155
4415065	Restaurant or retail vending facility	152
4669730	Automated sweepstakes-type game	151
4262632	Electronic livestock identification system	106

Chapter 4

Conclusions and Discussion

4.1 Evaluation and Discussion

4.1.1 Evaluation

Recall that the initial goals of the project were firstly to try to see if expected the inverted U-shape curve in the technology-distance/patent-value relationship exists in our data; and secondly to use the optimal distance from this curve to guide recommendations for new technology areas.

Our results show no evidence of the a clear inverted U-shape in our patent data. This is a useful result in itself, challenging

The project results showed that the inverted U-shaped relation between radically (technological distance) and patent value is not apparent. This may be due to a number of factors, including:

- The IPC and CPC classifications do not adequately represent the domain.
- The forward citation measure of patent value is too unreliable
- Firms self-select innovations to patent, effectively avoiding the extreme of the U-shape, with only an almost linear central section remaining in the patent data.
- The inverted U-shape relation does not actually occur in the real world. Other factors in real innovation affect the relationship.

There is evidence that the IPC codes (and by extension the similar CPC codes) may not be good indicators of intellectual content. Leydesdorff (2008)

In Chapter 3, we saw that the patent count and forward citation ratios, although not revealing evidence of the inverted U-shape, do show features which reflect our intuitions

about the nature of technologies: the centrality of computing devices; the emergence of genomics in the period covered; and the isolationist attitude of the nuclear industry. So clearly, these classifications codes to embody valuable information about the technologies in the patent system.

Our results also show how the patent data can be used for more “local" guidance, showing nearby areas of higher citation and hence patent value.

It may be that the U-shape curve may be visible if there were six or more IPC3, but that firms simply do not investigate such innovation, or do not patent any that arise.

Criscuolo et al. (2017)[1] discuss intra-firm selection effects on what to patent. Could be relevant as we may not see any curve if firms as assignees avoid extremes of the curve. It may be that only a sestet of idea get into the patent database, excluding the extremes.

Ferguson and Carnabuci (2017) QUOTE: "We draw on insights from research into institutional gatekeeping to theorize that, to be granted, patent applications that span technological domains must have higher quality than otherwise comparable, narrower applications. Using data on failed and successful patent applications, we estimate an integrated, two–tage model that accounts for this differential selection. We find that more domain-spanning patent applications are less likely to be approved, and that controlling for this differential selection reduces the estimated effect of knowledge recombination on innovative impact by about one-third." So there is a selection bias, but it may not fully account for any effect we see.

Initial results seem to be unable to find the peak in the inverted U-shaped curve. Chasing this up in the literature came across Katila and Ahuja (2002). They undertake empirical studies and expect a curvilinear relationship for both search depth (exploitation) and search breadth (exploration). They observe the former, but not the latter which is a linear relationship, innovation going up with search breadth. They suggest that this may be that firms simply do not reach or go beyond the peak on the curve and only see the linear part. This is because wide breadth searching is so much more costlier than deep depth search, which can use the resources already available to the firm.

Alstott17 also says:"Controlling for spurious factors quantifies an intuitive fact: most technologies are not particularly proximate to each other. "

Mind you Novelli (2015) note that patents classified in many classes will also, and be more likely to be found in more firms’ patent searches. Thus, they will be more likely to be an input to others work and patents. These combined comments may show that my use of multiple sections, for example, is not a valid approach, possibly just reflecting the turning-up in patent searches

[1] https://sites.insead.edu/facultyresearch/research/file.cfm?fid=60330

4.1.2 Discussion

One aspect of this project that was proved to be a greater problem than expected was the quality of the data. It is generally acknowledged that obtaining and cleaning data are the most time-consuming parts of any data science project, so some time was expected on this. However, it was surprising that the data with which we began proved to be so full of errors.

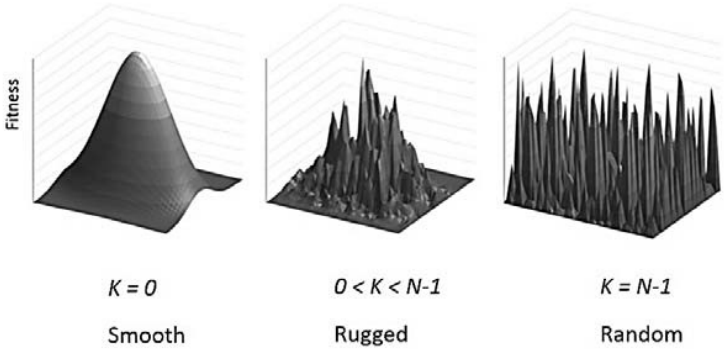

Fig. 4.1 from Alston and Mueller (2015).

The original project goals were to use the US patent data to identify any inverted U-shaped curve in the relation between “technology distance" (a measure of how radical an invention is from current technologies) and the value of a patent for this invention.

If present, the location of the optimal distance (the maximum patent value) was to be used to recommend a technology mix to a firm.

Early on, there was no evidence of any curvilinear reaction in any of the IPC codes that were investigated. Indeed, the relations between patent value (as measured by forward citations accrued by a patent) any any IPC classification seemed to follow no clear pattern.

Recall that we began using the patent data from the NBER, as this was a well used dataset, which we thought would have had most of the errors “shaken out". These are the results presented in the poster session (the poster is on SAFE). But, it then became apparent that the IPC classification data that we were using from that database were not from the actual patent documents, but derived from the US patent classification by a look-up to a concordance (a crosswalk). [2]

This issue with the NBER data led us to use the data from a collaborator (Bowen Yan)of the supervisory team. There were early doubts about this data too, as it took considerable cleaning to remove IPC codes that were invalid. Results from this data did not show any particular inverted U-shape either (these results are given in §3.2 above).

[2] On a personal note: having worked for many years as a research astronomer, I was somewhat shocked to find that many of the papers in the patent data field do not fully describe the provenance of their datasets. This make reproducible research very difficult, if not impossible.

Much of this work has been the search for a way of measuring distances – in this case between technologies/inventions. Metrics lie at the heart of machine learning.

The results indicate that the best recommendation to a firm is just to innovate across technologies, increasing the mix. Of course this is not too helpful. Although what does seem practical is to give local advice based on the heat maps.

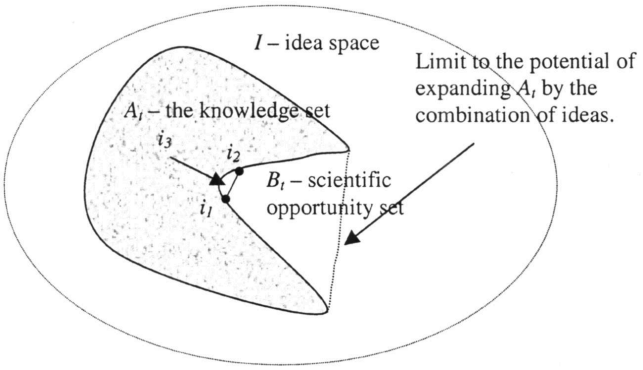

Fig. 4.2 Figure 1 from Olsson (2000).

4.2 Limitations of Study

Although somewhat premature, there are already a number of limitations that need to be noted in this study. Some of these limitations have arise from the literature, and have bedevilled other researchers – others have arisen during the coding to date. These limitations also point to future work that might be undertaken.

4.2.1 Provenance of the IPC classifications

The major issue that arose during the project

http://www.patentsview.org/download/ may be the best place to start. It uses the CPC, which is a new classification system, but may avoid some issues. One of the files is cpc_current.csv, which has the CPC classifications for all patnes

4.2.2 Issues of the patent system overall

One possible concern is the have evidence which suggests that many major innovations are not actually patented at all Fontana et al. (2013). Hopefully, the number of un-patented innovations is small compared to the total number of patents in the database, but if the former are significant patents (and radical) it would impact our recommendations.

Empirical evidence from the wider US patent data indicates that a major change in the US patent system seems to have occurred between the years 1990 and 2000 Strandburg (2009). This is right within the time-frame of the NBER used in this study.

One possible concern is the evidence that the most important innovations are *not* patented Fontana et al. (2013), making our search for valuable patents less interesting (although their

study was rather limited in scope). However, Fontana et al. (2013) also suggests that certain sectors have a greater propensity than others, indicating those economic sectors on which we should focus. They also compare their results with the propensity values from other studies and show that there is still much uncertainty anyway. Similarly, citeGress10 found that the various technology categories that they studied in the US patent data exhibited differed in their overall statistical properties. Hence, drawing inferences across technology sectors may not be valid. This result further support the case to break the data into economic/technology sectors.

It may best to look at industry sectors, by firm, rather than by patent classifications systems, as a way of generating coherent communities that then we apply the patent analysis to.

? compares the IPC for specific medical patents with a full medical ontology, the Medical Subject Headings ontology (MeSH). Although they are structurally similar there are many differences and they suggest that the patent classification needs to be improved. One way to do this my be by text analysis rathe that by the patent examiner. New technologies hold out the hope that this can be done in an automated fashion.

Historical, not current

The NBER data covers the years 1997 to 2006. These limits make it difficult to look at backward and forward citations outside these years.

Assignees not at clean as one would have hoped

Despite extensive cleaning of assignee names, working with the data has shown that there are still many assignees that should be given he same assignee ID but do not. There are, in particular, misspelling and other naming issues that were found during coding for many firms. Examples include Volkswagen and Hewlett Packard. As the assignee is one of the key elements of our study and recommendation, these errors could have a major and unpredictable influence on the recommendations we make.

As mentioned in chapter **??**, there are obvious limitations.

Firstly, to what extent do patents really represent innovative activity.

Secondly, many of the measures used to get value of a patent rely on forward citation, but these are not available for recent patents so make use in the current epoch a problem.

4.3 Future Work

The current study can be extended in a number of ways. It would be interesting to move the timeframe of the study to the most recent patent data, so that it could be used in the present day.

Kim, Daniel et al. (2016) also note, of interest to us, that they only use the forward citation counts unto 5 years only from publication as “it is well known" that only recent citation data works in prediction of a patents impact, see Benson CL, Magee CL (2015).

4.3.1 Heat diffusion

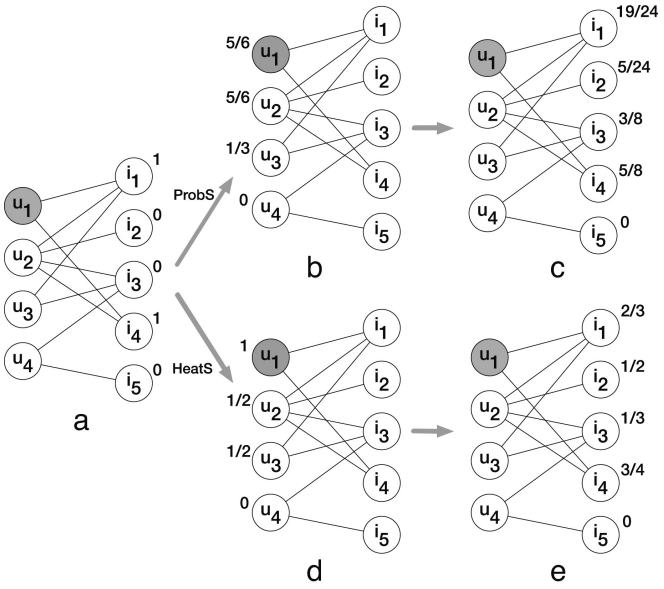

Fig. 4.3 Figure 1 from **?**

Zhou10 claim that the heat diffusion model generates the best predictions in terms of novelty (surprisal), a measure which they estimate from purely theoretical considerations. It would be good to test this idea by actually looking at predictions to see if they are novel, and maybe by other measures as given in this chapter. Vidmer15 also use the heat diffusion model which they equally argue avoids the preferential attachment found in traditional recommendations (e.g. Netflix) and so better captures the need for resources found in nation-product export data. So does this mean that playing with heat diffusion I can cover both my concern in the same way? And then spend a week testing the impacts of this approach?

If time is available, I will also look at alternative approaches to increasing the novelty of recommendations. In particular, Zhou et al. (2010) and Arnaboldi et al. (2017) use a heat

diffusion algorithm, which they argue/show increases diversity of recommendations. Zhou et al. (2010) claim that the heat diffusion model generates the best predictions in terms of novelty (surprisal), a measure which they estimate from purely theoretical considerations. It would be good to test this idea by actually looking at predictions to see if they are novel by other measures (such as those listed below).

Vidmer et al. (2015) also use the heat diffusion model which they argue that it avoids the preferential attachment found in traditional recommendations (e.g. Netflix,) and so better captures the need for resources found in their nation-product export data.

Note that Vidmer uses a heat diffusion add-on the link prediction. This is also mentioned by Zhou10, and in more general terms by Lu+12. see also Zhang07

Vidmer et al. (2015) uses a measure, as part of their link prediction, the “Revealed Competitive Advantage" (RCA). This is used to set a link between a country and a product (i.e. a link only exists if the export is important to a country?), if the RCS is greater than 1.

$$RCA_{i\alpha} = \frac{e_{i\alpha}}{\sum_{\beta} e_{j\beta}} \Big/ \frac{\sum_{j\alpha} e_{j\alpha}}{\sum_{j\beta} e_{j\beta}} \tag{4.1}$$

4.3.2 Patent text analysis

The use of the patent classification codes, such as the IPC and the CPC, may be one source of the difficulties in identifying technology distances. These classification systems were designed for use in a epoch before substantial computing was available. In the last few years extensive text analysis of document databases has become readily available.

Early work in text analysis of patents tried to classify patents into the standard classification codes from the text in the patent (e.g. Fall et al., 2003; Wu et al., 2008). These are example os supervised learning, in which the system is trained using existing US or IPC classifications. If the classification systems are not adequate, an unsupervised approach would be more appropriate, finding underlying structure in the technology description of the patents.

An excellent review of patent analysis, including pointers to work on analysis of patent texts is Abbas et al. (2014)

Perlich and Lawrence (2008) and Venugopalan and Rai (2015) used the textual content of the patents. Venugopalan and Rai (2015) suggest that the problem lies with the classifications systems. Such studies typically analyses only a small subset of all patents held, due to the vast size of the full data, and the complexity of text analysis. Moreover, as Venugopalan and Rai (2015) notes, evidence suggest that the full text of the patent is not required: the patent’s abstract and claims are sufficient.

Recent work had revealed possible weaknesses of the US patent classifications. For example, in one study reported by Barirani et al. (2013), for the time period in their investigation, the USPTO code for nanotechnology (code number 977) was used for about 4000 patents, while a lexical search of the patent texts returned over 50,000 patents.

4.3.3 Scientific networks

Increasingly, the patent system is becoming more aligned to the scientific literature. Modern studies of technology development and innovation use both scientific paper and patents together due to their mutual interaction and similarities (e.g. Mina et al. (2007)). Indeed, Narin et al. (1997) show how citation by patents to the scientific literature are important. There is a similar relevant study Mina et al. (2007) of both scientific citation, and accompanying patent citation, in the field of treatments for coronary artery disease.

General comment: it seems (see Trajtenberg (1990), footnote 3, that many aspects of patent system study draw from analogy with scientific publishing and citation. It would be interesting if we could return the honour and apply some of the aspects that were later used only for the patent system to the scientific paper system (e.g. 'economic' aspects?).

In this section we review the literature on scientific collaboration and citation, as these have many similarities with the patent system, and they interact greatly. For example, from Italian patent data it was found that academic inventors play a more central role, they have more collaborators, than non-academic inventors Balconi et al. (2004). They help connect the network, having a high betweenness-centrality. Such a result would suggest that, atlas for this limited sample inventors, the academics would also produce patents of greater generality and possibly impact. It would seem that we could use this to help prioritise our link predictions.

Indeed, many patents of interest cite not only earlier patents but also scientific papers. For example, techniques to predict possible breakthrough scientific papers based on paper citations Ponomarev et al. (2014). Others have sought predict future citation counts in the scientific literature Yan et al. (2011).

Gruber et al. (2013) find that inventors with a scientific education are more likely to generate patents that span technological boundaries than those with an engineering degree. Moreover, having a doctorate is associated with a greater degree of recombination for all groups of inventors.

For example, Liben-Nowell and Kleinberg (2007) use co-citation networks in science for link prediction, and discuss a number of problems that they encounter which will likely have a bearing on our patent dataset example. And Yan et al. (2011) predict from the citations list of scientific journals having built a graph of papers and citations. This is relevant as the patents that are likely to be our main concern are those that not only link to previous patents,

but also to the scientific literature. This is because we are interested in those technical fields that are likely to have major impact.

More speculatively, Ponomarev et al. (2014) look at predicting journal papers that will be highly cited, to help identify candidate “breakthroughs". The measure they employ is the rate of citation over the first few years, as all their breakthrough papers had high citations in the beginning. They use citation data, and this may have some bearing on whether we can predict valuable patents. By analogy, we would be looking for patents that garner many citations early on. We could add such forward citation statistics to our database, and use this property as part of the link prediction and/or prioritisation.

The recombination of diverse ideas are found not only in the patent literature, but also in that of innovation in science (e.g. Uzzi et al. (2013)). Some have developed techniques to try and predict possible breakthrough scientific papers based on paper citations Ponomarev et al. (2014). Of course, as noted above, the realms of patent citation and journal-paper citation are becoming interwoven as R&D moves increasing to a science-based mode.

Issue of recombination of diverse ideas found all over the patent literature, and also in that of innovation in science (e.g. Uzzi et al. (2013)). There is a mix of conventionality and novelty. Perhaps (partly) the conventionality is needed for the result/idea to have impact, i.e. be accepted and exploited by the community.

Kim et al. (2016) adopt the same method as used previously on scientific papers Uzzi et al. (2013), and get a similar result Kim, Daniel et al. (2016). Namely that the highest impact patents comes from a mix of conventionality and novelty.

It is interesting that a similar relation has been found in the behaviour of science project review panels Boudreau et al. (2016). It was found that reviewers give lower scores to project close to their own area of expertise, and also to those that are highly novel.

References

Abbas, A., Zhang, L., and Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. *World Patent Information*, 37:3 – 13.

Acs, Z. J., Anselin, L., and Varga, A. (2002). Patents and innovation counts as measures of regional production of new knowledge. *Research Policy*, 31(7):1069 – 1085.

Adamopoulos, P. and Tuzhilin, A. (2014). On unexpectedness in recommender systems: Or how to better expect the unexpected. *ACM Trans. Intell. Syst. Technol.*, 5(4):54:1–54:32.

Ahuja, G. and Lampert, C. M. (2001). Entrepreneurship in the large corporation: a longitudinal study of how established firms create breakthrough inventions. *Strategic Management Journal*, 22(6-7):521–543.

Alston, L. and Mueller, B. (2015). Towards a more evolutionary theory of property rights. *Iowa Law Review*, 100(6).

Alstott, J., Triulzi, G., Yan, B., and Luo, J. (2017). Mapping technology space by normalizing patent networks. *Scientometrics*, 110(1):443–479.

Ardito, L., Petruzzelli, A. M., and Panniello, U. (2016). Unveiling the breakthrough potential of established technologies: an empirical investigation in the aerospace industry. *Technology Analysis & Strategic Management*, 28(8):916–934.

Arnaboldi, V., Campana, M. G., Delmastro, F., and Pagani, E. (2017). A personalized recommender system for pervasive social networks. *Pervasive and Mobile Computing*, 36:3 – 24. Special Issue on Pervasive Social Computing.

Balconi, M., Breschi, S., and Lissoni, F. (2004). Networks of inventors and the role of academia: an exploration of italian patent data. *Research Policy*, 33(1):127 – 145.

Barirani, A., Agard, B., and Beaudry, C. (2013). Discovering and assessing fields of expertise in nanomedicine: a patent co-citation network perspective. *Scientometrics*, 94(3):1111–1136.

Benner, M. and Waldfogel, J. (2008). Close to you? bias and precision in patent-based measures of technological proximity. *Research Policy*, 37(9):1556 – 1567.

Benson, C. L. and Magee, C. L. (2015). Quantitative determination of technological improvement from patent data. *PLoS One*, 10(4):e0121635.

Benson, C. L. and Magee, C. L. (2016). *Using Enhanced Patent Data for Future-Oriented Technology Analysis*, pages 119–131. Springer International Publishing.

References

Block, J., Miller, D., Jaskiewicz, P., and Spiegel, F. (2013). Economic and technological importance of innovations in large family and founder firms. *Family Business Review*, 26(2):180–199.

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., and Riedl, C. (2016). Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science*, 62(10):2765–2783.

Breschi, S., Lissoni, F., and Malerba, F. (2003). Knowledge-relatedness in firm technological diversification. *Research Policy*, 32(1):69 – 87.

Bronwyn H. Hall, Adam Jaffe, M. T. (2005). Market value and patent citations. *The RAND Journal of Economics*, 36(1):16–38.

Bruck, P., Réthy, I., Szente, J., Tobochnik, J., and Érdi, P. (2016). Recognition of emerging technology trends: Class-selective study of citations in the u.s. patent citation network. *Scientometrics*, 107(3):1465–1475.

Carpenter, M. P., Narin, F., and Woolf, P. (1981). Citation rates to technologically important patents. *World Patent Information*, 3(4):160–163.

Cassi, L. and Plunket, A. (2014). Proximity, network formation and inventive performance: in search of the proximity paradox. *The Annals of Regional Science*, 53(2):395–422.

Cecere, G. and Ozman, M. (2014). Innovation, recombination and technological proximity. *Journal of the Knowledge Economy*, 5(3):646–667.

Cloodt, M., Hagedoorn, J., and Kranenburg, H. V. (2006). Mergers and acquisitions: Their effect on the innovative performance of companies in high-tech industries. *Research Policy*, 35(5):642 – 654.

Czarnitzki, D., Hussinger, K., and Schneider, C. (2011). Wacky patents meet economic indicators. *Economics Letters*, 113(2):131 – 134.

Dalum, B., Pedersen, C. O. R., Villumsen, G., Dalum, B., Pedersen, C. O. R., and Villumsen, G. (2005). Technological life-cycles, lessons from a cluster facing disruption. In *European Urban and Regional Studies*, pages 229–246.

Della Malva, A. and Riccaboni, M. (2014). (un)conventional combinations: At the origins of breakthrough inventions.

Dolfsma, W. and Leydesdorff, L. (2011). Innovation systems as patent networks: The netherlands, india and nanotech. *Innovation*, 13(3):311–326.

Ejermo, O. (2005). Technological diversity and jacobs' externality hypothesis revisited. *Growth and Change*, 36(2):167–195.

Elliott, G. (2007). Basics of US patents and the patent system. *AAPS Journal*, 9(3):E317–24.

Falk, N. and Train, K. (forthcoming). The relation of patent characteristics to forward citations. *Journal of Business Valuation and Economic Loss Analysis*.

Fall, C. J., Torcsvari, A., Benzineb, K., and Karetka, G. (2003). Automated categorization in the international patent classification. *SIGIR Forum*, 37(1):10–25.

Ferguson, J.-P. and Carnabuci, G. (2017). Risky recombinations: Institutional gatekeeping in the innovation process. *Organization Science*, 28(1):133–151.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Manage. Sci.*, 47(1):117–132.

Fleming, L. and Sorenson, O. (2004). Science as a map in technological search. *Strategic Management Journal*, 25(8-9):909–928.

Fontana, R., Nuvolari, A., Shimizu, H., and Vezzulli, A. (2013). Reassessing patent propensity: Evidence from a dataset of r&d awards, 1977–2004. *Research Policy*, 42(10):1780 – 1792. Economics, innovation and history: Perspectives in honour of Nick von Tunzelmann.

Fornahl, D., Broekel, T., and Boschma, R. (2011). What drives patent performance of German biotech firms? the impact of R&D subsidies, knowledge networks and their location. *Papers in Regional Science*, 90(2):395–418.

Gilsing, V., Nooteboom, B., Vanhaverbeke, W., Duysters, G., and van den Oord, A. (2008). Network embeddedness and the exploration of novel technologies: Technological distance, betweenness centrality and density. *Research Policy*, 37(10):1717 – 1731. Special Section Knowledge Dynamics out of Balance: Knowledge Biased, Skewed and Unmatched.

Gress, B. (2010). Properties of the USPTO patent citation network: 1963-2002. *World Patent Information*, 32(1):3–21.

Gruber, M., Harhoff, D., and Hoisl, K. (2013). Knowledge recombination across technological boundaries: Scientists vs. engineers. *Management Science*, 59(4):837–851.

Hall, B. H., Jaffe, A. B., and Trajtenberg, M. (2001). The NBER patent citation data file: Lessons, insights and methodological tools. Working Paper 8498, National Bureau of Economic Research.

Han, Y.-J. and Park, Y. (2006). Patent network analysis of inter-industrial knowledge flows: The case of korea between traditional and emerging industries. *World Patent Information*, 28(3):235 – 247.

Harhoff, D., Narin, F., Scherer, F. M., and Vopel, K. (1997). Citation Frequency and the Value of Patented Innovation. CIG Working Papers FS IV 97-26, Wissenschaftszentrum Berlin (WZB), Research Unit: Competition and Innovation (CIG).

Harris, C. G., Arens, R., and Srinivasan, P. (2010). Comparison of IPC and USPC classification systems in patent prior art searches. In *Proceedings of the 3rd International Workshop on Patent Information Retrieval*, PaIR '10, pages 27–32, New York, NY, USA. ACM.

Hsu, D. H. and Lim, K. (2014). Knowledge brokering and organizational innovation: Founder imprinting effects. *Organization Science*, 25(4):1134–1153.

Jaffe, A. B. (1986). Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits, and market value. *The American Economic Review*, 76(5):984–1001.

References

Jaffe, A. B., Trajtenberg, M., and Fogarty, M. S. (2000). The meaning of patent citations: Report on the nber/case-western reserve survey of patentees. Working Paper 7631, National Bureau of Economic Research.

Kasmire, J., Korhonen, J. M., and Nikolic, I. (2012). How radical is a radical innovation? an outline for a computational approach. *Energy Procedia*, 20(Supplement C):346 – 353.

Katila, R. and Ahuja, G. (2002). Something old, something new: A longitudinal study of search behavior and new product introduction. *The Academy of Management Journal*, 45(6):1183–1194.

Kauffman, S., Lobo, J., and Macready, W. G. (2000). Optimal search on a technology landscape. *Journal of Economic Behavior & Organization*, 43(2):141 – 166.

Keijl, S., Gilsing, V., Knoben, J., and Duysters, G. (2016). The two faces of inventions: The relationship between recombination and impact in pharmaceutical biotechnology. *Research Policy*, 45(5):1061 – 1074.

Kim, Daniel, Cerigo, Daniel Burkhardt, Jeong, Hawoong, and Youn, Hyejin (2016). Technological novelty profile and invention's future impact. *EPJ Data Sci.*, 5(1):8.

Lanjouw, J. and Schankerman, M. (2004). Patent quality and research productivity: Measuring innovation with multiple indicators. *Economic Journal*, 114(495):441–465.

Leydesdorff, L. (2008). Patent classifications as indicators of intellectual organization. *JASIST*, 59(10):1582–1597.

Li, G.-C., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., Yu, A. Z., and Fleming, L. (2014a). Disambiguation and co-authorship networks of the u.s. patent inventor database (1975–2010). *Research Policy*, 43(6):941 – 955.

Li, R., Chambers, T., Ding, Y., Zhang, G., and Meng, L. (2014b). Patent citation analysis: Calculating science linkage based on citing motivation. *Journal of the Association for Information Science and Technology*, 65(5):1007–1017.

Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, 58(7):1019–1031.

Lu, L., Medo, M., Yeung, C. H., Zhang, Y.-C., Zhang, Z.-K., and Zhou, T. (2012). Recommender systems. *Physics Reports*, 519(1):1 – 49. Recommender Systems.

Luan, C., Hou, H., Wang, Y., and Wang, X. (2014). Are significant inventions more diversified? *Scientometrics*, 100(2):459–470.

Mastrogiorgio, M. and Gilsing, V. (2016). Innovation through exaptation and its determinants: The role of technological complexity, analogy making & patent scope. *Research Policy*, 45(7):1419 – 1435.

Mehta, N. (2005). Measuring organizational scientific productivity: a study at NCL. *Current Science*, 88(2):223–230.

Mina, A., Ramlogan, R., Tampubolon, G., and Metcalfe, J. (2007). Mapping evolutionary trajectories: Applications to the growth and transformation of medical knowledge. *Research Policy*, 36(5):789 – 806.

Moore, K. A. (2005). Worthless patents. *Berkeley Technology Law Journal*, 20:1521–1552.

Narin, F., Hamilton, K. S., and Olivastro, D. (1997). The increasing linkage between u.s. technology and public science. *Research Policy*, 26(3):317 – 330.

Nemet, G. F. and Johnson, E. (2012). Do important inventions benefit from knowledge originating in other technological domains? *Research Policy*, 41(1):190 – 200.

Nooteboom, B., Haverbeke, W. V., Duysters, G., Gilsing, V., and van den Oord, A. (2007). Optimal cognitive distance and absorptive capacity. *Research Policy*, 36(7):1016 – 1034.

Novelli, E. (2015). An examination of the antecedents and implications of patent scope. *Research Policy*, 44(2):493 – 507.

Olsson, O. (2000). Knowledge as a set in idea space: An epistemological view on growth. *Journal of Economic Growth*, 5(3):253–75.

Olsson, O. (2005). Technological opportunity and growth. *Journal of Economic Growth*, 10(1):31–53.

Perlich, C., S. and Lawrence, R. (2008). Content-based link prediction for patent marketing. Technical Report RC24857, IBM Research Division, Thomas J. Watson Research Centre, P.O. Box 218, Yorktown Heights, NY.

Petruzzelli, A. M., Rotolo, D., and Albino, V. (2015). 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.

Ponomarev, I. V., Williams, D. E., Hackett, C. J., Schnell, J. D., and Haak, L. L. (2014). Predicting highly cited papers: A method for early detection of candidate breakthroughs. *Technological Forecasting and Social Change*, 81:49 – 55.

Schadt, M. (1992). Field-effect liquid-crystal displays and liquid-crystal materials: key technologies of the 1990s. *Displays*, 13(1):11 – 34.

Schoenmakers, W. and Duysters, G. (2010). The technological origins of radical inventions. *Research Policy*, 39(8):1051 – 1059.

Schumpeter, J. A. (1939). *Business cycles : a theoretical, historical, and statistical analysis of the capitalist process*. Martino Publishing.

Shane, S. (2001). Technological opportunities and new firm creation. *Management Science*, 47(2):205–220.

Sorensen, J. B. and Stuart, T. E. (2000). Aging, obsolescence, and organizational innovation. *Administrative Science Quarterly*, 45(1):81–112.

Spulber, D. F. (2015). How patents provide the foundation of the market for inventions. *Journal of Competition Law & Economics*, 11(2):271–316.

References

Sternitzke, C., Bartkowski, A., and Schramm, R. (2008). Visualizing patent statistics by means of social network analysis tools. *World Patent Information*, 30(2):115–131.

Strandburg, K. J. e. a. (2009). Patent citation networks revisited: Signs of a twenty-first century change? *North Carolina Law Review*, 87(6):1657.

Strumsky, D. and Lobo, J. (2015). Identifying the sources of technological novelty in the process of invention. *Research Policy*, 44(8):1445 – 1461.

Stuart, T. E. and Podolny, J. M. (1996). Local search and the evolution of technological capabilities. *Strategic Management Journal*, 17(S1):21–38.

Suh, J. H. (2015). Exploring the effect of structural patent indicators in forward patent citation networks on patent price from firm market value. *Technology Analysis & Strategic Management*, 27(5):485–502.

Torrance, A. W. (2009). Patents and the regress of useful arts. *Columbia Science and Technology Law Review*, 10.

Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations. *The RAND Journal of Economics*, 21(1):172–187.

Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342(6157):468–472.

Venugopalan, S. and Rai, V. (2015). Topic based classification and pattern identification in patents. *Technological Forecasting and Social Change*, 94:236 – 250.

Verhoeven, D., Bakker, J., and Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. *Research Policy*, 45(3):707 – 723.

Vidmer, A., Zeng, A., Medo, M., and Zhang, Y.-C. (2015). Prediction in complex systems: The case of the international trade network. *Physica A: Statistical Mechanics and its Applications*, 436:188 – 199.

vom Stein, N., Sick, N., and Leker, J. (2015). How to measure technological distance in collaborations, the case of electric mobility. *Technological Forecasting and Social Change*, 97:154 – 167.

Weitzman, M. L. (1998). Recombinant growth*. *The Quarterly Journal of Economics*, 113(2):331–360.

Wu, C. H., Ken, Y., and Huang, T. (2008). The support vector machine classification system for patent document information importance analysis. In *2008 International Conference on BioMedical Engineering and Informatics*, volume 1, pages 375–379.

Wuyts, S., Colombo, M. G., Dutta, S., and Nooteboom, B. (2005). Empirical tests of optimal cognitive distance. *Journal of Economic Behavior & Organization*, 58(2):277 – 302.

Yan, B. and Luo, J. (2017). Filtering patent maps for visualization of diversification paths of inventors and organizations. *J. Assoc. Inf. Sci. Technol.*, 68(6):1551–1563.

Yan, R., Tang, J., Liu, X., Shan, D., and Li, X. (2011). Citation count prediction: Learning to estimate future citations for literature. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1247–1252, New York, NY, USA. ACM.

Youn, H., Strumsky, D., Bettencourt, L. M. A., and Lobo, J. (2015). Invention as a combinatorial process: evidence from US patents. *Journal of The Royal Society Interface*, 12(106).

Zhou, T., Kuscsik, Z., Liu, J.-G., Medo, M., Wakeling, J. R., and Zhang, Y.-C. (2010). Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515.

4.4 Sample Python Codes

Due to page length constraints, sample codes from the Notebooks are listed below. These codes generate examples of the key plots and heat maps, as used in the dissertation.

FILE: PatentsView_plots_analysis.ipynb

```
# PatentsView data
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
get_ipython().magic('matplotlib inline')

file_to_load = '/Users/ahuxor/MSc/Thesis_Project/Data/patentsView/cpc_current.tsv'
# read in CPC data, but only for columns with patent number, section, IPC3 and IPC4
# takes 30secs
pv_cpc_df = pd.read_csv(file_to_load,sep='\t',usecols=[1,2,3,4])

print(pv_cpc_df.head())
print('length of pv_cpc_df = ',len(pv_cpc_df))

# load in data used for Yan sample, just to get patent numbers for 1997-1990
ipc3_1976_1990_df = pd.read_csv('ipc3_1976_1990_df',usecols=[0,1])
print(ipc3_1976_1990_df.head())
print('length of ipc3_1976_1990_df =',len(ipc3_1976_1990_df))

#rename column for PID to patent_id
ipc3_1976_1990_df.rename(columns={'PID':'patent_id'}, inplace=True)
# now join this IPC3 file with the PV file to get their IPC codes for these patent numbers
pv_76_90_df = pd.merge(ipc3_1976_1990_df, pv_cpc_df, how='left', on= 'patent_id')
sample_patents = pv_76_90_df.patent_id.unique()
len(sample_patents)

# needed to match for citations in PV CPC file
pat76_06_assg_df = pd.read_stata('pat76_06_assg.dta')

# calculate section counts for PV data, for each patent
def get_section_counts(focal_patent):
    sections = list(set(pv_76_90_df[pv_76_90_df.patent_id==focal_patent].section_id.values))
    num_unique_elements = len(sections)
    return(num_unique_elements)

# get sections for PV data
# 1047550 (full sample) takes 2h 50min
patent_list = []
sections_counts_list = []
num_forward_cites_list = []
for patent in sample_patents:
    xx = get_section_counts(patent)
    record = pat76_06_assg_df[pat76_06_assg_df.patent==patent]
    if not record.empty:
        num_forward_cites = record.allcites.values[0] # get number of forward cites for each patent
        num_forward_cites_list.append(num_forward_cites)
        patent_list.append(patent)
        sections_counts_list.append(xx)

# create dataframe from these lists
section_counts_df = pd.DataFrame(
    {'patent': patent_list,
     'number_of_sections': sections_counts_list,
     'forward_citations': num_forward_cites_list
    })

# save this as a csv file for later use (as each takes about an hour to generate)
section_counts_df.to_csv('section_counts_PV_df.csv', index=False)

plt.scatter(section_counts_df.number_of_sections,section_counts_df.forward_citations,alpha=1)
plt.ylabel('# of forward citations')
plt.xlabel('# of sections')
plt.ylim([0,2000])
plt.savefig("/Users/ahuxor/MSc/Thesis_Project/msc_thesis/Figs/hist_sections_forward_cites_PV.jpg",
 bbox_inches = 'tight')

# look at specific patents with many sections
six_sections = section_counts_df[section_counts_df.number_of_sections == 6]
print('number of patents with six sections = ',len(six_sections))

# look for number forward cites > 50
high_cited = six_sections[(six_sections.forward_citations>50)]
hi_cited_patents = high_cited.patent.tolist()
# get details on these highly cited patents

for a_patent in hi_cited_patents:
    temp = ipc3_1976_1990_df[ipc3_1976_1990_df.patent_id == a_patent]
    print('patent # = ', a_patent)
print(high_cited.sort_values(by='forward_citations', ascending=False))

# get a set of plots for the slopes

# first, get max number of sections for any patent
max_num_sections = max(section_counts_df.number_of_sections)

slope_list = []
intercept_list = []
R_squared_list = []
legend_text1 = ['1 sections','2 sections','3 sections', '4 sections','5 sections','6 sections']
fig = plt.figure(figsize=(10,10))
for plot_number in range(1,(max_num_sections)+1):
    print('plot_number = ',plot_number)
    # get subset of those with 'x' sections
    x_sections = section_counts_df[section_counts_df.number_of_sections == plot_number]
    print('number patents = ', len(x_sections))
```

```
    x_forward_citations = x_sections.forward_citations
    #print(x_forward_citations)
    hist_data = np.histogram(x_forward_citations,bins=50)
    #print(hist_data)
    x = hist_data[1][1:]
    y = hist_data[0][0:]
    logx = np.log10(x)
    logy = np.log10(y)
    # create dataframe from log values
    # as this may help get them into a regression
    logged_df = pd.DataFrame({'logx':logx, 'logy':logy})
    # remove rows with NaN/Inf etc
    clean_logged_df = logged_df.replace([np.inf, -np.inf], np.nan)
    clean_logged_df = clean_logged_df.dropna()
    clean_logged_df_trim = clean_logged_df[clean_logged_df.logy>0]
    from sklearn import linear_model
    linear = linear_model.LinearRegression()
    trainX = np.asarray(clean_logged_df_trim.logx[:]).reshape(-1, 1)
    trainY = np.asarray(clean_logged_df_trim.logy[:]).reshape(-1, 1)
    testX = np.asarray(clean_logged_df_trim.logx).reshape(-1, 1)
    testY = np.asarray(clean_logged_df_trim.logy).reshape(-1, 1)
    linear.fit(trainX, trainY)
    linear.score(trainX, trainY)
    slope_list.append(linear.coef_[0][0])
    #print('Coefficient: \n', linear.coef_[0][0])
    intercept_list.append(linear.intercept_[0])
    #print('Intercept: \n', linear.intercept_[0])
    R_squared_list.append(linear.score(trainX, trainY))
    #print('R? Value: \n', linear.score(trainX, trainY))
    predicted = linear.predict(trainX)
    linear.fit(trainX, trainY)
    bin_name = legend_text1[(plot_number-1)]
    slope_text = 'slope =' + format(linear.coef_[0][0],'.2f')
    ax = fig.add_subplot(5,2,plot_number)
    ax.scatter(logged_df.logx, logged_df.logy)
    ax.plot(testX, linear.predict(testX), color='red', linewidth=2)
    ax.set_xlim([0.4,3])
    ax.set_ylim([-1.0,6])
    ax.text(0.5, 0, bin_name)
    ax.text(2, 5, slope_text)
    ax.set_xlabel('log(# of forward citations)')
    ax.set_ylabel('log(# of patents)')
fig.show()
plt.savefig("/Users/ahuxor/MSc/Thesis_Project/msc_thesis/Figs/plots_slopes_sections_PV.pdf",
bbox_inches = 'tight')

# # IPC3
## calculate IPC3 classes in a patent
def get_ipc3_counts(focal_patent):
    ipc3s = list(set(pv_76_90_df[pv_76_90_df.patent_id==focal_patent].subsection_id.values))
    num_unique_elements = len(ipc3s)
    return(num_unique_elements)

# new sample for testing IPC3 classes plots
sample_patents_ipc3s = sample_patents # make same as sections, for now
len(sample_patents_ipc3s)

# get sections for all IPC3 classes data
# 1047550 (full sample) takes  2hr 44min
patent_list = []
ipc3_counts_list = []
num_forward_cites_list = []
for patent in sample_patents_ipc3s:
    xx = get_ipc3_counts(patent)
    record = pat76_06_assg_df[pat76_06_assg_df.patent == patent]
    if not record.empty:
        num_forward_cites = record.allcites.values[0] # get number of forward cites for each patent
        #print(forward_cites)
        num_forward_cites_list.append(num_forward_cites)
        patent_list.append(patent)
        ipc3_counts_list.append(xx)

# create dataframe from these lists
ipc3_counts_df = pd.DataFrame(
    {'patent': patent_list,
     'number_of_ipc3': ipc3_counts_list,
     'forward_citations': num_forward_cites_list
    })

# save this as a csv file for later use (as each takes about an hour to generate)
ipc3_counts_df.to_csv('ipc3_counts_df_PV.csv', index=False)

# plot histogram of forward citatiosn against number of ipc3 classes in any given patent
plt.scatter(ipc3_counts_df.number_of_ipc3,ipc3_counts_df.forward_citations,alpha=0.5,c='g')
plt.ylabel('# of forward citations')
plt.xlabel('# of CPC3 classes')
plt.ylim([0,2000])
plt.savefig("/Users/ahuxor/MSc/Thesis_Project/msc_thesis/Figs/hist_ipc3_V_forward_cites_PV.jpg",
bbox_inches = 'tight')

# get a set of plots for the slopes PV data
# first, get max number of sections for any patent
max_num_ipc3 = max(ipc3_counts_df.number_of_ipc3)

slope_list = []
intercept_list = []
R_squared_list = []
legend_text1 = ['1 CPC3','2 CPC3','3 CPC3', '4 CPC3','5 CPC3','6 CPC3','7 CPC3','8 CPC3','9 CPC3','
10 CPC3','11 CPC3','12 CPC3','13 CPC3']
fig = plt.figure(figsize=(10,10))
for plot_number in range(1,10):
    print('plot_number = ',plot_number)
```

```
    # get subset of those with 'x' ipc3 classes
    x_ipc3 = ipc3_counts_df[ipc3_counts_df.number_of_ipc3 == plot_number]
    print('number patents = ', len(x_ipc3))
    x_forward_citations = x_ipc3.forward_citations
    #print(x_forward_citations)
    hist_data = np.histogram(x_forward_citations,bins=50)
    #print(hist_data)
    x = hist_data[1][1:]
    y = hist_data[0][0:]
    logx = np.log10(x)
    logy = np.log10(y)
    # create dataframe from log values
    # as this may help get them into a regression
    logged_df = pd.DataFrame({'logx':logx, 'logy':logy})
    # remove rows with NaN/Inf etc
    clean_logged_df = logged_df.replace([np.inf, -np.inf], np.nan)
    clean_logged_df = clean_logged_df.dropna()
    clean_logged_df_trim = clean_logged_df[clean_logged_df.logy>0]
    from sklearn import linear_model
    linear = linear_model.LinearRegression()
    trainX = np.asarray(clean_logged_df_trim.logx[:]).reshape(-1, 1)
    trainY = np.asarray(clean_logged_df_trim.logy[:]).reshape(-1, 1)
    testX = np.asarray(clean_logged_df_trim.logx).reshape(-1, 1)
    testY = np.asarray(clean_logged_df_trim.logy).reshape(-1, 1)
    linear.fit(trainX, trainY)
    linear.score(trainX, trainY)
    slope_list.append(linear.coef_[0][0])
    #print('Coefficient: \n', linear.coef_[0][0])
    intercept_list.append(linear.intercept_[0])
    #print('Intercept: \n', linear.intercept_[0])
    R_squared_list.append(linear.score(trainX, trainY))
    #print('R? Value: \n', linear.score(trainX, trainY))
    predicted = linear.predict(trainX)
    linear.fit(trainX, trainY)
    bin_name = legend_text1[(plot_number-1)]
    slope_text = 'slope =' + format(linear.coef_[0][0],'.2f')
    ax = fig.add_subplot(5,2,plot_number)
    ax.scatter(logged_df.logx, logged_df.logy)
    ax.plot(testX, linear.predict(testX), color='red', linewidth=2)
    ax.set_xlim([0.4,3.5])
    ax.set_ylim([-1.0,6])
    ax.text(0.5, 5, bin_name)
    ax.text(2, 5, slope_text)
    ax.set_xlabel('log(# of forward citations)')
    ax.set_ylabel('log(# of patents)')
fig.show()
plt.savefig("/Users/ahuxor/MSc/Thesis_Project/msc_thesis/Figs/plots_slopes_ipc3_PV.pdf",
bbox_inches = 'tight')

stats = pd.DataFrame(
    {'slope': slope_list,
     'intercept': intercept_list,
     'R_squared': R_squared_list
    })

plt.scatter(stats.index.values,stats['slope'],marker='o',color='red',linewidth=4.0)
plt.plot(stats.index.values,stats['slope'],marker='.',color='blue',linewidth=1.0)
plt.title(' ')
plt.xlabel('# CPC3 codes')
plt.ylabel('slope of fit to power-law')
plt.savefig("/Users/ahuxor/MSc/Thesis_Project/msc_thesis/Figs/slopes_aggregate_plot_CPC3_PV.pdf",
bbox_inches = 'tight')
plt.show()
```

```
FILE : heatmaps_PV_data.ipynb

# ## Get distances from co-occurance, using PV data
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from scipy.spatial.distance import cosine
from matplotlib.colors import LogNorm
import seaborn as sns

get_ipython().magic('matplotlib inline')
get_ipython().magic('reload_ext line_profiler')

file_to_load = '/Users/ahuxor/MSc/Thesis_Project/Data/patentsView/cpc_current.tsv'
# read in CPC data, but only for columns with patent number, section, CPC3 and CPC4
# takes 30secs
pv_cpc_df = pd.read_csv(file_to_load,sep='\t',usecols=[1,2,3,4])

# a sanity check on the CPC sections indicates that the CPC codes have all been cleaned
unique_cpc_section_list = pv_cpc_df.section_id.unique().tolist()
print(unique_cpc_section_list)

# load in data used for Yan sample, just to get patent numbers for 1997-1990
ipc3_1976_1990_df = pd.read_csv('ipc3_1976_1990_df',usecols=[0,1])

#rename column for PID to patent_id
ipc3_1976_1990_df.rename(columns={'PID':'patent_id'}, inplace=True)

# now join this IPC3 file with the PV file to get the CPC codes for these patent numbers
pv_76_90_df = pd.merge(ipc3_1976_1990_df, pv_cpc_df, how='left', on= 'patent_id')
print('length of pv_76_90_df = ',len(pv_76_90_df))

# create list for unique_icl_section. Should just be the A to H sections
unique_cpc_section_list = ['A','B','C','D','E','F','G','H','Y']
# create list of unique assignee numbers
unique_patents = pv_76_90_df.patent_id.unique()
```

```
# create starter df for feature vectors based on section
feature_vectors_df = pd.DataFrame({'cpc_section' : ['A','B','C','D','E','F','G','H','Y']})
len(unique_patents)

# num = number of patents to look at (to test code with small samples first)
def get_patent_PVsection_counts_for_patents(num):
    feature_vectors_list_of_lists =[]
    # first get list of (unique) assignees (try first five first to test code)
    range_of_patents = unique_patents[0:num]
    # then for each one of them,
    for a_patent in range_of_patents:
        #print('patent number = ',a_patent)
        # look at all the patent sections each patent has #
        temp = pv_76_90_df[pv_76_90_df.patent_id==a_patent].section_id
        section_codes_in_current_list = list(temp.values)
        #print(section_codes_in_current_list)
        #  and count instances of these if multiple instances of any
        n = len(unique_cpc_section_list)
        cpc_codes_per_patent_list = []
        for m in range(0,n):
            current_cpc_section = unique_cpc_section_list[m]
            counted = section_codes_in_current_list.count(current_cpc_section)
            cpc_codes_per_patent_list.append(counted)
        feature_vectors_list_of_lists.append((cpc_codes_per_patent_list))
    return(feature_vectors_list_of_lists)

# 1047550 (full sample) takes 1h 33min
j=1047550
# %%prun -f get_patent_section_counts_for_patents feature_lists = get_patent_section_counts_for_patents(j)
PVsection_counts_per_patent = get_patent_PVsection_counts_for_patents(j)

PVsection_counts_per_patent_df = pd.DataFrame(PVsection_counts_per_patent)

# add column names to df from unique_cpc_section_list names
PVsection_counts_per_patent_df.columns = unique_cpc_section_list
# add patent numbers as index
# first get required number (function of j, remember), then convert to integer
patents_for_index = unique_patents[:len(PVsection_counts_per_patent_df)]
patents_for_index
# then change index for df
indexed_PVsection_counts_per_patent_df = PVsection_counts_per_patent_df.set_index(patents_for_index)
indexed_PVsection_counts_per_patent_df.tail()
# create clipped version of df with value =1 for any value of 1 and above
indexed_PVsection_counts_per_patent_df_clipped = indexed_PVsection_counts_per_patent_df.clip(0,1)
indexed_PVsection_counts_per_patent_df_clipped.tail()
# load data with forward citations
pat76_06_assg_df = pd.read_stata('pat76_06_assg.dta')

len(indexed_PVsection_counts_per_patent_df_clipped)

# full sample takes 2hrs 23 min
list_of_sections = ['A','B','C','D','E','F','G','H','Y']
#create empty lists to hold values generated
firsts_list_PVsections = []
seconds_list_PVsections = []
ratio_list_PVsections = []
patent_counts_list_PVsections = []
# set threshold value for forward citations, to derive ratio of high to low forward citations
threshold = 20
# create short-name version of the long but descriptive df, for use in loop
short_name_df = indexed_PVsection_counts_per_patent_df_clipped
for y in range(0,9):
    first = list_of_sections[y]
    for z in range(0,9):
        second = list_of_sections[z]
        #print('dyad=',first,'/', second)
        #get patents which have both of these sections
        matches_dyad = short_name_df[(short_name_df[first]==1) & (short_name_df[second]==1)]
        # throw out those that also have other sections (for clean sampling)
        matches_dyad_no_others = matches_dyad[matches_dyad.sum(axis=1)<3]
        #get actual patent numbers (used as index for thes dfs)
        patents = matches_dyad_no_others.index.values
        #print(patents)
        count_patents = len(patents)
        # obtain the number of forward citations
        forward_cites_list = []
        for a_patent in patents:
            # get record for this patent from 76-06-assg
            record = pat76_06_assg_df[pat76_06_assg_df.patent==a_patent]
            # have a case where the patent number has no record in pat76_06_assg_df, so exclude such cases
            if not record.empty:
                forward_cites = record.allcites.values[0] # get number of forward cites for each patent
                #print(forward_cites)
                forward_cites_list.append(forward_cites)
        # derive ratio citations above/below threshold
        num_above_threshold = len([i for i in forward_cites_list if i >= threshold])
        num_below_threshold = len([i for i in forward_cites_list if i < threshold])
        # need to handle num_below_threshold =0, as get divide by zero error in some cases
        if (num_below_threshold == 0):
            num_below_threshold = 1.0
        citations_ratio = num_above_threshold/num_below_threshold
        #print('ratio = ', citations_ratio)
        firsts_list_PVsections.append(first)
        seconds_list_PVsections.append(second)
        ratio_list_PVsections.append(citations_ratio)
        patent_counts_list_PVsections.append(count_patents)
print('DONE')

# create dataframe from these lists for the forward citations
forw_cite_ratio_df_PVsections = pd.DataFrame(
    {'first_section': firsts_list_PVsections,
     'second_section': seconds_list_PVsections,
```

```
    'forw_cite_ratio': ratio_list_PVsections
    })

# create dataframe from these lists for patent counts
patent_counts_df_PVsections = pd.DataFrame(
    {'first_section': firsts_list_PVsections,
     'second_section': seconds_list_PVsections,
     'patent_counts': patent_counts_list_PVsections
    })

# convert the forward citation dataframe into another representing the matrix
PVsection_forw_cite_ratios = forw_cite_ratio_df_PVsections.pivot(index='first_section',
columns='second_section', values='forw_cite_ratio')
# normalize the counts to account for different patent proposensity
normalized_counts_list_PVsections = []
for n in range(0,len(patent_counts_df_PVsections)):
    record = patent_counts_df_PVsections[n:n+1]
    first = record.first_section.values[0]
    first_lone = patent_counts_df_PVsections[(patent_counts_df_PVsections.first_section==first) &
     (patent_counts_df_PVsections.second_section==first) ]
    first_lone_counts = first_lone.patent_counts.values[0]
    first_lone_counts
    second = record.second_section.values[0]
    second_lone = patent_counts_df_PVsections[(patent_counts_df_PVsections.first_section==second) &
     (patent_counts_df_PVsections.second_section==second) ]
    second_lone_counts = second_lone.patent_counts.values[0]
    second_lone_counts
    if first==second:
        sum_lone_values = first_lone_counts
    else:
        sum_lone_values = first_lone_counts+second_lone_counts
    normalized_counts = record.patent_counts.values[0]/sum_lone_values
    normalized_counts_list_PVsections.append(normalized_counts)

# add the normalised counts to the df
patent_counts_df_PVsections['normalized_counts'] = normalized_counts_list_PVsections
patent_counts_df_PVsections.head()

# convert the forward citation dataframe into another representing the matrix
PVsection_patent_counts = patent_counts_df_PVsections.pivot(index='first_section',
columns='second_section',values='patent_counts')
# as above for for normalized counts
PVsection_patent_counts_norm =
patent_counts_df_PVsections.pivot(index='first_section',columns='second_section',
values='normalized_counts')

# plot raw counts
fig = plt.figure(figsize=(8,8))
sns.heatmap(PVsection_patent_counts, cmap='coolwarm', linewidths=0.5, annot=True, fmt="d",
cbar=False, norm=LogNorm(vmin=PVsection_patent_counts.min(),
vmax=PVsection_patent_counts.max()))
plt.savefig("/Users/ahuxor/MSc/Thesis_Project/msc_thesis/Figs/PVsection_patent_counts_heatmap.pdf",
bbox_inches = 'tight')
fig.show()

# plot normalised counts
fig = plt.figure(figsize=(8,8))
sns.heatmap(PVsection_patent_counts_norm, cmap='coolwarm', linewidths=0.5,cbar=False,annot=True,
norm= LogNorm(vmin=PVsection_patent_counts.min(),  vmax=PVsection_patent_counts.max()))
plt.savefig("/Users/ahuxor/MSc/Thesis_Project/msc_thesis/Figs/PVsection_patent_counts_norm_heatmap.pdf",
bbox_inches = 'tight')
fig.show()

# plot forward citation ratios above/below threshold of 20
fig = plt.figure(figsize=(8,8))
sns.heatmap(PVsection_forw_cite_ratios, cmap='coolwarm', linewidths=0.5, annot=True, cbar=False)
plt.savefig("/Users/ahuxor/MSc/Thesis_Project/msc_thesis/Figs/PVsection_cite_ratios_heatmap.pdf",
bbox_inches = 'tight')
fig.show()

# create multiplot of the above
cpc_section_codes = ['A','B','C','D','E','F','G','H','Y']

fig = plt.figure(figsize=(15,20))

for plot_number in range(1,10):
    print('plot_number = ',plot_number)
    code = cpc_section_codes[plot_number-1]
    result = pd.concat([PVsection_patent_counts_norm[code], PVsection_forw_cite_ratios[code]], axis=1)
    result.columns = ['norm_patent_counts', 'forward_citation_ratio']
    result['distance'] = 1-result.norm_patent_counts
    # sort on norm_patent_counts
    result_sorted = result.sort_values(by='distance')
    result_to_plot = result_sorted[1:]
    self_citation_rate = result_sorted[:1].forward_citation_ratio.values[0]
    ax = fig.add_subplot(5,2,plot_number)
    ax.scatter(result_to_plot['distance'],result_to_plot['forward_citation_ratio'],c='black', label=code)
    ax.plot(result_to_plot['distance'],result_to_plot['forward_citation_ratio'],'g--',label='')
    ax.set_xlabel('Distance (normalised co-reference)')
    ax.set_ylabel('Forward citation ratio')
    ax.legend(loc='upper right')
    plt.axhline(y=self_citation_rate,ls='dotted')
plt.savefig("/Users/ahuxor/MSc/Thesis_Project/msc_thesis/Figs/cpc_sections_plot_distance_forward_cites.pdf",
bbox_inches = 'tight')plt.show()
```